Supervised Event Coding from Text Written in Arabic: Introducing Hadath

Javier Osorio,[1] Alejandro Reyes,[2] Alejandro Beltran,[1] Atal Ahmadzai[1]

[1]University of Arizona, [2]AUDI Mexico,
School of Government and Public Policy, University of Arizona, United States
Corresponding author: josorio1@email.arizona.edu

Abstract

This article introduces Hadath, a supervised protocol for coding event data from text written in Arabic. Hadath contributes to recent efforts in advancing multi-language event coding using computer-based solutions. In this application, we focus on extracting event data about the conflict in Afghanistan from 2008 to 2018 using Arabic information sources. The implementation relies first on a Machine Learning algorithm to classify news stories relevant to the Afghan conflict. Then, using Hadath, we implement the Natural Language Processing component for event coding from Arabic script. The output database contains daily geo-referenced information at the district level on who did what to whom, when and where in the Afghan conflict. The data helps to identify trends in the dynamics of violence, the provision of governance, and traditional conflict resolution in Afghanistan for different actors over time and across space.

Keywords: event data, information extraction, Arabic, Afghanistan, conflict

1. Introduction

Over the last few years, collaborative work between Political Scientists and Computer Scientists produced considerable contributions to understanding socio-political conflict processes around the world thanks to the production of computerized event data (Althaus et al., 2019; Bond et al., 2003; Hürriyetoğlu et al., 2019; Mohiuddin et al., 2016; O'Brien, 2012; Schrodt et al., 2014; Schrodt, 2012; Subrahmanian, 2013; Wang et al., 2016). These event coding projects take advantage of the vast availability of online news reports to extract information about *who did what to whom, where, and when*, thus producing a wealth of data about incidents of socio-political conflict around the world. Despite the enormous contributions of these research projects, the large majority of these projects primarily use text written in English as their source of information.

Unfortunately, coding socio-political incidents of conflict in foreign locations where English is not a native language using text written in English is likely to generate discrepancies that affect the quality of the data output. To address this Anglo-centric approach, in recent years, a handful of research projects started engaging in multi-lingual event coding. Some of these efforts rely on automated translation from non-English text into English (Boschee et al., 2016), while other code event data directly from text written in non-English languages (Schrodt et al., 2014; Osorio and Reyes, 2017; Osorio et al., 2019b).

In line with ongoing research advancing multi-lingual event coding, this article introduces Hadath, a supervised protocol for coding event data from text written in Modern Standard Arabic. At its core, Hadath uses shallow parsing to identify events in the corpus based on entries provided by dictionaries of actors, actions, and locations. In this way, Hadath follows other sparse coding protocols (Schrodt, 2009) while contributing to pioneering work of computerized event coding from Arabic text (Open Event Data Alliance, 2016; Halterman et al., 2018).

This application focuses on extracting event data about the Afghan conflict between 2008 and 2018 using narratives from Arabic newspapers. Following the literature on wartime order and rebel governance (Arjona, 2016; Staniland, 2012), our conceptualization of event data in the Afghan conflict includes acts of violence, as well as the provision of governance, and traditional conflict resolution. This helps moving beyond a narrow focus on violence and including other non-violent behaviors taking place in conflict settings.

The methodology consists of two main steps. First, we deploy a Machine Learning (ML) classifier to identify the specific news stories relevant to the Afghan conflict from a vast collection of news articles written in Arabic. Second, we rely on Hadath to implement a Natural Language Processing (NLP) protocol for event coding. The resulting database presents geo-referenced information at the daily district level on who did what to whom, when, and where in the Afghan conflict.

2. Recent Developments

2.1. NLP tools in Arabic

Although English is the dominant language in NLP processing tools, in recent years, scholars have been advancing non-English NLP developments including tools for Modern Standard Arabic. These emerging resources in Arabic include news articles corpora such as *i*ArabicWeb16 (Suwaileh et al., 2016; Khaled Yasser and Elsayed, 2018); annotation of social media in Arabic through the EveTAR and ArSAS projects (Hasanain et al., 2018; AbdelRahim Elmadany and Magdy, 2018); tools to identify dialects and regional variations in Arabic (Zaghouani and Charfi, 2018; Alshutayri and Atwell, 2018); and lexical disambiguation resources for Arabic diacritics (Sawsan Alqahtani and Zaghouani, 2018).

In line with these developments, the field of computerized event coding has been advancing tools for processing text written in Arabic. The Open Event Data Alliance (OEDA) is adapting Universal PETRARCH for coding event data from Arabic using the CAMEO ontology (Open Event

Data Alliance, 2016; Gerner et al., 2002). As part of this project, OEDA developed a supervised tool for translating CAMEO dictionaries from English to Arabic (Halterman et al., 2018). Hadath contributes to these efforts to developing event coding capabilities for Arabic text.

2.2. Text-as-Data in Conflict Studies

The use of text-as-data is developing solid roots in the social sciences (Grimmer and Steward, 2013). Researchers in political science and public administration are rapidly catalyzing the leverage of computerized text analysis by exploring previously uncharted domains of inquiry and developing increasingly sophisticated text analysis tools (Wilkerson and Casas, 2017; Hollibaugh, 2018).

The use of computerized event data revolutionized the way in which researchers analyze conflict processes. Traditionally, the field primarily relied on manually coded databases of conflict processes around the world such as the Uppsala Conflict Data Program (UCDP) (Sundberg and Melander, 2013) and the Armed Conflict and Event Dataset (ACLED) (Raleigh et al., 2010). In contrast with these manually coded databases, some scholars took advantage of NLP tools and the vast availability of online news articles to develop computerized event coding protocols.

The pioneering work of Schrodt opened the door to machine-generated event data from news papers through the KEDS program (Schrodt, 1998). After developing TABARI, a second generation coder based on sparse parsing, Schrodt triggered an enormous production of event data (Schrodt, 2009). The third generation of coders is PETRARCH, which incorporates full parsing and Treebanks (Schrodt et al., 2014). Parallel to these projects, the Integrated Crisis Early Warning System (ICEWS) program (O'Brien, 2010) advanced its own event coding tools.

In addition to these main programs, the field of computerized event coding in conflict studies is now populated with a variety of data generation approaches of increasing coverage, sophistication, and accuracy (Althaus et al., 2019; Bond et al., 2003; Halterman et al., 2018; Hammond and Weidmann, 2014a; Hürriyetoğlu et al., 2019; Subrahmanian, 2013; Osorio and Reyes, 2017; Osorio et al., 2019a). With a few exceptions (Osorio and Reyes, 2017; Piskorski et al., 2011), most protocols almost exclusively rely on news stories written in English, thus neglecting the richness and detail of vast amounts of information produced in foreign locations in their native languages. The field is barely opening to the possibility of processing text in non-English languages for event coding. To advance this potential, the Open Event Data Alliance has been spearheading research enabling the generation of event data in Arabic (Open Event Data Alliance, 2016; Halterman et al., 2018). Hadath contributes to these recent research endeavors to process Modern Standard Arabic for event coding.

3. Conflict in Afghanistan

The current insurgency in Afghanistan is the continuation of a decades-long warfare in the country. Starting with the invasion of the Soviet Union in late 1970s, followed by the civil war of the 1990s, and then confounded by the emergence of the Taliban in 1996, the war in Afghanistan has a long history and multiple actors. However, the roots of the current Taliban insurgency go back to late 2001 when the U.S. led international coalition forces toppled the Taliban's Emirates of Afghanistan (CFR, 2020). This intervention was in response to the 9/11 terrorist attacks that Al-Qaeda carried out in the U.S. Though, the Emirates of the Taliban were not directly involved in conducting the attacks, they provided sanctuaries and safe havens for the leadership and strategic infrastructure of Al-Qaeda in Afghanistan (Kean, 2011). Using Afghanistan as its base, the Al-Qaeda organization coordinated the 9/11 suicide attacks in the U.S.

After being defeated by the U.S. military intervention, the Taliban re-emerged as an armed insurgency against the newly establish Afghan government and the international forces in 2003 (Kenneth and Thomas, 2017). The resurgence, which started with scattered run and hit attacks in different remote parts of the country, has systematically grown both in magnitude and geographic scope in the subsequent years. By 2006, the scattered attacks had developed to a full-fledged insurgency in different parts of the country (Jones, 2008). In addition, The Taliban also strategically included different methods of inflicting targeted and indiscriminate violence including suicide bombings, implanting Improvised Explosive Devices (IEDs), and conducting well-coordinated military attacks against strategic targets.

4. Text Gathering and Classification

4.1. Gathering News Stories

As part of a larger research project funded by the United States Department of Defense - Minerva Research Initiative (71623-LS-MRI), this paper focuses on the Afghan conflict to test Hadath. Although Dari and Pashto are the official languages in Afghanistan, these languages are still part of the Arab sign-language family, which is common to several languages in Asia, Africa, and the Middle East. Since there are more resources available to the researchers in terms of text, NLP tools, and human coders in Arabic than in Dari or Pashto, we developed first the capacity to code event data in Arabic as a practical matter. Having a functional software for event coding in Arabic serves as the "possible adjacent" (Jhonson, 2011) that can be extended to other Arabic script languages, including Dari and Pashto.

To generate this collection of news articles written in Arabic, we relied on the Nexis Uni global news platform, an online repository hosting vast collections of newspapers in different languages. Nexis Uni contains 17 different newspapers published in Arabic with over 2.5 million articles collected between 2008 and 2018. To gather relevant news stories, we ran a robust query in Nexis Uni's search engine to identify articles potentially related to the conflict in Afghanistan. A team of three research assistants manually downloaded all the articles that the search output produced. This text gathering effort generated a collection of 100,857 individual stories.

For Hadath to identify the day in which an event occurred, the filename of each article must incorporate its date of publication. To enable this feature, we used the *date_extractor* Python package (Dufour, 2020) to identify the publication date, and then used this information to modify the file name of each news story so that it indicates the date of publication. This process facilitates grouping news stories in year-specific folders that served to build the training data corpus described in the section below.

4.2. Machine Learning Classifier

The Nexis Uni search engine returned several news articles not directly relevant to our selection criteria. Although the query included *boolean* exclusions to filter out unrelated stories, a considerable portion of the retrieved articles were tangential to our study.

To resolve this ambiguity, we initially tasked a team of 6 human coders to manually classify each article. Each coder was randomly assigned to a folder from which they briefly read each article and coded as “accept” or “exclude.” Coders classified the stories based on a direct reference to a conflict-related incident occurring within Afghanistan. To “accept” an article, the story must include a description of an actual event or incident related to any of the three dimensions of interest: acts of violence, provision of governance in the context of war, or traditional conflict mitigation. The most challenging aspect of this task is identifying relevant news stories that explicitly report on factual incidents of events occurring in the country, not just opinions or broad discussion loosely related to Afghanistan. In consequence, we excluded statements or opinions made by foreign persons or entities about Afghanistan and other international reports summarizing the conflict.

To asses the reliability of our human coders, the 6 coders applied the same classification routine to a random sample of 1,000 articles and we evaluated their agreement. This revealed that three coders were producing unreliable classifications. To resolve this issue we utilized a Machine Learning (ML) text classifier model described below, trained on the tags assigned by the coders with the highest agreement. The Fleiss’ Kappa for these three coders reached .817 and a paired inter-coder agreement averaged of 92%. Given this assessment of inter-coder reliability, we used their initial classification on the universe of articles as training data.

The resulting training data consists of 55,870 articles that contain no explicit biases in classification. This includes 17,426 articles tagged as “accept,” and 38,444 classified as “exclude”. Although the categories are unbalanced, we decided against balancing the data because we expect the universe of observations to follow the same distribution.

The classification pipeline first normalized the text and removed any English language characters, digits and stop words. The light stemming feature of the *Tashaphyne* python package (Zerrouki, 2012) served to reduce words to their stem. We used *TfidfVectorizer* from *sci-kit learn* (Pedregosa et al., 2011) to convert the Arabic characters into a features matrix based on the recurrence of relevant words, with the maximum number of features capped at 5,000.

To improve the accuracy of the training data and to resolve ambiguities between human coders and the machine, we trained a Logistic Regression (LR) that reported an F1 of 0.87 and used the parameters from this model to classify the training data. This process generated 3,929 articles where the human coder and model were in disagreement. Three human coders focused on resolving these ambiguities by correctly classifying each article. This subset only represents 7% of the training data but it greatly improved model performance. This step modified the number of articles in the “accept” category to 16,339.

We evaluate the performance of each model using *k*-fold cross-validation (CV) that shuffles and splits the data into 5 subsets, and leaves out 10% for validation. Figure 1 shows the average F1 performance for each models. We used a random grid search to evaluate different Convolutional Neural Network (CNN) specifications and reported the two best performing parameters. CNN 1 and CNN 2 share a vocabulary of 150,166, 128 filters and an embedding dimension of 50. CNN 1 has a kernel size of 3 and averaged an F1 of 0.923, in contrast CNN 2 has a kernel size of 5 and averaged an F1 of 0.922. The Random Forest (RF) model averaged 0.924 F1. The Multinomial Naive Bayes (NB) model averaged 0.835 F1. The Extreme Gradient Boosting (XGB) model averaged 0.919 F1. The linear Support Vector Machine (SVM) classifier averaged 0.94 F1. Using this updated training data, the LR model performance increased to 0.94 F1. To decide between LR and SVM, we re-trained the model using the entire training data rather than the *k*-fold CV approach, and left out 10% of the data for testing. This process resulted in an F1 of 0.934 for SVM and an F1 of 0.938 for LR. Given the slightly better performance of the latter, we use LR to classify the universe of articles.

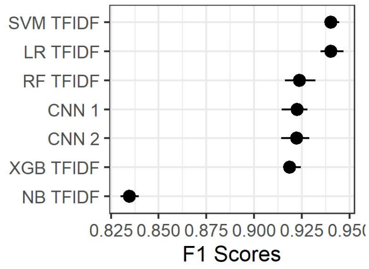

Figure 1: Machine Learning models

For the collection of 100,857 articles, we normalized and processed them following the same procedure implemented for the training data. Then we applied the parameters of the LR model to classify the entire collection, which resulted in a classification of 26,235 relevant articles.

5. Event Coding

5.1. Introducing Hadath

After selecting the relevant news stories, the next step is to extract events from the text collected. For this task we developed Hadath, a supervised NLP application for extracting events and their geographic location from text in Arabic. Hadath stems from a long tradition of coders using sparse parsing to extract event data. The first of these programs used to collect conflict data was Tabari (Schrodt, 2009). This software served as the source code or inspiration for a long line of coders in the social sciences (Best et al., 2013; Hammond and Weidmann, 2014b; Chojnacki et al., 2012; Schrodt, 2001; Schrodt et al., 2004; Schrodt, 2006; Schrodt et al., 2014; Schrodt and Gerner, 2012)

An event can be defined as the categorical description of someone doing something to someone else, at a specific time and location based on the explicit information contained in text. Event data contains five components: a **source** performing an action, the **action** observed, and the **target** of the observed action, a specific **date**, and an identifiable geographic **location**. Hadath is capable of extracting all five features from text in Arabic, and to the extent of our knowledge is the first program with these capabilities.

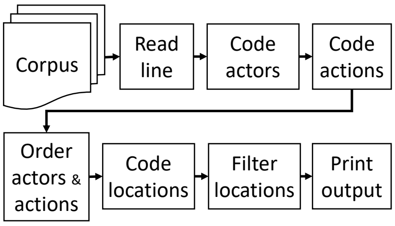

Figure 2: Hadath system

Figure 2 presents the Hadath algorithm:

1. **Read the text**. Take the Arabic corpus as an input file. The corpus contains a unique identifier per news article and for each paragraph within each news story. Paragraphs are formatted as a long line reading from right to left. Hadath uses the line as the coding unit.

2. **Coding actors**. Load the dictionary of actors containing named entities used as search criteria. Each entry has an associated numeric code. Search for those entities in the corpus, looking first for the longest names.[1] Record each entry in textual and numeric format for every match between the dictionary and the corpus.

3. **Coding actions**. Load the actions dictionary containing verb phrases. Use those entries as search criteria for detecting actions. Record the actions in the same way it does with the actors.

4. **Ordering actors and actions**. After finishing coding actors and verbs in each line, reorder the coding output according to the order in which actors and actions appear in the script (from right to left).

5. **Code locations**. For each paragraph with a matching actor or action, use the locations dictionaries to look for the provinces (states) and districts (counties) mentioned in the corpus. Save the matching locations.

6. **Filter locations**. To minimize the problem of geographic ambiguity, use the locations filter to discard false positives for geographic locations.

7. **Print output**. Print the output indicating first the date of the event, the matching actors an actions in the order they appear in the text, print any matching locations.

For this application, we use the corpus of articles on Afghanistan discussed in the previous section. This collection contains news articles in Arabic script related to the Afghan conflict from 2008 through 2018. The text was preprocessed and reformatted, producing a corpus with about 90 MB of text across 394,690 sentences.

The dictionary of actors used in this implementation contains 318 named entities in Arabic related to organizations or individuals including the main insurgent groups, coalition forces, international and local actors relevant to the Afghan conflict. This list is based on knowledge of the case, available list of relevant actors made by country experts, and the discovery of additional actors identified using Named Entity Recognition (NER) (Abdelali et al., 2016). Table 1 below reports the main actor categories.

Table 1: Main Categories of Actors

Domestic Armed Actor	**Civil Society**
Afghan Taliban	Civilians
Warlords (Mujahideen)	Ethnic
Other	Religious
Int. Armed Actors	Women
Al-Qaeda	**Local organizations**
Hamas	Educational
Hezbollah	Private sector
Int. Jihadi Groups	Political organizations
ISIS	Political Parties
Muslim Brotherhood	Other
Int. Security Forces	**Int. Actors**
ISAF	Foreign governments
US	Multilateral Organizations
Other	International NGOs
National Security Forces	Political Parties
Army	Private Corporations
Intelligence	**State**
Police	Executive
	Judicial
	Legislative

The actions dictionary comprises 4,694 Arabic verb phrases associated with violence, governance provision, or traditional conflict resolution. The verbs used in this dictionary omit pronunciation diacritics, thus leaving plain Arabic script. This required deduplicating different verb conjugations that end up with the same plain Arabic script after removing diacritics. Table 2 shows the three main categories considered in the study: Violence, Governance, and Pashtoonwali (traditional lifestyle). To build this dictionary, we first applied Part of Speech (POS) tagging on the corpus to generate an initial list of verbs. Human coders then filtered them out based based on their relevance to the Afghan conflict. Coders then added variations and synonyms of each verb to build up redundancy, which is necessary for unstructured text. To do so, we used online web resources to consider all possible verb conjugations.

To geo-reference events, Hadath uses the dictionaries of Provinces (equivalent to states) and Districts (equivalent to

[1] Prioritizing long strings improves the coding efficient by not devoting resources looking for shorter words or sub-strings.

Table 2: Main Action Categories

Violence	**Governance**
Physical violence	Judicial governance
Economic extortion	Policing
Pashtoonwali	Taxation
Conflict mitigation	
Reconciliation	

counties) to identify the location of the event. The dictionaries comprise all 34 provinces and 400 districts of Afghanistan, including variations on Province and District spelling as well as potential abbreviations.

To reduce the risk of false positives in the identification of locations, Hadath uses the locations filter to confirm that the location identified is in fact a physical location. For example, this dictionary contains nuances that Hadath uses to distinguish between Kabul street and the city of Kabul.

The final step consists of a post-coding process for cleaning the output, validating to asses performance, and removing duplicate events. The resulting output constitutes a georeferenced database of event data throughout Afghanistan at the daily-district level between 2008 and 2018.

Below, we illustrate how Hadath uses dictionaries to identify events, actors, and locations in text. Consider an example where the actors dictionary includes the following two groups: the "Taliban" as `[10100]` طالبان and "NATO soldiers" as `[20100]` عساكر الناتو. In addition, the actions dictionary includes reference to the verb "attack" `[101]` هاجم. The locations dictionary may include the district of "Khanabad" `[1401]` خان آباد and the province of "Kunduz" `[14]` كونوز. Using a sparse parsing approach, Hadath searches through the corpus to identify explicit matches of these words within the text. Table 3 presents a basic coding example of text in Arabic followed by the numeric output that Hadath generates. The table also includes the English translation for the reader to follow.

Table 3: Hadath Coding Example

الطالبان هاجمت عساكر الناتو في الخان اباد بكونوز

`20100 101 10100 1401 14`

Taliban attacked NATO soldiers in Khanabad, Kunduz

In this example, Hadath would read from right to left the Arabic sentence and identify the "Taliban" (طالبان) and "NATO soldiers" (عساكر الناتو) as the relevant actors and assign the codes `10100` and `20100`, respectively. The program would also match on the verb "attacked" (هاجم) and record the code `101`. After confirming that the location names should not be filtered out, it would identify the district of "Khanabad" (خان آباد) as `1401` and the province of "Kunduz" (كونوز) as `14`. In this way, Hadath identifies event data.

6. Results

This section describes the event data that Hadath extracted from the collection of Arabic news stories on Afghanistan. To avoid artificial inflation of event data caused by multiple sources reporting the same event, we used the "one-per-day" rule and deduplicated events using a pipeline that reduces multiple mentions of the same event per day.

Overall, Hadath detected 17,614 actors in this coding exercise. Figure 3 shows the frequency of actor records by category. Results show that the Taliban is the most active armed group in the conflict with 16.8% of the total records. Other domestic armed groups like the Mujahideen warlords (with 3.7%) and international armed groups such as Al-Qaeda (with 5.5%) and the Islamic State of Iraq and Syria, ISIS (with 4.5%) are also present in Afghanistan, but their salience in the conflict is secondary when compared to the Taliban. In general, the combination of all domestic and international non-state armed actors amounts to 30.8% of the actors recorded in Afghanistan, which is indicative of the high degree of complexity of this conflict.

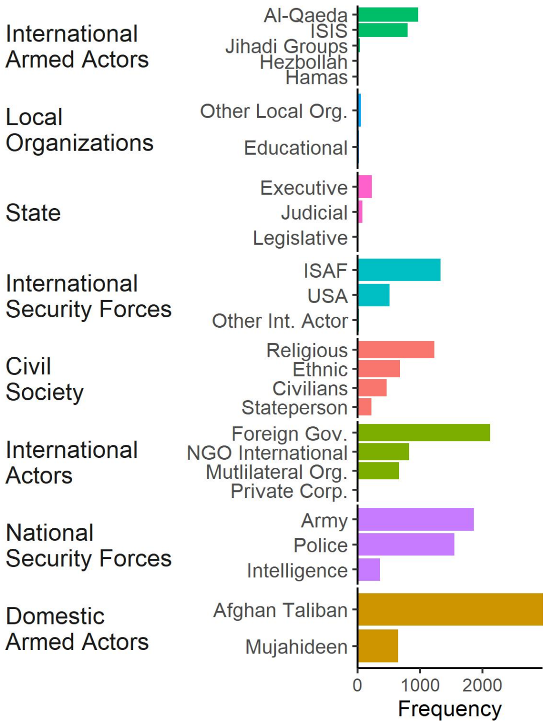

Figure 3: Frequency of Actors.

The data also reveals the central role of the Afghan security forces (with 21.4% of the total) as well as the international forces (with 10.5%) active in Afghanistan. It is noticeable, that the Afghan Army (10.6%) and the Police (8.8%) have a more active role than United States troops (2.9%) and International Security Assistance Forces, ISAF (7.6%). In general, the combined contribution of Afghan and international security forces amounts to 31.9% of the records. This share of the total detections is comparable to that of the insurgent organizations, which is indicative of the balance of forces and the high level of contestation in the Afghan conflict.

In addition to outlining the main contenders in the Afghan conflict, the data highlights the involvement of the international community (20.5% of the total) in Afghanistan. There is intense activity form foreign government representatives, diplomatic missions, international NGOs, and multilateral organizations. Finally, the data also is suggestive of the high degree of diversity in the Afghan civil society sector (14.7%) with the relevant presence of religious figures as well as representatives of different ethnic groups.

Figure 4 details the types of actions that Hadath identified in this coding exercise. In contrast to over-simplistic characterizations of the Afghan conflict, the data reveals the importance of Governance provision in the midst of war. In general, about 54.4% of the actions relate to the supply of Governance in the form of judicial services (18.4%), policing (19.9%), and taxation (16.1%). The second most relevant action category relates to the use of violence in the context of the armed conflict, with 28.6% of the records. As expected, the category of physical violence reports the highest proportion of incidents with 23.5% of all recorded actions. Acts of economic extortion amount to 5.1% of the cases. Finally, the data highlights the importance of traditional practices of conflict resolution and reconciliation at the community level (Pashtoonwali), which accounts for 17% of the data.

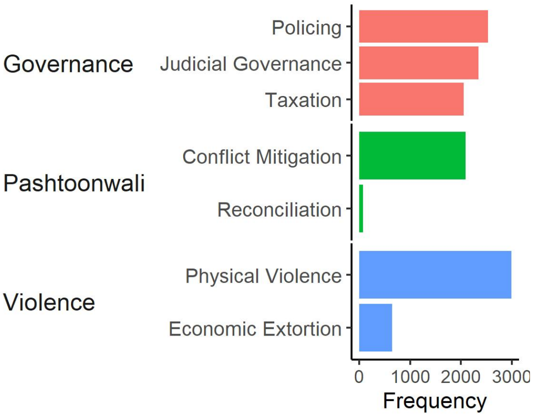

Figure 4: Frequency of Actions

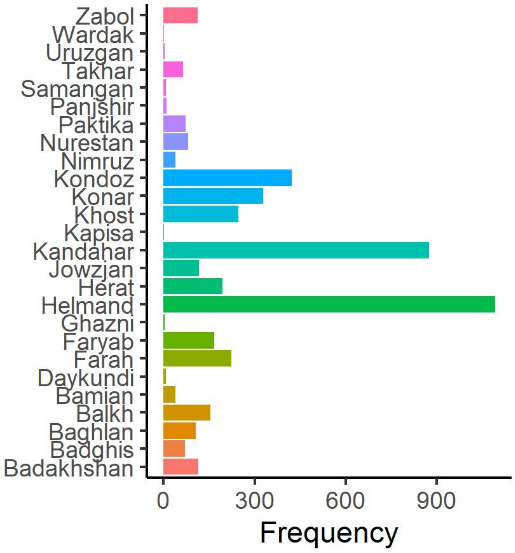

Figure 5: Provinces

Figure 5 presents the geo-location of events at the province level. As the graph indicates, the majority of incidents take place in the Helmand province and in Kandahar, which are well known hot-spots of conflict in Afghanistan.

Figure 6 showing the trends of actors over time. This graph helps to identify some insightful dynamics in the Afghan conflict. The first escalation of activity between 2008 and 2015 directly relates to the Taliban expansion from rural areas to populated urban centers such as Kabul, thus increasing its influence over the country (Masadykov et al., 2010). In addition, the Islamic State of of Iraq and Syria (ISIS or Daesh) emerged in Afghanistan around 2015 and expanded its activities to the Eastern regions of the country (Gambhir, 2015). The sharp decline in event detection in 2016 reflects the withdrawal of U.S. military personnel from Afghanistan in 2015 (CFR, 2020). Between 2011 and 2016, U.S. forces declined from 140,000 to 10,000 troops. As U.S. troops were reassigned, the number of news articles associated with this actor experience a drop from 2015 to 2016, but then it increases as the Islamic State gained traction in the region. Later on, the United States reconsidered its Afghan strategy in 2017 and increased its military presence under the new South Asia strategy. This second surge of U.S. military activity in Afghanistan reinvigorated the dynamics of conflict between 2017 and 2018.

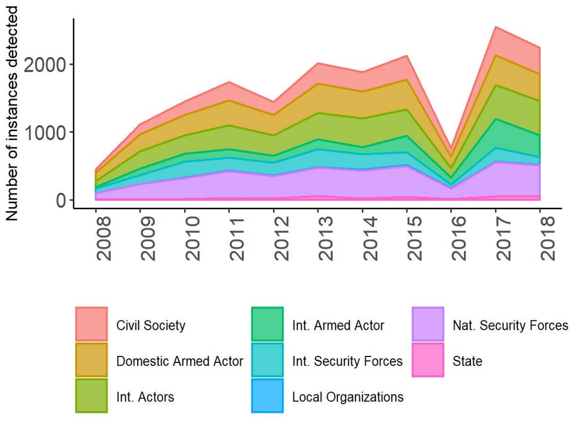

Figure 6: Temporal Trends

7. Future Research

Future research will focus on validating the coding output against manually annotated Gold Standard Records. This is will be an iterative process to improve the dictionaries of actors, actions, and locations in order to reduce the discrepancies between the computer output and the human annotation. Additional developments will consider enabling

Part of Speech tagging (POS) and Treebanks for semantic role assignment. This will help identifying of directionality of an event by defining the relationship of source-action-target. Finally, future research will enable a broader search criteria for detecting geographic locations. In its current version, Hadath only looks for a toponym (province or district) in any line containing an actor or an action. However, many instances do not mention locations in the paragraph where the event was extracted because news reports often indicate the location at the beginning of the article. In this way, future research will improve the functionality of Hadath to code event data from Arabic.

8. Conclusion

Hadath is a novel protocol for supervised event coding from text in Arabic. This software uses shallow parsing to match entries contained in dictionaries of actors, actions, and locations mentioned in a corpus written in Arabic script. In this way, Hadath contributes to research on Natural Language Processing by moving away from English-centered developments and advancing multi-lingual event data extraction. To test the functionality of Hadath in processing event data, this implementation focused on extracting events from Arabic news stories related to the conflict in Afghanistan between 2008 and 2018. To compile the collection of news articles used as corpus for this application, we developed a Machine Learning classifier to identify the specific news stories relevant to the Afghan conflict. After compiling the corpus, we used Hadath to generate event data identifying who did what to whom, when and where. The coding output allows identifying the salience of different actors related to the Afghan conflict as well as their behavioral dynamics. In this way, Hadath opens the door to future ML and NLP advances for generating event data from text written in Arabic.

9. Acknowledgements

This research was possible thanks to the support of The United States Department of Defense - Minerva Research Initiative, Award No. 71623-LS-MRI.

10. Bibliographical References

Althaus, S., Bajjalieh, J., Carter, J. F., Peyton, B., and Shalmon, D. A. (2019). Cline Center Historical Phoenix Event Data.

Arjona, A. (2016). *Rebelocracy: Social Order in the Colombian Civil War*. Cambridge University Press, NY.

Best, R. H., Carpino, C., and Crescenzi, M. J. C. (2013). An analysis of the TABARI coding system. *Conflict Management and Peace Science*, 30(4):335–348, jul.

Bond, D., Bond, J., Oh, C., Jenkins, C. J., and Taylor, C. L. (2003). Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development. *Journal of Peace Research*, 40(6):733–745.

Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., and Ward, M. (2016). ICEWS Coded Event Data.

CFR. (2020). A timeline of the u.s. war in afghanistan.

Chojnacki, S., Ickler, C., Spies, M., and Wiesel, J. (2012). Event Data on Armed Conflict and Security: New Perspectives, Old Challenges, and Some Solutions. *International Interactions*, 38(4):382–401.

Gambhir, H. (2015). Isis in afghanistan. *Backgrounder, Institute for the Study of War*, 6.

Gerner, D., Schrodt, P., Yilmaz, O., and Abu-Jabr, R. (2002). The Creation of CAMEO (Conflict and Mediation Event Observations): An Event Data Framework for a Post Cold War World. In *The Annual Meeting of the American Political Science Association*. 01/10/2013.

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

Halterman, A., Irvine, J., Grant, C., Jabr, K., and Yand, L. (2018). Creating and Automated Event Data System for Arabic Text.

Hammond, J. and Weidmann, N. B. (2014a). Using machine-coded event data for the micro-level study of political violence. *Research & Politics*, 1(2):2053168014539924, jul.

Hammond, J. and Weidmann, N. B. (2014b). Using machine-coded event data for the micro-level study of political violence. *Research and Politics*, 1(2):1–8.

Hollibaugh, G. E. (2018). The use of text as data methods in public administration: A review and an application to agency priorities. *Journal of Public Administration Research and Theory*, 29(3):474–490.

Hürriyetoğlu, A., Yörük, E., Yüret, D., Yoltar, Ç., Gürel, B., Duruşan, F., and Mutlu, O. (2019). A Task Set Proposal for Automatic Protest Information Collection Across Multiple Countries. In Leif Azzopardi, et al., editors, *Advances in Information Retrieval. 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II*. Springer.

Jhonson, S. (2011). *Where Good Ideas Come From*. Riverhead Books, New York.

Jones, S. (2008). The rise of afghanistan's insurgency: State failure and jihad. *International Security*, 32(4):7–40.

Kean, T. (2011). *The 9/11 commission report: Final report of the national commission on terrorist attacks upon the United States*. Government Printing Office.

Kenneth, K. and Thomas, C. (2017). Afghanistan: Post-taliban governance, security, and us policy. *Congressional Research Service, available at: www. fas. org/sgp/crs/row/RL30588. pdf*.

Masadykov, T., Giustozzi, A., and Page, J. M. (2010). Negotiating with the taliban: toward a solution for the afghan conflict.

Mohiuddin, S., Salam, S., Mustafa, A. M., Khan, L., Brandt, P. T., and Bhavani, T. (2016). Near Real-Time Atrocity Event Coding. *IEEE Intelligence and Security Informatics (ISI) 2016*.

O'Brien, S. (2012). A multi-method approach for near real time conflict and crisis early warning. In V. S. Subrahmanian, editor, *Handbook of Computational Approaches to Counterterrorism*, pages 401–418. Springer, NY.

Open Event Data Alliance. (2016). Universal Petrarch.

Osorio, J. and Reyes, A. (2017). Supervised event coding from text written in spanish: Introducing eventus id. *Social Science Computer Review*, 35(3):406–416.

Osorio, J., Mohamed, M., Pavon, V., and Brewer-Osorio, S. (2019a). Mapping violent presence of armed actors in colombia. *Advances of Cartography and GIScience of the International Cartographic Association*.

Osorio, J., Pavon, V., Salam, S., Holmes, J., Brandt, P. T., and Khan, L. (2019b). Translating CAMEO verbs for automated coding of event data. *International Interactions*, 45(6):1049–1064.

O'Brien, S. P. (2010). Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12(1):87–104.

Piskorski, J., Tanev, H., Atkinson, M., Van Der Goot, E., and Zavarella, V. (2011). Online news event extraction for global crisis surveillance. In *Transactions on computational collective intelligence V*, pages 182–212. Springer.

Raleigh, C., Linke, A., Hegre, H., and Karlsen, J. (2010). Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Schrodt, P. and Gerner, D. (2012). Fundamentals of Machine Coding. In *Analyzing International Event Data: A Handbook of Computer-Based Techniques*.

Schrodt, P., Gerner, D., and Yilmaz, O. (2004). Using Event Data to Monitor Contemporary Conflict in the Israeli-Palestine Dyad. In *Annual Meeting of the International Studies Association*.

Schrodt, P., Beieler, J., and Idris, M. (2014). Three's a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance. In *International Studies Association*, Toronto.

Schrodt, P. (1998). Kansas Event Data System.

Schrodt, P. (2001). Automated Coding of International Event Data Using Sparse Parsing Techniques. Chicago. Paper presented at the annual meeting of the International Studies Association.

Schrodt, P. (2006). Twenty Years of the Kansas Event Data System Project. *The Political Methodologist*, 14(1):2–6.

Schrodt, P. (2009). TABARI. Textual Analysis by Augmented Replacement Instructions.

Schrodt, P. a. (2012). Precedents, Progress, and Prospects in Political Event Data. *International Interactions*, 38(4):546–569, sep.

Staniland, P. (2012). States, Insurgents, and Wartime Political Orders. *Perspectives on Politics*, 10(2):243–264.

Subrahmanian, V. S. (2013). *Handbook of Computational Approaches to Counterterrorism*. Springer, New York.

Sundberg, R. and Melander, E. (2013). Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Wang, W., Kennedy, R., Lazer, D., and Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1504.

Wilkerson, J. and Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20(1):529–544.

11. Language Resource References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16.

AbdelRahim Elmadany, H. M. and Magdy, W. (2018). Arsas: An arabic speech-act and sentiment corpus of tweets. In Hend Al-Khalifa, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Alshutayri, A. and Atwell, E. (2018). Creating an arabic dialect text corpus by exploring twitter, facebook, and online newspapers. In Hend Al-Khalifa, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Dufour, D. J. (2020). date-extractor.

Hasanain, M., Suwaileh, R., Elsayed, T., Kutlu, M., and Almerekhi, H. (2018). Evetar: building a large-scale multi-task test collection over arabic tweets. *Information Retrieval Journal*, 21(4):307–336.

Khaled Yasser, Reem Suwaileh, A. S. Y. B. M. K. and Elsayed, T. (2018). Iarabicweb16: Making a large web collection more accessible for research. In Hend Al-Khalifa, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Sawsan Alqahtani, M. D. and Zaghouani, W. (2018). A large scale comprehensive lexical inventory for modern standard arabic. In Hend Al-Khalifa, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Suwaileh, R., Kutlu, M., Fathima, N., Elsayed, T., and Lease, M. (2016). Arabicweb16: A new crawl for today's arabic web. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 673–676.

Zaghouani, W. and Charfi, A. (2018). Guidelines and annotation framework for arabic author profiling. In Hend Al-Khalifa, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Zerrouki, T. (2012). Tashaphyne, arabic light stemmer.