Global Studies Quarterly (2022) **2**, 1–13

Bias in Text Analysis for International Relations Research

Leah C. Windsor
The University of Memphis, USA

How international is political text-analysis research? In computational text analysis, corpus selection skews heavily toward English-language sources and reflects a Western bias that influences the scope, interpretation, and generalizability of research on international politics. For example, corpus selection bias can affect our understanding of alliances and alignments, internal dynamics of authoritarian regimes, durability of treaties, the onset of genocide, and the formation and dissolution of non-state actor groups. Yet, there are issues along the entire "value chain" of corpus production that affect research outcomes and the conclusions we draw about things in the world. I identify three issues in the data-generating process pertaining to discourse analysis of political phenomena: information deficiencies that lead to corpus selection and analysis bias; problems regarding document preparation, such as the availability and quality of corpora from non-English sources; and gaps in the linguist analysis pipeline. Short-term interventions for incentivizing this agenda include special journal issues, conference workshops, and mentoring and training students in international relations in this methodology. Longer term solutions to these issues include promoting multidisciplinary collaboration, training students in computational discourse methods, promoting foreign language proficiency, and co-authorship across global regions that may help scholars to learn more about global problems through primary documents.

À quel point les recherches d'analyse des textes politiques sont-elles internationales? Dans l'analyse computationnelle de textes, la sélection des corpus penche fortement vers les sources en anglais et reflète un parti pris occidental qui influence l'étendue, l'interprétation et la généralisabilité des recherches sur la politique internationale. La sélection des corpus peut par exemple affecter notre compréhension des alliances et des alignements, des dynamiques internes des régimes autoritaires, de la durabilité des traités, du déclenchement des génocides, et de la formation et de la dissolution des groupes d'acteurs non étatiques. Il y a des problèmes tout au long de la « chaîne de valeur » de la production de corpus qui affectent les résultats des recherches et les conclusions que nous tirons sur différents éléments du monde. J'ai identifié trois problèmes dans le processus de génération de données afférent à l'analyse discursive des phénomènes politiques: des lacunes dans les informations qui entraînent des biais dans la sélection et l'analyse des corpus, des problèmes concernant la préparation des documents, notamment en termes de disponibilité et de qualité des corpus provenant de sources qui ne sont pas en anglais, et des lacunes dans le pipeline d'analyse des linguistes. Les interventions à court terme pour encourager ce programme comprennent des numéros spéciaux de revues, des ateliers de conférences, ainsi que le mentorat et la formation des étudiants en RI à cette méthodologie. Les solutions à plus long terme à ces problèmes comprennent la promotion de la collaboration multidisciplinaire, la formation des étudiants aux méthodes d'analyse computationnelle des discours, la promotion de la maîtrise des langues étrangères et la co-rédaction à travers différentes régions du monde, ce qui peut aider les chercheurs à en savoir plus sur les problèmes mondiaux grâce à des documents primaires.

¿Qué tan internacional es la investigación política del análisis de textos? En el análisis computacional de textos, la selección del corpus se inclina, en gran medida, a las fuentes del idioma inglés y refleja un sesgo occidental que influencia el alcance, la interpretación y las posibilidades de generalización de la investigación de la política internacional. Por ejemplo, el sesgo en la selección del corpus puede afectar cómo entendemos las alianzas y alineaciones, las dinámicas internas de los regímenes autoritarios, la durabilidad de los tratados, el inicio del genocidio, y la formación y disolución de grupos de actores no estatales. Yet there are issues along the entire "value chain" of corpus production that affect research outcomes and the conclusions we draw about things in the world. Puedo identificar tres problemáticas en el proceso de generación de datos que corresponden al análisis del discurso del fenómeno político: deficiencias de información que conllevan a sesgos en la selección y el análisis del corpus; problemas vinculados a la preparación de documentos, tales como la disponibilidad y la calidad de los corpus que provienen de fuentes que no son del inglés; y diferencias en el canal del análisis lingüístico. Entre las intervenciones a corto plazo orientadas a incentivar este plan de acción se incluyen ediciones especiales de revistas y seminarios, así como también la asesoría y capacitación de estudiantes de relaciones internacionales en relación a esta metodología. Por otro lado, algunas soluciones a largo plazo para estos problemas incluyen la promoción de la colaboración multidisciplinaria, la capacitación de estudiantes en métodos computacionales del discurso, la estimulación de la competencia lingüística extranjera y la coautoría en las regiones globales, lo que podría ayudar a los investigadores a aprender más sobre los problemas globales a partir de documentos primarios.

Introduction

In studying international relations (IR), how do we know what we know about the world? The Sapir–Whorf hypothesis suggests that language shapes our worldview (Whorf 1957, 1940). Following this, the language we use to investigate sociopolitical phenomena in the international system influences our understanding and analyses of these processes. This type of language bias is encapsulated in the phrase "history is written by the winners" as well as "one man's terrorist is another man's freedom fighter." Perspective, and its accompanying narrative that reflects grammatical and lexical choices, varies by language, culture, and country. The corpora we choose do affect the answers we get: people in different contexts perceive the political world around them differently (Geddes 1990). In this article, I explore how the methodological area of computational text analysis informs our understanding of global political events, where the field of "text as data" has room for growth, and provide some suggestions for broadening the scope of corpus

Windsor, Leah C. (2022) Bias in Text Analysis for International Relations Research. *Global Studies Quarterly*, https://doi.org/10.1093/isagsq/ksac021

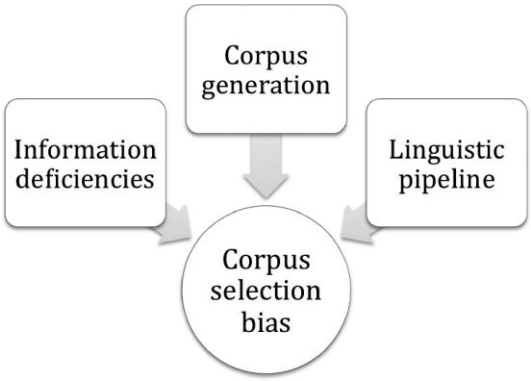

Figure 1. Overview of corpus selection bias factors

generation to incorporate the range of document sources to reflect greater linguistic and political diversity. In this epistemic analysis of computational text analysis, I suggest that democratizing the language and political perspectives will provide insight into the ways that implicit and explicit language bias affects what we know about IR.

Because non-English corpora are underrepresented in IR research using text-as-data methods, scholars' understanding of international political phenomena is filtered through an ethnocentric, Western lens. This can lead to scholars misdiagnosing or failing to understand political phenomena since the experiences of political processes omit the perspectives offered by accounts in non-English languages. If we democratize corpus collection and diversify the sources of information to include non-English languages, then scholars can begin to understand IR from a truly international perspective. If we work across disciplines to develop better computational tools to process and analyze non-English documents, we will begin to de-bias the research process and understand the variation in perspectives on IR across languages, cultures, and countries. This will fundamentally improve our understanding of quantities of interest such as cross-national and civil conflicts, collective action problems, human rights, populism, and peacemaking processes.

In this article, I identify three problems in the data-generating process (DGP) pertaining to discourse analysis of political phenomena as shown in figure 1: information deficiencies, document preparation, and linguistic analysis pipeline. The DGP goes by many names, often used interchangeably, such as computational text analysis, discourse analysis (either quantitative or qualitative), text-as-data, and quantitative text analysis, even though they may be derived via different processes. While these processes differ in some ways, they are commonly linked by the three problems facing scholars who study society, governance, and politics through the lens of language.

This article will proceed as follows. I first summarize the current text-as-data approach in context, including the increased number of IR papers using computational text-as-data methods. It merits noting that the point is not that no IR research is being done using these methods, but rather that what it represents is a mere fraction of what could and should be done given the vastness of potential corpora and phenomena to investigate. I then turn to the issue of information deficiencies, including why corpora are missing and underrepresentation of low resource languages. Next, I discuss the process of corpus generation and some of the technical problems facing researchers. I then look at the linguistic analysis pipeline—the tools available for analyzing documents. Next, I provide two examples—event data generation and ethnolinguistic fractionalization—which underscore the need for more representative corpora in IR. Finally, I provide future directions and conclusions for advancing text-as-data methods in IR research.

Text-as-Data in Context

In computational text analysis, corpus selection skews heavily toward English-language sources and reflects a Western bias that influences the scope, interpretation, and external validity of research on international politics (Colgan 2019a). Text-as-data is a methodology that can be applied to any area of study, including US government, comparative politics (CP), and international relations (IR). It is also interdisciplinary, drawing on techniques from computer science and linguistics among others, especially related to extracting political information from social media (Chiovaro, Windsor, and Paxton 2021; Chiovaro et al. 2021). While I focus most on issues related to IR, it bears noting that the Venn diagram overlap between IR and CP is substantial and the line demarcating one from the other is blurry. One delineation might be between qualitative and quantitative approaches, although this is neither necessary nor sufficient to land in either camp. To the extent that comparative and IR scholars study overlapping phenomena, the critique of textual bias applies.

Text mining of social media transcends disciplines and fields, as it is used extensively to understand sociopolitical processes such as social mobilization and protest, government repression, and decisions to close the Twitter or Weibo spigots in an attempt to thwart collective anti-government action (King, Pan, and Roberts 2013, 2014). Siegel and Pan (2018) extend the social mobilization literature popularized by McAdam reflecting on the civil rights movement (McAdam 1986, 1989) to the context of Saudi Arabia, where the find that government repression deterred the imprisoned but not those who still had their freedom from engaging in online dissent. Carter and Carter (2020) find similarly nuanced patterns around pro-democracy movement anniversaries in China but not for other holidays. Scholars have much yet to understand about the relationships between online behavior, civil society, and government responses that diversifying source material can help reveal. Social media scholars must also take care to ensure that they are measuring what they think they are measuring. For example, social media can also be problematic for the fact that default location settings for social media apps may be misleading, and diaspora groups far from the epicenter of conflict can influence local dynamics.

Most political research using this approach originates in the study of Western democratic institutions, focusing mostly on US institutions such as the federal courts and legislatures (Goh 2019). There are several reasons for this: most published scholarship in political science is authored by scholars in global north/Organisation for Economic Co-operation and Development (OECD) countries (Breuning et al. 2018a, 2018b), English-language and Western documents are generally archived in user-friendly formats that do not require extensive pre-processing, and most computational tools are available only for English-language corpora and at the disposal of. As a result, we have gained a lot of knowledge about Western democratic legislatures, courts, state politics, voting, and social and mainstream media. To a lesser extent, text-as-data workflows have been applied to non-Western inquiries, generally in the subfield of CP. Least studied are IR phenomena. The Global South Action Network has been instrumental in connecting scholars across the global north/south divide (Yannitell-Reinhardt 2021).

The relative ease with which many English-language documents can be downloaded, processed, and analyzed facilitates computational discourse research, since there is minimal preprocessing required. Western countries have institutional, technological, and infrastructural advantages to developing countries in their ability to produce, capture, process, and store linguistic data from political sources. Unsurprisingly, this research skews heavily toward US government institutions such as the judiciary and legislatures at the federal and state levels, as these documents are maintained in easily accessible databases and in user-friendly formats (Hill and Hurley 2002; Grimmer 2009; Osborn and Mendez 2010; Owens and Wedeking 2011; Rice and Zorn 2014; Hinkle and Nelson 2016). Through the application of text analysis to legislative language, Supreme Court deliberations, and Western democratic leadership, there has been much greater focus on participatory inclusion and gender, decision-making and reasoning, and rationales and deliberations regarding conflict participation. However, because this has often only considered English language texts and Western-style institutions, we are missing information about rising regional powers, the bargaining processes and capabilities of rebel and insurgent groups, and the deterioration of human rights practices.

Increasing International Relations Representation in Quantitative Text Analytics

Language reveals information about the most fundamental questions in IR research, such as those regarding war and peace: Is a leader bluffing or making a credible threat? Is a nuclear-armed state stable or unstable? Will states uphold their international commitments even amid political and economic strife? Are human rights likely to worsen or improve in a given country? However, the current state of the text-as-data field is presently underprepared to address these questions of international importance because of the dearth of textual information about most of the countries in the world. Data in non-Western societies are challenging to collect, clean, and analyze even given recent advancements in computational text manipulation. Yet, this should not deter IR scholars from collecting corpora from non-English and non-Western sources for the purpose of engaging in research about core IR questions. These overlooked regions represent a form of selection bias in IR scholarship that limits our ability to understand political processes from non-Western points of view.

Efforts to increase representation of IR scholarship using computational text-as-data methods stand alongside other discipline-wide concerted efforts to increase participation from scholars from the global south. IR research has largely been developed by Western scholars theorizing about non-Western phenomena. Including and incorporating the histories, narratives, and documents from non-Western settings will help scholars to better understand the root causes and consequences of conflict and peace. By exploring the ways in which the research of Western IR scholars fosters an epistemological bias given the reliance on English-language sources, we can identify ways forward to diversify not only political corpora, but academic partnerships across country and language divides.

Harnessing Language for International Relations Insights

The 2008 Political Analysis Special Issue on text analytics demarcates a renewed interest in computational discourse analysis for political research. As a result, the issue increased interest in the field, and in 2009 Harvard hosted the inaugural New Direction in Analyzing Text as Data conference that features innovative applications of discourse analysis. This conference has served as a conduit for cross-pollination of methodologies between the fields of computer science, linguistics, and political science. The fields of computer science and computational linguistics address questions of political importance but from varying disciplinary perspectives. The conference organizers' goals include bringing together researchers from the fields of political science and computer science to encourage cross-pollination of methodologies, foster interdisciplinary collaboration, and expand the realm of political research questions. While computational discourse analysis has been well established in the fields of computer science and linguistics, this approach is relatively new to the field of political science, emerging alongside the data science and "big data" revolution enabled largely by social media.

The field of IR lags behind both US government and, to a lesser extent, CP, in its use of computational text-analysis methods. Yet, in spite of an English-language bias in corpus selection, there is a robust and growing literature using discourse analysis to understand international political processes. This includes the language of lying and deception among dictators and terrorists, whereby neural networks and automated feature selection correctly predicted deceptive language two-thirds of the time (Hancock et al. 2010; Abrahms, Beauchamp, and Mroszczyk 2017). Similarly, it also includes threat detection (Hancock et al. 2010), statements of resolve whereby leaders who make more resolved statements tend to prevail in international disputes (Dyson 2006; Dyson and Preston 2006; McManus 2014, 2016, 2017; Kydd and McManus 2017), violent extremism (Windsor 2017b), rebel group image management (Jones and Mattiacci 2017), and international politics broadly defined (King and Lowe 2003; Cohen et al. 2008; Lowe 2008; Jurka et al. 2012).

Specific research has focused on leaders and countries, such as Saddam Hussein in Iraq, that show symmetry between personal and public statements (Brands and Palkki 2012; Dyson and Raleigh 2014; says Smithc2 2014; Shala, Rus, and Graesser 2014; Windsor et al. 2017); positively valenced language improving public opinion ratings of Hugo Chavez in Venezuela (Love and Windsor 2017), Vladimir Putin (Dyson 2001; Labzina and Nieman 2017), and Medvedev in Russia (Baturo and Mikhaylov 2014); the survival and longevity of long-term leaders such as Mao Tse Tung in China and how Chinese social media operatives strategically obscure political events that might cast negative light on the country (King, Pan, and Roberts 2013; Kreutz and Croicu 2014; Roberts et al. 2014; Windsor, Dowell, and Graesser 2014; Dowell, Windsor, and Graesser 2015; Li et al. under review); and explanations for the rise of populism in France and Belgium (Jagers and Walgrave 2007; Alduy and Wahnich 2015).

Political dynamics in the United Nations have been a recent focus of study (Windsor 2016; Baturo, Dasandi, and Mikhaylov 2017), and social media continues to dominate much of sociopolitical research, especially focusing on large-scale social mobilizations (Tumasjan et al. 2010; Segerberg and Bennett 2011; Aharony 2012; Qiu et al. 2012; Gupta et al. 2014; Ito et al. 2015; Beauchamp 2017). The Comparative Manifesto Project data include party platforms for more than fifty countries, for democracies in mostly OECD and Central and Eastern European countries. Studies using this data have made advances in measuring political party positions and in structural topic modeling—a process that

can help scholars working with multilingual documents with the effect of reducing corpus selection bias (Mikhaylov and Laver 2008; Volkens, Bara, and Budge 2009; Lucas et al. 2015; Lehmann et al. 2018).

During height of leadership trait analysis (LTA) and role theory in foreign policy research (Hermann 1987; Hermann and Preston 1994), scholars focused on the language of leaders to derive estimates of policy preferences. Scholars of LTA argue that particular leadership features and leaders matter, and these features are derived using linguistic and psychological personality metrics (Hermann et al. 2001). LTA has been used to describe how leadership is context-dependent (Cuhadar et al. 2017) and which leaders are likely to take risks (Kowert and Hermann 1997). LTA has provided insight into the psychological profiles of leaders (Hermann 1980; Hermann and Post 2003), what factors indicate stress in political leaders (Hermann 1979), how leadership traits influence bureaucratic choices (Preston and 't Hart 1999), and how leadership features of prime ministers and members of parliament vary (Kaarbo 1997; Kaarbo and Hermann 1998). While there has been much debate about how to calibrate the role that leaders play in international politics, even some realists have acknowledged that individuals matter (Jervis 2013). Recent LTA scholarship has provided opportunities to substantially diversify analyses in non-Western languages, including work by Brummer et al. (2020), Rabini et al. (2020), Canbolat (2021), and Thiers (2021).

The majority of text-as-data research in IR has focused on leaders and individuals. Thus, through the application of text analysis to the issues of leadership traits, we now better understand leaders' psychological calculus for various decisions including foreign and domestic policy choices and rhetoric. However, the reliance on translated or English-language sources means that we are forced to ignore what leaders have said—or what has been said about them—in the original language. This has implications for understanding domestic coalitions, coup-proofing, and changes in regime dynamics.

The event data domain is experiencing a reorientation of efforts to diversify the sources of information used to understand patterns of political events, especially contentious political events such as protests and riots (Raleigh et al. 2010; Lorenzini et al. 2016; Tubishat et al. 2019; Birch and Muchlinski 2020; Dowd et al. 2020; Sobolev et al. 2020). Social media can also liberate information in ways that traditional news media cannot (Steinert-Threlkeld 2017; Goebel and Steinhardt 2019; Zhang and Pan 2019; Driscoll and Steinert-Threlkeld 2020).

Information Deficiencies

The first problem in corpus selection bias in IR is information deficiencies. The origins of these deficiencies includes the following: a deficit of speakers or translated and validated documents from low-resource as well as widely spoken languages; unavailability of documents due to conflict in situ, lack of support for document preservation and archiving; opaque political environments that closely guard documents of interest such as terrorist groups or authoritarian regimes; and a dwindling proportion of polyglot IR researchers. There is even a dearth of research using primary sources from widely spoken languages such as Mandarin and Hindi. The lack of primary source research in Chinese, Russian, and Hindi means that scholars may miss political nuances and shifts in the status quo because they are not using the target language. These nuances and shifts may include improving or worsening relations between ethnic groups, against the central government, or against a foreign state, for example.

Moreover, the grievances and rationales may only be best expressed in the local language. For example, in the Guatemalan human rights trial of Efrain Rios Montt, some of the concepts were misinterpreted:

> The Ixil interpreter regularly translated ch'ich' to a Spanish word whose meaning was less context dependent (for example, "at chalab' tzok'el kan, ili'b'aj kan naj, ta'n ch'ich'" interpreted as 'there are some of them who were left cut up, he left them hanging, with machetes'). However, in cases in which either a plane or a helicopter could be a possible referent for ch'ich', interpreters translated ch'ich' as instrumento, (instrument), a confusing translation for Spanish-speaking lawyers and judges. Lawyers often mistakenly assumed that "instrument" indicated that a speaker could not identify the aircraft in question or that the speaker was being purposefully evasive. Both prosecution and defense lawyers frequently requested clarification of this use. (García 2019, 242)

For IR scholars quantifying human rights abuses, researching peace and reconciliation processes, and quantifying civil war dynamics, these confusing distinctions between the varying interpretations of "ch'ich" may be important. Scholars of IR may not be fluent in Ixil, but this gap can be bridged through partnerships across disciplines, thus enriching the depth and quality of scholarship.

Political discourse data are not missing at random, and corpus collection and production suffer from similar challenges to observational data collection in IR. Either some corpora do not exist in written form, or their written form is presently unsupported by computational analytical tools, or the documents are well guarded and nearly impossible to access. The DGP for political corpora suffers from reporting bias similar to the way that observational data are missing from databases such as the World Bank Human Development Index. Countries may not report annual data due to ongoing conflict or active violence, to technical or infrastructure impediments, or to group-level bias whereby some portions of the country are not surveyed or counted adequately, such as minority groups.

Records-keeping may be sporadic, underfunded, and collected and stored in user-unfriendly formats or suspended due to lack of personnel or active conflict. In the case of civil conflicts, records may be looted or destroyed. Data archiving are often a low priority, and they may not engage in best practices for document archiving, such as formatting documents in consistent structures, which I discuss in the next section. For scholars, accessing these documents may require in-country field research and extensive preprocessing to clean the data in preparation for analysis. In politically unstable countries, archival data may be deliberately destroyed by outgoing regimes as a safeguard against future political liability or prosecution. Problems such as these manifested in countries such as Rwanda in 1994, Iraq in 2003, Bosnia between 1992 and 1995, and Guatemala after the thirty-year civil war (1960–1996).

Addressing the Missing Data/Omitted Variable Bias

Missing data influence the entire "value chain" of corpus production that affect research outcomes and the conclusions we draw about things in the world. The original documents from non-English and non-Western sources often

originate in countries lacking either the technology or incentives—or both—to curate politically relevant documents. Intrastate conflict within countries can hamper, delay, or even deliberately prevent document archiving. Relevant political documents may fall victim to actors' who use pillaging as a tactic of war.

Alternately, actors may destroy their own documents for fear of information being captured or used against them or their confederates. Active conflict interrupts the process of records-keeping, as other conflict-related activities are prioritized. Similarly, developing countries not experiencing conflict may also fall short in implementing the best practices for document archiving due to lack of technological capacity, improper storage, and exposure to degrading elements or natural disasters (Mnjama 2005, 2010; Dube 2011).

Low-Resource and High-Resource Languages

An example of an information deficit is the underrepresentation of "low-resource languages" in text-as-data political inquiry; these are rare linguistic groups that can be conceptualized as "low density, less commonly taught, underresourced, less resourced, low resourced, endangered, and vulnerable language" (Cieri et al. 2016). On the other hand, high-resource languages (HRL) are those for which computational translation programs and parallel corpora exist, such as the official languages of the United Nations (Eisele and Chen 2000). Low Resource Languages (LRLs) include more obscure and dwindling linguistic groups representing politically tenuous populations such as the Rohingya in Burma, the Acehnese in Indonesia, and the Uyghurs in China. Many of the quantities of interest—such as political processes within authoritarian regimes, marginalization of minority populations, issues of environment and climate-related human insecurity, and civil conflict—transpire in countries with LRLs. In the absence of investigating the political phenomena in the languages of a majority of the world's population, political scientists must make assumptions about what people in these environments talk about and believe. This represents a large foreign policy blind spot as well as an opportunity for ample future scholarship (Colgan 2019b).

Large swaths of world politics remain uncovered by computational corpus linguistics, including Africa, Latin America, Asia, and Southeast Asia. Similar to the identification of minorities at risk (MAR), United Nations Educational, Scientific and Cultural Organization (UNESCO) identifies languages at risk, with more than 1.5 million people speaking threatened or vulnerable languages, that is, those with declining rates of intergenerational transmission where children do speak the language, but it is restricted to certain venues (such as home) (Brenzinger et al. 2003; Lewis and Simons 2010). While the universe of cases for corpus selection is large, most languages remain vastly underrepresented in the data even though their countries of origin serve as the motivation for much of IR research. These countries are the source of emerging power politics, such as the BRICS (Brazil, Russia, India, China, and South Africa) countries (Thies and Nieman 2017) as well as countries where multiple types of security threats originate, including human security (e.g., health, migration, and conflict-related displacement); country-level instability, MAR, civil conflict, and origins of terrorism; and intrastate conflict.

Corpus Generation

Countries in the developing world face obstacles in the production of political corpora, including data collection, formatting, storage, and sharing. The digital divide between wealthy and poor countries affects document preparation; countries experiencing conflict face additional difficulties. While scholars in developed countries are endowed with first-rate technology, such as computers, reliable electricity and Internet, and data-processing software, scholars in developing countries face multiple, deep challenges. Data generated and collected in the developing world are often not gathered with academic use in mind. For example, it is common to have typewritten documents with handwritten notes in the margins, scanned at an angle and saved as an image to a portable document format (.pdf) file, making the text more difficult to extract using optical character recognition (OCR).

Behind Iron Curtains

Data in opaque political environments, such as authoritarian regimes and extremist groups, are also often difficult to obtain. In authoritarian regimes, the leadership thrives on maintaining a culture of secrecy, relying on ministries of information for intelligence gathering. For example, in the case of Iraq under Saddam Hussein, the Ministry of Information kept extensive documentation of meetings and records, but these records were private and highly guarded and not recovered until after the 2003 conflict. Similarly, some countries may collect and archive data efficiently but may be unwilling to share this information with a broad audience. Documents from Al Qaeda in Afghanistan were retrieved after the military intervention in 2001. These included internal communications as well as military strategy, private information that individuals in each case were unwilling to share or make public. These documents were ultimately processed, translated, and archived at the now-defunct Conflict Records Research Center at the National Defense University in Washington, DC, but accessing the full corpus required on-site access.

To this point, in 2016 the National Academy of Sciences and Intelligence Community began a decadal survey to assess research priorities and potential synergies between academics in the social and behavioral sciences, and government agencies (Windsor 2017a). Of note was the suggestion to liberate declassified data and facilitate scholarly research on timely and relevant issues, including multilingual corpora. Open-source intelligence initiatives such as the Central Intelligence Agency's Foreign Broadcast Information Service and World News Connection, and the BBC's Monitoring Service aggregated news reports from foreign media and meticulously translated them using human sources before automated computer translation was available. While the end result was an English-language corpus, the breadth of information these efforts generated could be used to study and model international processes from many multilingual, multicultural perspectives.

Knowing Where to Look: The Politics and Language Pipeline

What deep structural factors might influence whether text analysis of political corpora is undertaken in English versus non-English languages? Given that most political science scholarship originates in OECD countries, with the majority of doctoral programs located in North America (Ishiyama and Breuning 2006), it is somewhat unsurprising that English-language sources account for the bulk of political corpora. One factor that may contribute to the lack of linguistic diversity in political science research is the dwindling multilingual pipeline in the US primary, secondary, and postsecondary educational systems.

In most US public schools, foreign language instruction does not begin until high school, which poses problems for attaining proficiency or fluency in second languages (Pufahl and Rhodes 2011; Friedman 2015). Fewer high-school students studying foreign languages lead to fewer college students studying foreign languages; in turn, fewer doctoral students in political science are proficient in non-English languages. The Modern Language Association reports that enrollment in languages other than English declined by 9.2% between 2013 and 2016, the second largest drop since the census began (Lusin 2012; Looney and Lusin 2018). For this article, I conducted a survey of one hundred and thirty doctoral programs in Political Science in the United States from the list of American Political Science Association resources for students to determine the status of foreign language requirements across departments. With sixty respondents, the response rate was 46 percent. Of these sixty programs, only nine (15 percent) include a foreign language component for successful completion of the doctoral degree requirements.[1]

Corpus selection bias is a manifestation of other types of biases: studies of implicit bias show that people demonstrate subconscious preferences for cues that they like and dislike (Blanton et al. 2009; Jost et al. 2009; Tetlock and Mitchell 2009), and confirmation bias suggests that people seek out information that affirms preexisting beliefs (Nickerson 1998), including the selection of documents and texts (Lustick 1996). The set of functional polyglot researchers from PhD-granting US universities in the field of political science, much less in the smaller text-as-data subfield, is incredibly small. The limited mandatory second-language exposure for IR scholars influences textual data selection. Thus, scholars with little non-English language exposure may already be biased against looking for documents in low-resource languages.

Both the implicit and the explicit biases point to a troubling issue: given the lack of peer-reviewed scholarship originating in developing countries and about LRL communities, how can we be certain that the theories and hypotheses generated with linguistically biased corpora accurately represent real relationships and phenomena in the world? Clearly, it is unreasonable to expect that all IR scholars should gain proficiency in multiple languages; a more balanced approach, and one for which there is growing support in the IR community, is forging collaborations with scholars in non-Western countries who are either native speakers or proficient in the languages of interest. In recent years, IR scholars have partnered with colleagues in Latin America, Eastern Europe, and Asia to host conferences outside of North America and Western Europe to address the geographic bias that inhibits broader participation from scholars from the global south. These research partnerships can span not only borders but also academic disciplines.

Linguistic Analysis Pipeline

The linguistic pipeline for software and tools to analyze non-English corpora is underdeveloped. Non-English and LRL documents are largely unsupported by standard software translation or OCR programs used to convert.pdf files to ones supported by analytical tools, such as spreadsheets or text files. OCR programs vary in their ability and accuracy to process information embedded in .pdf files (Smith 2007; Heliński, Kmieciak, and Parkoła 2012). The process of extracting usable information from old, deteriorated, or irregularly formatted documents can be time-consuming and labor-intensive. Furthermore, once the information is extracted, the documents may need to be translated. Professional translation services are not only time-consuming but also expensive. Further, while recent research on automated translation programs has demonstrated their reliability, questions still exist about the issue of syntactic, lexical, and semantic fidelity to the author's original intent. Most computational linguistics programs—especially those examining syntax—are available for English-language corpora only.

Document Processing

For non-English and non-Western documents that *do* exist in usable formats for researchers to access, substantial preprocessing steps are required before they can be analyzed. Figure 2 illustrates this point: the first document is a typewritten and scanned transcript of the RTLM (Radio Television Libre des Milles Collines) broadcast during the Rwandan Genocide, the second is a scanned copy of the Malian peace agreement from 2015, and the third is a handwritten account of Guatemalan police records. Documents in this format present significant problems for open-source OCR programs such as Tesseract, as they are "noisy," meaning they have substantial amounts of non-essential information such as borders, stamps, and handwritten notes that clutter the textual signal (Heliński, Kmieciak, and Parkoła 2012). Cleaning text-as-data corpora is labor-intensive and requires human judgment throughout the preprocessing phase as documents are prepared for analysis; not all text-processing steps can be automated.

Preprocessing refers to the steps that researchers must take to prepare a corpus for analysis. This may include the process of OCR, scanning, translation, document conversion, and "cleaning" documents using regular expressions to omit erroneous characters that may interrupt or impede the computational software used to analyze corpora.

The text itself may also be blurry or smudged, and the document formats change over time making automated processing more challenging. This means that the OCR process requires labor-intensive human supervision to make decisions about accommodating document quirks. For example, documents from pre-Internet eras were typewritten or handwritten, or both, and have been subsequently subjected to heat, water, mold, fire, or other damage. Furthermore, they are often scanned askew and must be manually adjusted. Researchers also face computational impediments to accessing LRL corpora; the human language technology community notes that for low-density languages, few online resources are available for processing, translation, and analysis (Hogan 1999; Megerdoomian and Parvaz 2008). Some text-as-data analysis does not require substantial preprocessing and can accommodate more noisy corpora, such as topic modeling.

Substantial computational barriers remain for low-resource languages, especially those using non-Latin systems of writing. Given that most computational programs analyze English-language corpora, one solution is to first translate non-English corpora into English and then analyze them with computational linguistics tools. However, this workflow may risk obscuring the intent conveyed in the original language, as some details can get lost in translation; researchers

[1] These include Baylor University, Boston College, the University of Pennsylvania, Northern Illinois University, Johns Hopkins (Political Science), University of Virginia, The New School for Social Research, Northern Arizona University, and Florida International University. For a full list of respondents, see table 2 in the appendix.

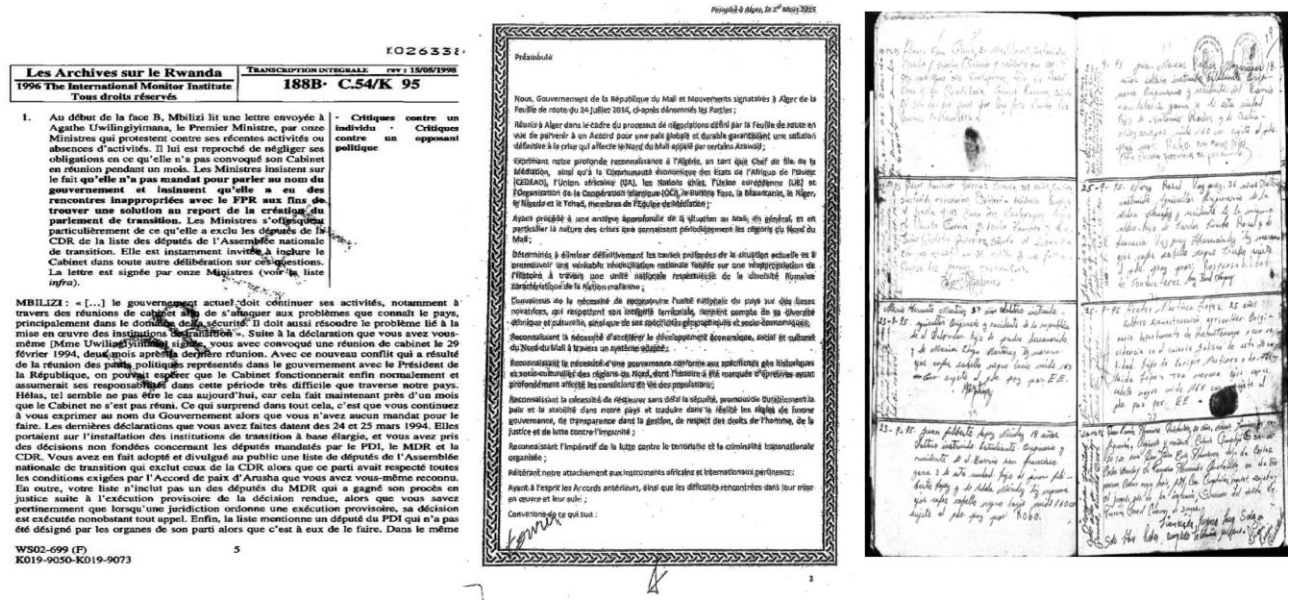

Figure 2. (Left to right) RTLM document from Rwanda, Malian peace agreement, Guatemalan historical archive

should confirm that the translated documents maintain fidelity to the original documents.

In one such analysis of dictionary-based translation using a United Nations parallel corpus, Google Translate proved to be fairly reliable, with the exception of character-based languages such as Chinese and Arabic, which showed the largest effect sizes between the original and target languages (Eisele and Chen 2000; Windsor, Cupit, and Windsor 2019).[2] Other non-Latin languages that use Cyrillic script (e.g., Slavic languages like Russian and Ukranian), Hebrew, Turkish, and Brahmic (common in India) are also challenging for computational translation and analytic programs as they introduce additional preprocessing steps prior to analysis. These preprocessing steps often require specialized software and extensive human-in-the-loop intervention in laborious and non-automatable workflows. Drawing on suggestions from the previous section, these problems with information deficiencies and document processing can be addressed by using language-agnostic methods as well as collaboration with scholars from underrepresented countries and linguistic groups.

To summarize, even once linguistically rare corpora have been located, these documents still often require substantial pre-processing due to their idiosyncratic and irregular nature. Preprocessing documents prior to analysis is both necessary and time-consuming. It is incumbent upon scholars working on "big-text-as-data" to demonstrate the volume of preprocessing steps and coding decisions that inform their workflow by documenting their workflow and annotating publications with the steps and decisions that produced the outcome corpus. This is especially important not only for transparency and reproducibility (Munafò et al. 2017) but also to address misperceptions about quantitative text research about the facility of batch downloading ready-made datasets and automating coding decisions with impersonal algorithms. To the contrary, as Nelson suggests, computers should expedite the investigative process and facilitate humans' analytical strengths, so researchers focus their time and efforts most efficiently on interpretation rather than classification (Nelson 2017, 2018).

[2] A parallel corpus is a collection of documents that are translated into one or more other languages than the original across meaning units, usually sentences.

Consequences of Linguistic Biases

Fundamental Research in Political Science

Given the recession of democracies and the rise of authoritarian populism in the world, and the growth of violent extremist organizations, IR scholars should pursue research using computational text-analysis methods to answer questions about fundamental areas of IR scholarship, such as the causes and consequences of war and peace among and within states, democratization and democratic backsliding, and leadership and regime changes. Yet, if we seek answers to political questions in only the convenient languages with the most accessible corpora requiring little preprocessing, we miss the richness, nuances, wisdom, and perspectives that other languages and cultures can offer about the political quantities of interest that drive our curiosity about the world.

One consequence of English-language bias is that we miss the opportunity to fill in the gaps for other indicators that suffer from similar systematic missing data. IR scholars are tasked with investigating political questions related to democratization, major geopolitical shifts, human rights practices, international and civil war onset, duration, termination, trade, and security between states. Investigations of these phenomena have largely been observational, drawing from datasets that represent a majority of countries in the world such as the Polity IV Project, Correlates of War, Cross-National Time Series, the World Bank, Archigos, Freedom House, and the Cingranelli–Richards Human Rights Data (Cingranelli 2006; Marshall, Jaggers, and Gurr 2006; Goemans, Gleditsch, and Chiozza 2009; Sarkees 2010; Banks 2011; Freedom House 2014).

It is important to note that these commonly used observational datasets do not represent the universe of cases in the international system and that data are missing

nonrandomly for countries that may be lacking bureaucratic infrastructure or a functioning central government to collect and curate this data. For example, recent work has examined bias in event data generation (Weidmann 2016) and terrorist activity reporting (Drakos and Gofas 2006), proposing methodological solutions to check for estimation sensitivity to missing data. Imputation, matching, and classifying have emerged as computational workarounds for filling in the gaps in datasets as well (Jackman 2000; Honaker and King 2010; Si and Reiter 2013). Because data in developing, non-Western, and non-democratic societies is often systematically missing—such as economic, health, education, and other indicators—language data can help to fill in the knowledge gaps. However, this cannot happen using English-only sources, since the information may only be available in the local language.

Event Data and Forecasting Accuracy

At its core, event data are text-as-data and face many of these same challenges. Researchers working on event data generation continue to grapple with the problem of syntax in assigning roles, relationships, and activities between international actors (Schrodt, Beieler, and Idris 2014; Norris 2016), and the problems are even more complex with multilingual corpora. News organizations with contracting budgets for international reporting have pulled journalists from foreign assignments and shuttered overseas bureaus; as a consequence, international news coverage has declined (Kaphle 2015; Gray 2017). Related, when journalists are sent to report on international matters, they often do so in the context of active conflict; in recent years, the Committee to Protect Journalists has recorded an increase in the number of journalists killed while on assignment. As a result, news coverage of smaller or less-populated countries has decreased, with implications not only for political corpus collection but also for event data analysis.

The NLP (natural language process) parsing of event data works best on newswire sources such as the Associated Press that use formulaic, straightforward phrasing rather than local news sources that may provide more nuanced—albeit with more complicated syntax, grammar, and named entities—information about politically important events. Thus, as the upstream sources of data winnows, the dearth of news is felt downstream: we know less about "on the ground" conditions that can lead to larger scale political disturbances, decreasing the accuracy of forecasts and predictions.

The critical connection between political MAR and the languages they speak is that what researchers often know about these groups comes from Western sources, filtered through cultural biases and interpretations rather than directly from the group source. For example, even widely used event data sources "missed" the onset of the Arab Spring uprisings due in part to its reliance on English-language newswire services, which may have been detected if local news sources were consulted (Ward et al. 2013; Wang et al. 2016). Nam (2006) also provides substantial evidence for the effectiveness of using local sources, giving examples from South Korea and Burma. Another potential oversight is in the coding of political events: the taxonomy of significant political events was created by Western scholars, using English-language sources. So, for example, even though cattle rustling and raiding is a significant factor in conflict escalation (Witsenburg and Adano 2009; Butler and Gates 2012), this activity is not included as an event subcategory.

Misunderstanding Group Mobilization

Scholars of social movements often rely on social media to measure popular sentiment, unrest, and mobilization (Kavanaugh et al. 2011). Microblogging platforms, such as Twitter, can provide real-time information about unfolding contentious politics. However, for countries or areas with a significant ex-patriate or diaspora population, the signals become muddled as it is not always obvious from where the social media posts are originating. This may lead to an over-estimation of actual, local, human mobilization, or a mischaracterization of the strength of participants. For example, the pathbreaking work by Chenoweth and Belgioioso found that social movements display Physics characteristics—mass and momentum (Chenoweth and Belgioioso 2019). Wouldn't we want to know if the results hold across languages, actors, and social media bounding boxes?

Future Directions

This article is intended to seed a conversation about computational text as data for IR research. I have discussed the challenges facing computational text-as-data methods for IR research, including sourcing linguistically diverse corpora, preparing and preprocessing documents, and generating corpora to foster theoretically driven hypotheses to test about how international actors use language strategically. At present, the text analytics for IR research is limited in scope and reflects an English language and Western corpus bias.

While English language sources have expedited the process of computational advancements and methodological discoveries, they have come at the expense of uncovering new insights drawn from research using low-resource languages that represent themes and regions of political interest in the world. Some information deficiencies are difficult to remedy, as sources are unavailable in irretrievable ways such as missing data due to irregular records-keeping during conflicts, and inaccessibility post-conflict. Document-processing problems also contribute to the lack of linguistic diversity in political corpora.

Three approaches will help internationalize the study of IR in the field of computational text analysis. First, we must collaborate better across borders and partner with local scholars who speak the target language. Local scholars will likely have insight into corpora that will deepen our understanding of international political processes, and the cross-national scholarly collaboration will help to diversify the field of IR (Breuning et al. 2018a). Second, we must utilize the existing methodologies that are language-agnostic, such as topic modeling, to analyze documents in non-English languages. As Nelson suggests, computers are useful for sorting and humans are good at interpreting (Nelson 2017); we should leverage this reciprocal relationship to bridge quantitative and qualitative IR scholarship using text-analytic methods. Finally, we should forge interdisciplinary relationships, especially between social scientists and computer scientists who are doing pathbreaking research using Bidirectional Encoder Representations from Transformers (BERT) and its multilingual cousins, TBERT and mBERT, which handle multilingual corpora (Devlin et al. 2019; Chau, Lin, and Smith 2020; Gonen et al. 2020; Peinelt, Nguyen, and Liakata 2020). The former are trained in the methods of social science inquiry and the latter in developing applied technologies and workflows to facilitate research on sociopolitical phenomena.

We not only need more representative international corpora but more representative scholars to ensure fidelity

to the intended meaning of the documents. The mission of the Global South Academic Network (Reinhardt 2018) is to foster collaboration between scholars across the academic divides; this association can help facilitate partnerships from low-resource languages. Alongside more opportunities for foreign language study and fieldwork experiences, these interdisciplinary and multinational collaborations should help minimize the risks of misinterpreting output of non-English and multilingual corpora. These collaborations are essential and valuable, as language shapes the way people see and experience the world around them and concepts can become lost in translation (Boroditsky 2011; Cibelli et al. 2016).

Other strategies for remedying selection bias in IR text-as-data corpora include participating in conference working groups, aggregating IR corpora in a centralized repository, encouraging the development of text-as-data courses and workshops, and incentivizing this line of scholarship through special issues in journals. For example, the recent Preconference on Politics and Computational Social Science at Northeastern University featured a large selection of research using text-as-data methods to explore questions related to political protest (Eubank 2018), deliberative democracy (Chen 2018), online dissent in absolute monarchies (Siegel and Pan 2018), and threats of violence toward civilians in China (Carter and Carter 2018).

Specializations within IR, such as international political economy, international and sub-national conflict, peace building, treaties and alliances, and international cooperation more broadly all lend themselves to computational text analysis using presently available data. International organizations, such as the United Nations and its branches, and regional institutions, such as MERCOSUR or the Organization for African Unity, can facilitate a common platform for analyzing political language and begin to address the gap in corpus selection by committing to providing documents such as resolutions, reports, and treaties on publicly available and widely accessible repositories, ideally in multiple formats including plain text.

Text selection bias is a problem fundamentally because it poses restrictions on the types of research questions that IR scholars in particular are able to ask and answer. It filters the experiences of ethnically, linguistically, and politically diverse people in the world through a Western lens. Researchers must make decisions about distinguishing between important and erroneous information that may or may not be critically important in the target language. Through interdisciplinary collaborations, scholars can begin to address the deficits of having depended on English language to learn about the world. As a discipline, we must be willing to revisit our assumptions about how the world works by seeing it through the lens of different languages.

References

Abrahms, Max, Nicholas Beauchamp, and Joseph Mroszczyk. 2017. "What Terrorist Leaders Want: A Content Analysis of Terrorist Propaganda Videos." *Studies in Conflict & Terrorism* 40 (11): 899–916.

Aharony, Noa. 2012. "Twitter Use by Three Political Leaders: An Exploratory Analysis." *Online Information Review* 36 (4): 587–603.

Alduy, Cécile, and Stéphane Wahnich. 2015. "Marine Le Pen Prise Aux Mots." In *Décryptage Du Nouveau Discours Frontiste.*, 94–98. Paris: Seuil.

Banks, A. 2011. "Cross-National Times Series Dataset." *State University of New York*. Accessed November 15, 2021. http://www. databanksinternational. com/53.html.

Baturo, Alexander, Niheer Dasandi, and Slava J. Mikhaylov. 2017. "Understanding State Preferences with Text as Data: Introducing the UN General Debate Corpus." *Research & Politics* 4 (2): 2053168017712821.

Baturo, Alexander, and Slava Mikhaylov. 2014. "Reading the Tea Leaves: Medvedev's Presidency through Political Rhetoric of Federal and Sub-National Actors." *Europe-Asia Studies* 66 (6): 969–92.

Beauchamp, Nicholas. 2017. "Predicting and Interpolating State-Level Polls Using Twitter Textual Data." *American Journal of Political Science* 61 (2): 490–503.

Birch, Sarah, and David Muchlinski. 2020. "The Dataset of Countries at Risk of Electoral Violence." *Terrorism and Political Violence* 32 (2): 217–36.

Blanton, Hart, James Jaccard, Jonathan Klick, Barbara Mellers, Gregory Mitchell, and Philip E. Tetlock. 2009. "Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT." *Journal of Applied Psychology* 94 (3): 567–82.

Boroditsky, Lera. 2011. "How Language Shapes Thought." *Scientific American* 304 (2): 62–65.

Brands, Hal, and David Palkki. 2012. "'Conspiring Bastards': Saddam Hussein's Strategic View of the United States." *Diplomatic History* 36 (3): 625–59.

Brenzinger, Matthias, Akira Yamamoto, Noriko Aikawa, Dmitri Koundiouba, Anahit Minasyan, Arienne Dwyer, and Colette Grinevald et al. 2003. "Language Vitality and Endangerment." *UNESCO Intangible Cultural Unit, Safeguarding Endangered Languages, Paris*. Accessed July 1, 2010. Accessed November 15, 2021. http://www.Unesco. Org/Culture/Ich/Doc/Src/00120-En.Pdf.

Breuning, Marijke,Feinberg, Ayal, Benjamin Isaak Gross, Melissa Martinez, Ramesh Sharma, and John Ishiyama. 2018a. "How International Is Political Science? Patterns of Submission and Publication in the American Political Science Review." *PS: Political Science & Politics*. 51 (4): 789–98.

———. 2018b. "Clearing the Pipeline? Gender and the Review Process at the American Political Science Review." *PS: Political Science & Politics* 51 (3): 629–34.

Brummer, Klaus, Michael D. Young, Özgur Özdamar, Sercan Canbolat, Consuelo Thiers, Christian Rabini, Katharina Dimmroth, Mischa Hansel, and Ameneh Mehvar. 2020. "Coding in Tongues: Developing Non-English Coding Schemes for Leadership Profiling." *International Studies Review* 22 (4): 1039–67.

Butler, Christopher K., and Scott Gates. 2012. "African Range Wars: Climate, Conflict, and Property Rights." *Journal of Peace Research* 49 (1): 23–34.

Callison-Burch, Chris. 2009. "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, 286–95. Stroudsburg, PA: Association for Computational Linguistics.

Canbolat, Sercan. 2021. "Deciphering Deadly Minds in Their Native Language: The Operational Codes and Formation Patterns of Militant Organizations in the Middle East and North Africa." In *Operational Code Analysis and Foreign Policy Roles*, edited by Mark Schafer and Stephen G. Walker, 69–92. New York: Routledge.

Carter, Brett L., and Erin Baggott Carter. 2018. "When Autocracies Threaten Citizens with Violence: Evidence from China." *British Journal of Political Science*. 1–26.

Carter, Erin Baggott, and Brett L. Carter. 2020. "Focal Moments and Protests in Autocracies: How Pro-Democracy Anniversaries Shape Dissent in China." *Journal of Conflict Resolution* 64 (10): 1796–1827.

Chau, Ethan C., Lucy H. Lin, and Noah A. Smith. 2020. "Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank." *ArXiv Preprint ArXiv:2009.14124*.

Chen, Kaiping. 2018. "Can the Mass Public Deliberate in a Semi-Authoritarian Setting? Examining Deliberative Reasoning in Macau's Deliberative Poll." In *ICA, Washington, DC*.

Chenoweth, Erica, and Margherita Belgioioso. 2019. "The Physics of Dissent and the Effects of Movement Momentum." *Nature Human Behaviour*. 3: 1088–1095.

Chiang, David. 2007. "Hierarchical Phrase-Based Translation." *Computational Linguistics* 33 (2): 201–28.

Chiovaro, Megan, Leah C. Windsor, and Alexandra Paxton. 2021. "Vector Autoregression, Cross-Correlation, and Cross-Recurrence Quantification Analysis: A Case Study in Social Cohesion and Collective Action." Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 43.

Chiovaro, Megan, Leah C. Windsor, Alistair Windsor, and Alexandra Paxton. 2021. "Online Social Cohesion Reflects Real-World Group Action in Syria during the Arab Spring." *PLoS One* 16 (7): e0254087.

Cibelli, Emily, Yang Xu, Joseph L. Austerweil, Thomas L. Griffiths, and Terry Regier. 2016. "The Sapir–Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color." *PLoS One* 11 (7): e0158725.

Cieri, Christopher, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. "Selection Criteria for Low Resource Language Programs." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4543–49. Portorož: European Language Resources Association (ELRA).

Cingranelli, D.L. 2006. "The Cingranelli–Richards (CIRI) Human Rights Dataset." Retrieved September 15.

Cohen, Alex S., Kyle S. Minor, Lauren E. Baillie, and Amanda M. Dahir. 2008. "Clarifying the Linguistic Signature: Measuring Personality from Natural Speech." *Journal of Personality Assessment* 90 (6): 559–63.

Colgan, Jeff D. 2019a. "American Bias in Global Security Studies Data." *Journal of Global Security Studies* 4 (3): 358–71.

———. 2019b. "American Perspectives and Blind Spots on World Politics." *Journal of Global Security Studies* 4 (3): 300–309.

Cuhadar, Esra, Juliet Kaarbo, Baris Kesgin, and Binnur Ozkececi-Taner. 2017. "Personality or Role? Comparisons of Turkish Leaders across Different Institutional Positions." *Political Psychology* 38 (1): 39–54.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805 [Cs]*, May. Accessed November 15, 2021. http://arxiv.org/abs/1810.04805.

Dowd, Caitriona, Patricia Justino, Roudabeh Kishi, and Gauthier Marchais. 2020. "Comparing 'New' and 'Old' Media for Violence Monitoring and Crisis Response: Evidence from Kenya." *Research & Politics* 7 (3): 2053168020937592.

Dowell, Nia M., Leah C. Windsor, and Arthur C. Graesser. 2015. "Computational Linguistics Analysis of Leaders during Crises in Authoritarian Regimes." *Dynamics of Asymmetric Conflict* 9 (1–3): 1–12.

Drakos, Konstantinos, and Andreas Gofas. 2006. "The Devil You Know but Are Afraid to Face: Underreporting Bias and Its Distorting Effects on the Study of Terrorism." *Journal of Conflict Resolution* 50 (5): 714–35.

Driscoll, Jesse, and Zachary C. Steinert-Threlkeld. 2020. "Social Media and Russian Territorial Irredentism: Some Facts and a Conjecture." *Post-Soviet Affairs* 36 (2): 101–21.

Dube, Thembani. 2011. "Archival Legislation and the Challenge of Managing Archives in Zimbabwe." *ESARBICA Journal* 30: 279.

Dyson, Stephen Benedict. 2001. "J." *Policy Sciences* 34 (3/4): 329–46.

———. 2006. "Personality and Foreign Policy: Tony Blair's Iraq Decisions." *Foreign Policy Analysis* 2 (3): 289–306.

Dyson, Stephen Benedict, and Alexandra L. Raleigh. 2014. "Public and Private Beliefs of Political Leaders: Saddam Hussein in Front of a Crowd and behind Closed Doors." *Research & Politics* 1 (1): 2053168014537808.

Dyson, Stephen Benedict, and Thomas Preston. 2006. "Individual Characteristics of Political Leaders and the Use of Analogy in Foreign Policy Decision Making." *Political Psychology* 27 (2): 265–88.

Eisele, Andreas, and Yu Chen. 2000. "MultiUN: A Multilingual Corpus from United Nation Documents." Accessed November 15, 2021. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.682.4012&rep=rep1&type=pdf.

Eubank, Nick. 2018. "Peers and Protest." Politics and Computational Social Science Conference. *Northeastern University*.

Freedom House. 2014. "Freedom in the World. Selected Data from Freedom House's Annual Global Survey of Political Rights and Civil Liberties."

Friedman, Amelia. 2015. "America's Lacking Language Skills." *The Atlantic*. May 10, 2015. Accessed November 15, 2021. https://www.theatlantic.com/education/archive/2015/05/filling-americas-language-education-potholes/392876/.

García, María Luz. 2019. "Language, Culture, and Justice: Ixil Mayan Verbal Art in the 2013 Genocide Trial of José Efraín Ríos Montt in Guatemala." *Journal of Linguistic Anthropology* 29 (2): 239–48.

Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis* 2: 131–50.

Goebel, Christian, and Christoph Steinhardt. 2019. "Better Coverage, Less Bias: Using Social Media to Measure Protest in Authoritarian Regimes." *Department of East Asian Studies, University of Vienna*.

Goemans, Henk E., Kristian Skrede Gleditsch, and Giacomo Chiozza. 2009. "Introducing Archigos: A Dataset of Political Leaders." *Journal of Peace Research* 46 (2): 269–83.

Goh, Evelyn. 2019. "US Dominance and American Bias in International Relations Scholarship: A View from the Outside." *Journal of Global Security Studies* 4 (3): 402–10.

Gonen, Hila, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. "It's Not Greek to MBERT: Inducing Word-Level Translations from Multilingual BERT." ArXiv Preprint ArXiv:2010.08275.

Gray, Rosie. 2017. "The Wall Street Journal's Global Retrenchment." *The Atlantic*. February 16, 2017. Accessed November 15, 2021. https://www.theatlantic.com/politics/archive/2017/02/wall-street-journal-retrenches-around-the-world/516915/.

Grimmer, Justin. 2009. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18 (1): 1–35.

Gupta, Aditi, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. "TweetCred: Real-Time Credibility Assessment of Content on Twitter." In *Social Informatics*, edited by Luca Maria Aiello and Daniel McFarland, 228–43. Cham: Springer International Publishing.

Gurr, Ted Robert. 1995. *Minorities at Risk: A Global View of Ethnopolitical Conflicts*. Arlington, VA: United States Institute of Peace Press.

Hancock, Jeffrey, David Beaver, Cindy Chung, Joey Frazee, James Pennebaker, Art Graesser, and Zhiqiang Cai. 2010. "Social Language Processing: A Framework for Analyzing the Communication of Terrorists and Authoritarian Regimes." *Behavioral Sciences of Terrorism and Political Aggression* 2 (2): 108–32.

Heliński, Marcin, Miłosz Kmieciak, and Tomasz Parkoła. 2012."Report on the Comparison of Tesseract and ABBYY FineReader OCR Engines." Accessed November 15, 2021. http://www.digitisation.eu/fileadmin/Tool_Training_Materials/Abbyy/PSNC_Tesseract-FineReader-report.pdf.

Hermann, Margaret G. 1979. "Indicators of Stress in Policymakers during Foreign Policy Crises." *Political Psychology* 1 (1): 27–46.

———. 1980. "Explaining Foreign Policy Behavior Using the Personal Characteristics of Political Leaders." *International Studies Quarterly* 24 (1): 7–46.

———. 1987. "Foreign Policy Role Orientations and the Quality of Foreign Policy Decisions." In *Role Theory and Foreign Policy Analysis*, edited by Stephen Walker, 123–40. Durham, NC: Duke University Press.

Hermann, Margaret G., and Jerrold M. Post. 2003. *The Psychological Assessment of Political Leaders: With Profiles of Saddam Hussein and Bill Clinton*. Ann Arbor, MI: University of Michigan Press.

Hermann, Margaret G., and Thomas Preston. 1994. "Presidents, Advisers, and Foreign Policy: The Effect of Leadership Style on Executive Arrangements." *Political Psychology* 15 (1): 75–96.

Hermann, Margaret G., Thomas Preston, Baghat Korany, and Timothy M. Shaw. 2001. "Who Leads Matters: The Effects of Powerful Individuals." *International Studies Review* 3 (2): 83–131.

Hill, K.Q., and P.A. Hurley. 2002. "Symbolic Speeches in the US Senate and Their Representational Implications." *The Journal of Politics* 64 (1): 219–31.

Hinkle, Rachael K., and Michael J. Nelson. 2016. "The Transmission of Legal Precedent among State Supreme Courts in the Twenty-First Century." *State Politics & Policy Quarterly* 16 (4): 391–410.

Hogan, Christopher. 1999. "OCR for Minority Languages." Symposium on Document Image Understanding Technology, Annapolis, Maryland.

Honaker, James, and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54 (2): 561–81.

Ishiyama, John, and Marijke Breuning. 2006. "How International Are Undergraduate Political Science Programs at Liberal Arts and Sciences Colleges and Universities in the Midwest?" *PS: Political Science & Politics* 39 (2): 327–33.

Ito, Jun, Jing Song, Hiroyuki Toda, Yoshimasa Koike, and Satoshi Oyama. 2015. "Assessment of Tweet Credibility with LDA Features." In *Proceedings of the 24th International Conference on World Wide Web*, 953–58. New York: Association for Computing Machinery.

Jackman, Simon. 2000. "Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation." *Political Analysis* 8 (4): 307–32.

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): 319–45.

Jervis, Robert. 2013. "Do Leaders Matter and How Would We Know?" *Security Studies* 22 (2): 153–79.

Jones, Benjamin T., and Eleonora Mattiacci. 2017. "A Manifesto, in 140 Characters or Fewer: Social Media as a Tool of Rebel Diplomacy." *British Journal of Political Science* 49 (2): 739–61.

Jost, John T., Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, and Curtis D. Hardin. 2009. "The Existence of Implicit Bias Is beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies That No Manager Should Ignore." *Research in Organizational Behavior* 29: 39–69.

Jurka, Timothy P., Loren Collingwood, Amber Boydstun, Emiliano Grossman, and Wouter van Atteveldt. 2012. "RTextTools: Automatic Text Classification via Supervised Learning." *R Package Version* 1 (9). Accessed November 15, 2021. http://www.rtexttools.com/.

Kaarbo, Juliet. 1997. "Prime Minister Leadership Styles in Foreign Policy Decision-Making: A Framework for Research." *Political Psychology* 18 (3): 553–81.

Kaarbo, Juliet, and Margaret G. Hermann. 1998. "Leadership Styles of Prime Ministers: How Individual Differences Affect the Foreign Policymaking Process." *The Leadership Quarterly* 9 (3): 243–63.

Kaphle, Anup. 2015. "The Foreign Desk in Transition." *Columbia Journalism Review*. Accessed November 15, 2021. https://www.cjr.org/analysis/the_foreign_desk_in_transition.php.

Kavanaugh, A., S. Yang, S. Sheetz, L.T. Li, and E. Fox. 2011. "Microblogging in Crisis Situations: Mass Protests in Iran, Tunisia, Egypt." Workshop on Transnational Human–Computer Interaction, CHI. Accessed November 15, 2021. http://www.princeton.edu/~jvertesi/TransnationalHCI/Participants_files/Kavanaugh.pdf.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2): 326–43.

King, Gary, and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57 (3): 617–42.

———. 2014. "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation." *Science* 345 (6199): 1251722.

Kowert, Paul A., and Margaret G. Hermann. 1997. "Who Takes Risks? Daring and Caution in Foreign Policy Making." *The Journal of Conflict Resolution* 41 (5): 611–37.

Kreutz, Joakim, and Mihai Croicu. 2014. "Talking about Revolutions: Estimating the Chinese View on Protest Politics from Discrepancies in Xinhua News Reporting in Chinese and English." The Advantages and Disadvantages of Media-Sourced Data on Conflict. *Toronto, Canada.*

Kydd, Andrew H., and Roseanne W. McManus. 2017. "Threats and Assurances in Crisis Bargaining." *Journal of Conflict Resolution* 61 (2): 325–48.

Labzina, Elena, and Mark Nieman. 2017. "State-Controlled Media and Foreign Policy: Analyzing Russian-Language News." *European Political Science Association, Milan, Italy.*

Lehmann, Pola, Theres Matthieß, Nicolas Merz, Sven Regel, and Annika Werner. 2018. "Manifesto Corpus." *WZB Berlin Social Science Center.*

Lewis, M. Paul, and Gary F. Simons. 2010. "Assessing Endangerment: Expanding Fishman's GIDS." *Revue Roumaine de Linguistique* 55 (2): 103–20.

Li, Haiying, M. Conley, Z. Cai, A.C. Graesser, and Y. Duan. Under review. "Formality and Charismatic Leaders: A Study on Chinese Collective Leadership of the Communist Party of China." *The Leadership Quarterly.*

Looney, Dennis, and Natalia Lusin. 2018. "Enrollments in Languages Other Than English in United States Institutions of Higher Education, Summer 2016 and Fall 2016: Preliminary Report." *Modern Language Association.* Accessed November 15, 2021. https://www.mla.org/content/download/83540/2197676/2016-Enrollments-Short-Report.pdf.

Lorenzini, Jasmine, Peter Makarov, Hanspeter Kriesi, and Bruno Wueest. 2016. "Towards a Dataset of Automatically Coded Protest Events from English-Language Newswire Documents." Amsterdam Text Analysis Conference.

Love, Gregory, and Leah Windsor. 2017. "Populism and Popular Support: Vertical Accountability, Exogenous Events, and Leader Discourse in Venezuela." *Political Research Quarterly* 71 (3): 532–45.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16 (4): 356–71.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–77.

Lusin, Natalia. 2012. "The MLA Survey of Postsecondary Entrance and Degree Requirements for Languages Other Than English, 2009–10." Modern Language Association. *ERIC.*

Lustick, Ian S. 1996. "History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias." *American Political Science Review* 90 (3): 605–18.

Marshall, M.G., K. Jaggers, and T.R. Gurr. 2006. "'Polity IV Project: Political Regime Characteristics and Transitions, 1800–2009'." Accessed November 15, 2021. http://www. systemicpeace.org/polity/polity4.htm.

McAdam, Doug. 1986. "Recruitment to High-Risk Activism: The Case of Freedom Summer." *American Journal of Sociology* 92 (1): 64–90.

———. 1989. "The Biographical Consequences of Activism." *American Sociological Review* 54 (5): 744–60.

McManus, Roseanne W. 2014. "Fighting Words: The Effectiveness of Statements of Resolve in International Conflict." *Journal of Peace Research* 51 (6): 726–40.

———. 2017. "The Impact of Context on the Ability of Leaders to Signal Resolve." *International Interactions* 43 (3): 453–79.

Megerdoomian, Karine, and Dan Parvaz. 2008. "Low-Density Language Bootstrapping: The Case of Tajiki Persian." LREC.

Mikhaylov, Slava, M. Laver, and K. Benoit. 2008. Coder Reliability and Misclassification in Comparative Manifesto Project Codings." Midwest Political Science Association, Chicago, *IL.*

Mnjama, Nathan. 2005. "Archival Landscape in Eastern and Southern Africa." *Library Management* 26 (8/9): 457–70.

———. 2010. "Preservation and Management of Audiovisual Archives in Botswana." *African Journal of Library, Archives and Information Science* 20: 139–48.

Munafò, Marcus R., Brian A. Nosek, Dorothy V.M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P.A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1): 0021.

Nam, Taehyun. 2006. "What You Use Matters: Coding Protest Data." *PS: Political Science & Politics* 39 (2): 281–87.

Naseem, Tahira, and Regina Barzilay. 2011. "Using Semantic Cues to Learn Syntax." Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, Hyatt Regency San Francisco, August 7–11.

Nelson, Laura K. 2017. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research* 49 (1): 3–42.

———. 2018. "Finding Simple Patterns in Complex Political Movements." *Northeastern University.*

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2 (2): 175.

Norris, Clayton. 2016. "Petrarch 2: Petrarcher." *ArXiv Preprint ArXiv:1602.07236.*

Osborn, Tracy, and Jeanette Morehouse Mendez. 2010. "Speaking as Women: Women and Floor Speeches in the Senate." *Journal of Women, Politics & Policy* 31 (1): 1–21.

Owens, Ryan J., and Justin P. Wedeking. 2011. "Justices and Legal Clarity: Analyzing the Complexity of US Supreme Court Opinions." *Law & Society Review* 45 (4): 1027–61.

Peinelt, Nicole, Dong Nguyen, and Maria Liakata. 2020. "TBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7047–55. Stroudsburg, PA: Association for Computational Linguistics.

Preston, Thomas, and Paul 't Hart. 1999. "Understanding and Evaluating Bureaucratic Politics: The Nexus Between Political Leaders and Advisory." *Political Psychology* 20 (1): 49–98.

Pufahl, Ingrid, and Nancy C. Rhodes. 2011. "Foreign Language Instruction in US Schools: Results of a National Survey of Elementary and Secondary Schools." *Foreign Language Annals* 44 (2): 258–88.

Qiu, Lin, Han Lin, Jonathan Ramsay, and Fang Yang. 2012. "You Are What You Tweet: Personality Expression and Perception on Twitter." *Journal of Research in Personality* 46 (6): 710–18.

Rabini, Christian, Klaus Brummer, Katharina Dimmroth, and Mischa Hansel. 2020. "Profiling Foreign Policy Leaders in Their Own Language: New Insights into the Stability and Formation of Leadership Traits." *The British Journal of Politics and International Relations* 22 (2): 256–73.

Raleigh, C., A. Linke, H. Hegre, and J. Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47 (5): 651–660.

Reinhardt, Gina. 2018. "Global South Academic Network." Accessed November 15, 2021. https://gsan.essex.ac.uk/.

Rice, Douglas, and Christopher J. Zorn. 2014. "The Evolution of Consensus in the U.S. Supreme Court." Accessed November 15, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2470029.

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): 1064–82.

Sarkees, Meredith Reid. 2010. "The COW Typology of War: Defining and Categorizing Wars (Version 4 of the Data)." *Note with Version 4 of the Correlates of War Data.*

says, Smithc2. 2014. "Beware of Dog: Bluff and Denial in Saddam Hussein's Iraq." The Yale Review of International Studies *(blog)*. January 13.

Schrodt, Philip A., John Beieler, and Muhammed Idris. 2014. "Three's a Charm? Open Event Data Coding with EL: DIABLO, PETRARCH, and the Open Event Data Alliance." ISA Annual Convention.

Segerberg, A., and W.L. Bennett. 2011. "Social Media and the Organization of Collective Action: Using Twitter to Explore the Ecologies of Two Climate Change Protests." *The Communication Review* 14 (3): 197–215.

Shala, Lubna, Vasile Rus, and Arthur C. Graesser. 2014. "A Bilingual Analysis of Cohesion in a Corpus of Leader Speeches." FLAIRS Conference. Accessed November 15, 2021. https://pdfs.semanticscholar.org/1e5b/bbac72727b2f179dac89fde481cd16adcf27.pdf.

Si, Yajuan, and Jerome P. Reiter. 2013. "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys." *Journal of Educational and Behavioral Statistics* 38 (5): 499–521.

Siegel, Alexandra, and Jennifer Pan. 2018. "Online Dissent in an Absolute Monarchy: The Effect of Repression on Key Opinion Leaders in the Saudi Twittersphere." *Northeastern University.*

Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine." In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 629–33. Piscataway, NJ: IEEE.

Sobolev, Anton, M. Keith Chen, Jungseock Joo, and Zachary C. Steinert-Threlkeld. 2020. "News and Geolocated Social Media Accurately Measure Protest Size Variation." *American Political Science Review* 114 (4): 1343–51.

Steinert-Threlkeld, Zachary C. 2017. "Spontaneous Collective Action: Peripheral Mobilization during the Arab Spring." *American Political Science Review* 111 (2): 379–403.

Tetlock, Philip E., and Gregory Mitchell. 2009. "Implicit Bias and Accountability Systems: What Must Organizations Do to Prevent Discrimination?" *Research in Organizational Behavior* 29: 3–38.

Thiers, Consuelo. 2021. "One Step Forward, Two Steps Back: The Steering Effects of Operational Code Beliefs in the Chilean-Bolivian Rivalry." In *Operational Code Analysis and Foreign Policy Roles*, 129–48. New York: Routledge.

Thies, Cameron G., and Mark David Nieman. 2017. *Rising Powers and Foreign Policy Revisionism: Understanding BRICS Identity and Behavior through Time.* Ann Arbor, MI: University of Michigan Press.

Tubishat, Mohammad, Mohammad A.M. Abushariah, Norisma Idris, and Ibrahim Aljarah. 2019. "Improved Whale Optimization Algorithm for Feature Selection in Arabic Sentiment Analysis." *Applied Intelligence* 49 (5): 1688–707.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10 (1): 178–85.

Volkens, Andrea, Judith Bara, and Ian Budge. 2009. "Data Quality in Content Analysis. The Case of the Comparative Manifestos Project." *Historical Social Research /Historische Sozialforschung* 34 (1): 234–51.

Wallach, Hanna M. 2006. "Topic Modeling: Beyond Bag-of-Words." In *Proceedings of the 23rd International Conference on Machine Learning*, 977–84. New York: Association for Computing Machinery.

Wang, Wei, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. 2016. "Growing Pains for Global Monitoring of Societal Events." *Science* 353 (6307): 1502–503.

Ward, Michael D., Andreas Beger, Josh Cutler, Matt Dickenson, Cassy Dorff, and Ben Radford. 2013. "Comparing GDELT and ICEWS Event Data." *Analysis* 21 (1): 267–97.

Weidmann, Nils B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60 (1): 206–18.

Whorf, Benjamin Lee. 1940. *Science and Linguistics.* Indianapolis, IN: Bobbs-Merrill.

———. 1957. Benjamin Lee Whorf Papers.

Windsor, Leah. 2016. "Allies in Peace: Communication in the United Nations General Assembly." *Paper presented at the International Studies Association–International Communication conference (ISA-ICOMM), Atlanta, GA.*

———. 2017a. "The Predictive Power of Political Discourse." *White Paper. Social and Behavioral Sciences Decadal Survey.* Washington, DC: National Academy of Sciences.

———. 2017b. "The Language of Radicalization: Female Internet Recruitment to Participation in ISIS Activities." *Terrorism and Political Violence* 32 (3): 506–38.

Windsor, Leah, Nia Dowell, and Art Graesser. 2014. "The Language of Autocrats: Leaders' Language in Natural Disaster Crises." *Risk, Hazards, and Crisis in Public Policy* 5 (4): 446–67.

Windsor, Leah, Nia Dowell, Alistair Windsor, and John Kaltner. 2018. "Leader Language and Political Survival Strategies." *International Interactions* 44 (2): 321–36.

Windsor, Leah Cathryn, James Grayson Cupit, and Alistair James Windsor. 2019. "Automated Content Analysis Across Six Languages." *PLoS One* 14 (11): e0224425.

Witsenburg, Karen M., and Wario R. Adano. 2009. "Of Rain and Raids: Violent Livestock Raiding in Northern Kenya." *Civil Wars* 11 (4): 514–38.

Yannitell-Reinhardt, Gina. 2021. "GSAN: Global South Action Network." Accessed November 15, 2021. https://gsan.essex.ac.uk/.

Zhang, Han, and Jennifer Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology* 49 (1): 1–57.

Zhang, Yuan, and Regina Barzilay. 2015. "Hierarchical Low-Rank Tensors for Multilingual Transfer Parsing." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Stroudsburg, PA: Association for Computational Linguistics.