International sources, local language, and biased findings from text data

Andy Halterman (MIT)* Jill Irvine (University of Oklahoma)†
Khaled Jabr (University of Oklahoma)‡

11 February 2020

Abstract

Data from text is a cheap, accurate, and useful source of empirical data for researchers in political science, but most work relies on English language text from international sources. Differences in coverage between international and local sources raises concerns that findings may be biased by the source material used. We investigate the relationship between protests in 2011 Syria and subsequent violence and demonstrate that the results are substantially different using data coded from English and Arabic news sources. Data coded from Arabic sources recovers an estimate from gold standard data (Mazur 2018), but the relationship becomes insignificant when using English language data, both hand- and machine coded. These results suggest that applied researchers should include text from the local language when using automated text analysis to study subnational outcomes and methodologists should continue to develop text analysis tools for non-English text.

Text analysis, both manual and automated, is widely used in political science to generate data on political phenomena.[1] Researchers have long known that local news sources can contain very different information than international reporting, for reasons of access and audience interest (Davenport and Ball 2002; Schrodt, Simpson, and Gerner 2001; Baum and Zhukov 2015). Despite this, researchers often draw on English language sources when compiling datasets from text because English text is widely available, many automated tools are designed for English, and researchers may not speak the local language of an area. For instance, in the past two years, international

*PhD candidate, Department of Political Science and Security Studies Program, Massachusetts Institute of Technology.

†Presidential Professor of International and Area Studies and Vice Provost for Faculty, University of Oklahoma

‡Work conducted while an MA student in the Department of Computer Science, University of Oklahoma.

[1]We acknowledge the support of the National Science Foundation under award number SBE-SMA-1539302 in funding data acquisition and annotation. MIT's Political Methodology Lab provided computing resources.

news reporting in English has been used to study riots, protests, violence, and repression in Africa (Christensen 2018; Abbs 2019; PITF, n.d.), violent contestation around elections (Daxecker, Amicarelli, and Jung 2019), violence around coups (Easton and Siverson 2018; De Bruin 2019), resistance movements' strategies (Cunningham, Dahl, and Frugé 2019), and refugee-related violence (Gineste and Savun 2019). Other work uses automated text analysis methods applied to English language text to forecast political violence and instability (Tikuisis, Carment, and Samy 2013; Shellman, Levey, and Young 2013; Mueller and Rauh 2017). While many scholars use text in the local language (e.g., among many, Lyall 2010; Toft and Zhukov 2015; Osorio and Reyes 2017; Kim 2018; Sullivan 2019), scholars have long recognized that the availability of tools in other languages can lag behind English, making making the over-representation of English a growing concern as text analysis is increasingly automated (Lucas et al. 2015).

We demonstrate the risks of studying subnational political phenomena using only data from international, English-language sources. We show that data produced from two languages, English and Arabic, leads to different conclusions on a question of substantive importance: the relationship between anti-government protests and subsequent violence in the Syrian civil war. A hand-coded gold standard dataset of protests (Mazur 2018) shows a significant effect of protests on subsequent civilian casualties in the civil war. A semi-automated dataset that we compile using Arabic language text recovers this significant effect. In contrast, three separate datasets of protests coded from English text (manually and automated) fail to find any significant relationship between protests and subsequent civilian casualties.

Existing research has shown several problems with using news text to study subnational outcomes, and political violence in particular. Baum and Zhukov (2015) show that the home regime of an international news source affected how it covered the civil war in Syria. Others have shown a geographic bias in machine-coded event data from English, with capital cities and urban areas overrepresented (Hammond and Weidmann 2014; Weidmann 2016), which is a wider problem in conflict studies (Kalyvas 2004). Existing work has not studied whether substantive research findings vary depending on the text data used.

Protests and violence in Syria

Scholars are interested in the relationship between political mobilization and subsequent violence in civil war (Shellman, Levey, and Young 2013; Balcells 2017), arguing that wartime violence by the regime is used to target political opponents. Protests, by design, serve as a public signal of opposition to the regime (Ritter and Conrad 2016). While this signal is useful for motivating other participants (Petersen 2001) it also reveals information to the government that can be used for repression (Davenport 2007). Protests in Syria, in particular, have received a large degree of scholarly attention, with attempts by several scholars to measure pre-war political mobilization and protests (Mazur 2018; Halterman 2018; Ash and Obradovich 2020). Substantively, we would

وأشار عبد الرحمن إلى «تظاهر نحو ثلاثة آلاف شخص في ادلب (شمال غرب) وريفها كما المئات في مدينة **سراقب** (ريف ادلب) رغم التواجد الأمني الكثيف». [Abdel-Rahman pointed out that "about three thousand people demonstrated in Idlib (northwest) and its countryside, as well as hundreds in the city of **Saraqib** (Idlib countryside) despite the heavy security presence."] Source: *Al Watan*	Syrian security forces also moved to crush solidarity protests that broke out all over the country, especially in neighborhoods in Homs, the country's third-largest city; **Idleb**, in the country's northwest; and the suburbs of Damascus. Source: *The Star-Ledger (Newark, New Jersey)*

Figure 1: Differences in coverage between regional and international sources drawn from our collection of English and Arabic text from LexisNexis. The two sources are reporting on the same event, but the coverage in the Arabic-language Al-Watan provides much greater detail on the number of people involved and the specific location within Idlib where the protest occurred.

like to estimate the effect of a location experiencing protests in the early phrase of the conflict on subsequent violence during the war in Syria:

$$y_i = D_i\beta + X_i\gamma,$$

where y_i is a measure of civilian casualties, D_i is the number of protests, X_i is a vector of controls, and β is our quantity of interest. Work by previous scholars has provided several measures of the number and location of protests in Syria. If these datasets are equally valid, using their estimates of D_i will not produce meaningfully different estimates of $\hat{\beta}$. If, however, alternative measures of D_i produce different estimates, this indicates that the relationship is depending on the measurement technique used.

We use a dataset of hand-coded protest events provided in Mazur (2018) as our baseline estimate of protests. This data was collected from Arabic-language, pro-regime and pro-opposition newspapers and records the date, location, size, tactics, and response to protests in the early phases of the Syrian civil war. Because it was compiled from a range of sources by a scholar who specializes in the conflict, we treat this dataset as a gold standard report on protests in Syria. We take data for control variables, such as the size, ethnic composition, and percentage of government workers in each town from both Mazur (2018) and Khaddour and Mazur (2018). Out of 5,200 towns in the set of unique towns, 130 experience a protest according to Mazur's data.

Ash and Obradovich (2020) provide two other alternative measures of protest locations in Syria. The primary data that they use is a set of protests drawn from ICEWS, a machine coded dataset that codes events from English-language text (Boschee et al. 2015). Protests are recognized using a set of key terms in a dictionary and are automatically resolved to geographic locations (Lautenschlager, Starz, and Warfield 2017). As an alternative measure of protests, Ash and Obradovich (2020) also hand-code protest occurrences and locations from English language news reporting accessed

through LexisNexis.

We provide two further datasets for comparison. First, we generate a machine coded dataset of protest locations from English text using a supervised classifier trained on hand-labeled sentences and using off-the-shelf tools to recognize place names in text. We apply this to a set of English-language text drawn from LexisNexis.[2] We use this data to illustrate the easiest case of data creation: a supervised classifier on English language text. We also create a dataset of protests from Arabic text. We downloaded every Arabic-language article from LexisNexis published in 2011 and containing the word "Syria". We trained a classifier to detect protests using 700 labeled examples and used the classifier to return articles that are likely to describe protests. We then validated these reports by hand and extracted the locations of protests reported in the stories. Figure 1 shows an example of two sentences from our data, one in Arabic and one in English, reporting on the same protest in Idlib province.

Data on our outcome variable, civilian casualties in the Syrian civil war, comes from a dataset provided in Halterman (2018). This dataset reports the coordinates and causes of over 100,000 civilian deaths during the war. We limit the count of casualties during the first year of the armed conflict (July 2011–July 2012) because later phases of civil wars become increasingly driven by "endogenous" processes of civil war, rather than political factors at the beginning of the war (Kalyvas 2006).

Mazur's dataset reports coordinates for each location, but neither of the English-language datasets reported geographic coordinates. To merge the data together, we automatically geocoded each place name (Halterman 2017) and assigned protests or casualties to the closest location in Mazur's dataset.

Measurement-induced differences in the protest–violence relationship

Figure 2 shows the correlation between three datasets on protest locations: the English-language hand coded data from Ash and Obradovich (2020), data hand coded from English and Arabic sources by Mazur (2018), and our new set of semi-automated data coded from Arabic language text. The figure shows systematic underreporting of protests in the dataset produced from English language text compared to the "gold standard" data that Mazur provides.[3] The correlation between Mazur's data and the data we code from Arabic language text is much better, though both data sources have locations that they alone report protests occurring in.

The difference in measurement also produces substantively different findings. Figure 3 shows the coefficients of a linear probability model regression of whether a location experienced civilian casualties during the first 12 months of the war on the number of protests in that location, as reported by different datasets. The models control for population, the percentage Alawite, and the percentage of government workers

[2]Details on the classifier and named entity recognition system are available in the Appendix.
[3]The underreporting is similar for ICEWS data.

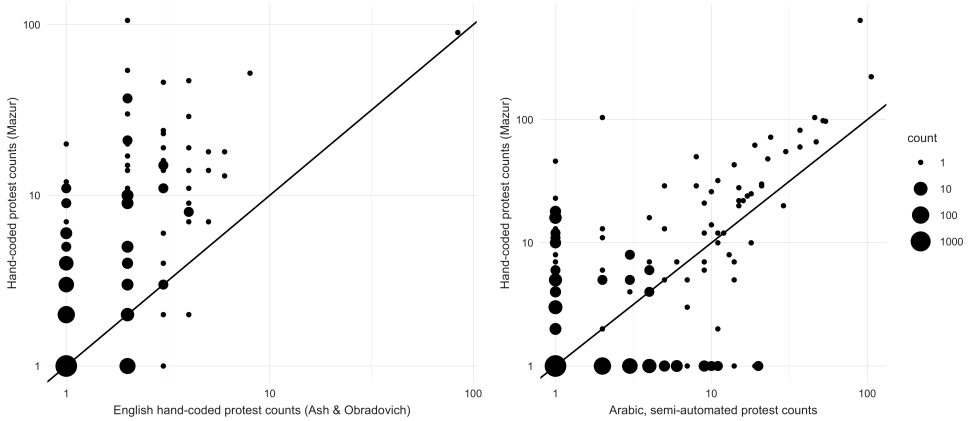

Figure 2: Correlation between the number of recorded protests in three datasets showing consistent under-reporting in English. Axes show the number of protests, plus one, logged. The left panel shows the correlation between Mazur's protest dataset coded from both English and Arabic and Ash and Obradovich's hand-coded English language dataset. The right panel shows the correlation between Mazur's dataset and our semi-automated dataset, coded from English. The 45 degree line marks perfect correlation: the points above the line represent protests that are under-reported in the alternative datasets.

(Khaddour and Mazur 2018). The two datasets that use Arabic text find a significant effect: protests are correlated with subsequent violence during the war. In contrast, the three datasets that rely exclusively on English language text find a positive but insignificant effect.

Implications

Our results demonstrate that the text sources used can affect the conclusions reached. A researcher with access only to international, English-language news sources would conclude that there is no significant relationship between protests and violence, regardless of whether the data was machine coded or hand coded. According to both hand coded and semi-automated data from Arabic text, this conclusion would be incorrect. In this context and for this question, the Arabic language data is more accurate in uncovering the relationship between protests and violence. This provides evidence that researchers using machine learning methods on text to measure subnational phenomena should use local sources in the local language, rather than relying only on international, English language reporting. Second, it suggests that while useful and cheap to produce, data derived automatically from text data is not without biases.

As researchers increasingly use text analysis to study subnational political phenomena, they will need to ensure that their conclusions are not colored by their reliance on international or English language sources. This concern will only become greater as automated techniques for text analysis become more widely used. Methodologists will need to continue their work in cross-lingual techniques for text analysis (e.g. Lucas

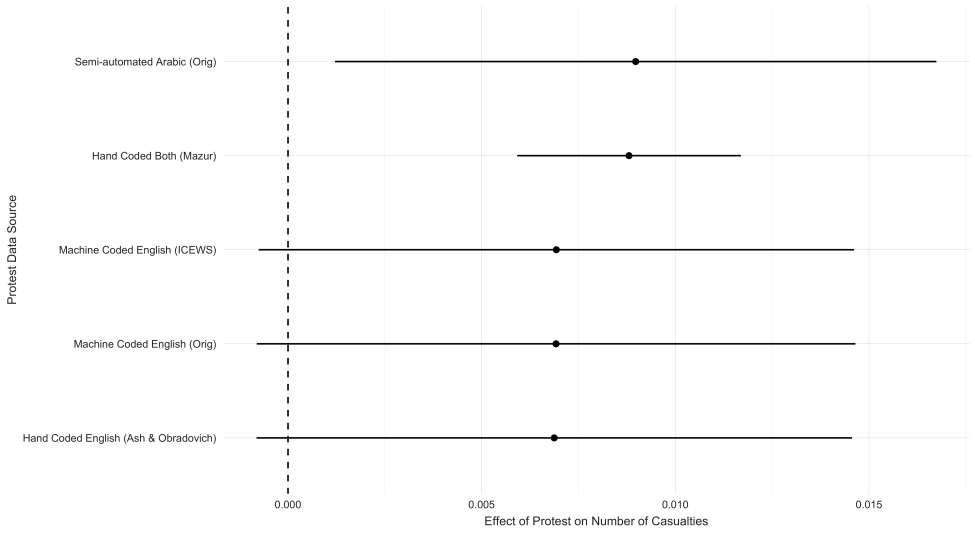

Figure 3: Coefficients from a linear probability model regression of whether a location experienced civilian casualties during the first year of the war on protest counts from five sources. The two sources that use Arabic text find a significant effect while the two sources using English-language text find no significant effect.

et al. 2015), providing guidance on decisions such as stemming and tokenization for non-English languages, research on how well different statistical classifiers work for highly-inflected languages, and employing models for natural language processing that are built for non-English languages.

References

Abbs, Luke. 2019. "The Hunger Games: Food Prices, Ethnic Cleavages and Nonviolent Unrest in Africa." *Journal of Peace Research*, 0022343319866487.

Ash, Konstantin, and Nick Obradovich. 2020. "Climatic Stress, Internal Migration, and Syrian Civil War Onset." *Journal of Conflict Resolution* 64 (1): 3–31.

Balcells, Laia. 2017. *Rivalry and Revenge: The Politics of Violence During Civil War*. Cambridge University Press.

Baum, Matthew A, and Yuri M Zhukov. 2015. "Filtering Revolution: Reporting Bias in International Newspaper Coverage of the Libyan Civil War." *Journal of Peace Research* 52 (3): 384–400.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Stephen M Shellman, James Starz, and Michael D Ward. 2015. "ICEWS Coded Event Data." In *Harvard Dataverse, V9, http://dx.doi.org/10.7910/DVN/28075*.

Christensen, Darin. 2018. "The Geography of Repression in Africa." *Journal of Conflict Resolution* 62 (7): 1517–43.

Cunningham, Kathleen Gallagher, Marianne Dahl, and Anne Frugé. 2019. "Introducing the Strategies of Resistance Data Project." *Journal of Peace Research*, 0022343319880246.

Davenport, Christian. 2007. "State Repression and Political Order." *Annual Review of Political Science* 10: 1–23.

Davenport, Christian, and Patrick Ball. 2002. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46 (3): 427–50.

Daxecker, Ursula, Elio Amicarelli, and Alexander Jung. 2019. "Electoral Contention and Violence (Ecav): A New Dataset." *Journal of Peace Research* 56 (5): 714–23.

De Bruin, Erica. 2019. "Will There Be Blood? Explaining Violence During Coups d'état." *Journal of Peace Research* 56 (6): 797–811.

Easton, Malcolm R, and Randolph M Siverson. 2018. "Leader Survival and Purges After a Failed Coup d'état." *Journal of Peace Research* 55 (5): 596–608.

Gineste, Christian, and Burcu Savun. 2019. "Introducing Posvar: A Dataset on Refugee-Related Violence." *Journal of Peace Research* 56 (1): 134–45.

Halterman, Andrew. 2018. "Violence Against Civilians in Syria's Civil War." *MIT Political Science Department Research Paper*.

———. 2017. "Mordecai: Full Text Geoparsing and Event Geocoding." *The Journal of Open Source Software* 2 (9). https://doi.org/10.21105/joss.00091.

Hammond, Jesse, and Nils B Weidmann. 2014. "Using Machine-Coded Event Data for the Micro-Level Study of Political Violence." *Research & Politics* 1 (2).

Kalyvas, Stathis N. 2004. "The Urban Bias in Research on Civil Wars." *Security Studies* 13 (3): 160–90.

———. 2006. *The Logic of Violence in Civil War*. Cambridge University Press.

Khaddour, Kheder, and Kevin Mazur. 2018. "Syria town database." Harvard Dataverse. https://doi.org/10.7910/DVN/YQQ07L.

Kim, Sung Eun. 2018. "Media Bias Against Foreign Firms as a Veiled Trade Barrier: Evidence from Chinese Newspapers." *American Political Science Review* 112 (4): 954–70.

Lautenschlager, Jennifer, James Starz, and Ian Warfield. 2017. "A Statistical Approach to the Subnational Geolocation of Event Data." In *Advances in Cross-Cultural Decision Making*, 333–43. Springer.

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*, mpu019.

Lyall, Jason. 2010. "Are Coethnics More Effective Counterinsurgents? Evidence from the Second Chechen War." *American Political Science Review* 104 (01): 1–20.

Mazur, Kevin. 2018. "State Networks and Intra-Ethnic Group Variation in the 2011 Syrian Uprising." *Comparative Political Studies*, 0010414018806536.

Mueller, Hannes, and Christopher Rauh. 2017. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review*, 1–18.

Osorio, Javier, and Alejandro Reyes. 2017. "Supervised Event Coding from Text Written in Spanish: Introducing Eventus ID." *Social Science Computer Review* 35 (3): 406–16.

Petersen, Roger D. 2001. *Resistance and Rebellion: Lessons from Eastern Europe*. Cambridge University Press.

PITF. n.d. *Political Instability Task Force Worldwide Atrocities Event Data Collection Codebook Version 1.0B2*. http://eventdata .parusanalytics.com/data.dir/atrocities.html.

Ritter, Emily Hencken, and Courtenay R Conrad. 2016. "Preventing and Responding to Dissent: The Observational Challenges of Explaining Strategic Repression." *American Political Science Review* 110 (1): 85–99.

Schrodt, Philip A, Erin M. Simpson, and Deborah J. Gerner. 2001. "Monitoring Conflict Using Automated Coding of Newswire Reports: A Comparison of Five Geographical Regions." *PRIO/Uppsala University/DECRG High-Level Scientific*

Conference on Identifying Wars: Systematic Conflict Research and Its Utility in Conflict Resolution and Prevention, Uppsala, Sweden.

Shellman, Stephen M, Brian P Levey, and Joseph K Young. 2013. "Shifting Sands: Explaining and Predicting Phase Shifts by Dissident Organizations." *Journal of Peace Research* 50 (3): 319–36.

Sullivan, Heather. 2019. "Sticks, Stones, and Broken Bones: Protest Violence and the State." *Journal of Conflict Resolution* 63 (3): 700–726.

Tikuisis, Peter, David Carment, and Yiagadeesen Samy. 2013. "Prediction of Intrastate Conflict Using State Structural Factors and Events Data." *Journal of Conflict Resolution* 57 (3): 410–44.

Toft, Monica Duffy, and Yuri M Zhukov. 2015. "Islamists and Nationalists: Rebel Motivation and Counterinsurgency in Russia's North Caucasus." *American Political Science Review* 109 (02): 222–38. https://doi.org/10.7910/DVN/DK4FXA.

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