Multi-language Event Coding Using Eventus ID

Javier Osorio *†

August 2016

Abstract

Recent innovations have facilitated the generation of massive amounts of data on conflict using automated coding protocols. Unfortunately, most of these approaches rely almost exclusively on English-language sources, thus neglecting valuable local sources written in their native language. This research analyzes protest events in Mexico, Brazil, France and Ukraine using Eventus ID 3.0, a novel event coding software capable of processing text written in Spanish, Portuguese, French and Russian. The study compares reports of The New York Times and narratives issued by local newspapers in their respective language. Results show that English-based event data suffers from problems of under-reporting, over-aggregation, and inaccuracy with respect the actors, tactics, and location when compared to local sources.

Keywords: Event data, protest, automated coding.

Word count: 5,974 words

*Assistant Professor, Department of Political Science, John Jay College of Criminal Justice, City University of New York. Contact: josorio@jjay.cuny.edu

†This work was supported by the National Science Foundation [NSF–RIDIR-1539302]. Special thanks to Yamilette Peguero, Natalia Nazarova, Anne-Sophie Rey, and Erik Alonzo for their valuable research assistance. The usual caveat applies.

Introduction

The massive availability of online news reports and recent advancements in natural language processing have favored the development of computerized methods for generating big data collections on conflict (Schrodt 2012, Schrodt, Beieler and Idris 2014, Leetaru and Schrodt 2013, Hanna 2014, Bond et al. 2003, Subrahmanian 2013, O'Brien 2012). However, most of these efforts primarily rely on English-language information for generating event data on conflicts taking place in foreign locations. Unfortunately, this Anglo-centric approach prevents researchers from taking advantage of vast volumes of information from local sources written in their native languages. In consequence, researchers are likely to miss timely, highly-detailed and accurate information about who did what to whom, when and where. Moreover, exclusively relying on English-based information sources is likely to distort the accuracy of the computer-generated data and point to misleading conclusions.

This research shows the consequences of generating event data on conflict in foreign locations by relying on English language information sources, and presents a technological solution for multi-language event coding using Eventus ID 3.0. The application analyzes the dynamics of dissent-repression in Mexico, Brazil, France and Ukraine by comparing reports of The New York Times (NYT) about protests taking place in those foreign locations and reports issued by local newspapers in Spanish, Portuguese, French and Russian, respectively.

Results show that event data generated from NYT suffers problems of under-reporting, over-aggregation, and inaccuracy when compared to event data generated from local information sources in their native language. The lack of accuracy in detecting actors (perpetrators and targets), tactics, dates and locations of occurrence in English based sources calls into question the validity of large event coding efforts.

The technological innovations of Eventus ID 3.0 extend the initial efforts of Osorio and Reyes (2016) for coding events from text written in Spanish. The most recent version of Eventus ID is capable of coding geo-referenced event data from other Romance languages such as Portuguese and French, as well as processing text in Cyrillic script, thus opening

the possibility of coding in Russian. In addition, the software goes beyond the traditional source-action-target event structure characteristic of other event coding programs such as Tabari (Schrodt 2009) and Petrarch (Schrodt, Beieler and Idris 2014). To do so, Eventus ID 3.0 includes a novel event coding algorithm that accommodates to more complex sentences including multiple sources as well as to incomplete sentences where the actor or the target are omitted. This multi-language event coding innovation contributes to ongoing efforts of the Open Event Data Alliance for generating mass event data in multiple languages (Schrodt, Beieler and Idris 2014), and potentiates other ancillary programs for natural language processing in non-English languages (The Stanford Natural Language Processing Group 2014).

The manuscript consists of four sections. The first segment discusses the characteristics of event data. The second presents the technological innovations included in Eventus ID 3.0. The third part presents an application for coding protest event data in Mexico, Brazil, France, and Ukraine using news reports in English and local newspapers in their native languages. The final section discusses future challenges in multi-language automated event coding.

What is an event?

Eventus ID is a software for supervised coding of event data from unstructured text written in multiple languages including Spanish, French, Portuguese, Russian, and English. The supervised character of the software requires human intervention for developing actor and verb dictionaries. Eventus ID's core function is to code event data using a system of pattern recognition that identifies actors, actions, and locations from unstructured text, thus providing information on who did what to whom, when and where. In addition to two event coding algorithms, the program includes an embedded function for geo-referencing event data at the sub-national level. To the best of the author's knowledge, this is the first program specifically designed for multi-language event coding.

An event is traditionally defined a set of categorical information describing the interaction between actors (Schrodt 1994). At its most basic level, an event provides a description of someone (*source*) doing something (*action*) to someone else (*target*). This tri-dimensional conceptualization of events, referred here as state-action-target (SAT) structure, can be expanded to include a *location* and a *date*.

Other event coding programs such as Tabari (Schrodt 2009) and Petrarch (Schrodt, Beieler and Idris 2014) work with strict compliance to the traditional SAT event structure, and only generate an event output when the source, action and target are explicitly stated in the text. However, if any of those elements is not directly mentioned in the corpus, then those programs would not generate any event coding output. Having a strict SAT requirement made sense in Tabari and Petrarch since they were developed for capturing state interactions at the international level. The SAT requirement also stems from an exclusive Anglo-centric approach. English is one of the few languages that explicitly requires having a subject in the sentence. In contrast, the vast majority of languages in the world are "pronoun-dropping" (pro-drop) since their grammar rules allow omitting pronouns while making them inferable through the verb conjugation. Unfortunately, imposing an strict SAT requirement is likely to generate false negatives when the source or target are not explicitly mentioned in the text.

There are two main reasons for omitting an actor in an event. First, some ontologies do not have a specific target. Consider for example the sentence: "protesters gathered in the main square." In this case, the nature of the event consists of an unilateral action, rather than a dyadic interaction. In consequence, there is no material target in the event. Since the target is omitted in the sentence, implementing a strict SAT requirement would not extract any event from the text, thus generating a false negative.

A second reason for omitting an actor from a sentence is grammatical rather than ontological. Languages can be categorized as "pro-drop" or "non-pro-drop" depending to the extent of which their grammar allows for dropping a pronoun in a sentence (Chomsky 1981, Cole 1987, Huang 1984). The majority of world languages have an elaborate system of verb-

subject agreement in which the verb conjugation includes the subject an implicit manner, thus allowing the subject to be explicitly omitted from the clause. In fact, strict subject-verb agreement is the anomaly rather than the rule, since only six languages including English require an explicit subject (Haspelmath 2001; 1500). All other languages have pro-drop grammatical rules.

A pervasive pro-drop grammatical rule is the present indicative. The use of this tense is particularly common in journalistic writing in Romance languages. For example, according to Guízar García (2004), 73% of newspaper headlines in Mexico use the present indicative. The popularity of the present indicative might relate to its capability of referring to actions that are taking place at the time they are being told, thus reinforcing the idea of novelty that media outlets try to convey.

In order to understand the essence of the present indicative in Spanish in comparison to English, it is easier to refer first to the present progressive. In English, the present progressive tense is a finite form of the verb that clearly defines the mood, tense, and person in the sentence. For example, in the phrase "they are arresting a criminal" the verb "to arrest" is conjugated in the indicative mood, present progressive tense, third person plural. The present progressive sentence that literally corresponds to this example in Spanish is "*ellos están arrestando a un criminal*". However, in Spanish one would simply use the present indicative tense which offers a more concise manner of conveying the same idea; thus the sentence would read "*arrestan a un criminal.*"

In Spanish, the present indicative is formed by removing the infinitive ending of the verb (e.g. taking out the final "*ar*" from the verb "*arrestar*") and replacing it with an ending that indicates the person performing the action. In this case, the stem "*arrest*" is complemented by the suffix "*an*," that refers to the third person in plural ("*ellos*" or "they"), so the resulting conjugation is "*arrestan*". What makes this conjugation form particularly challenging for event coding purposes is that the verb already provides implicit information about the person as part of the conjugation, and in consequence the subject of the action is often omitted from

the sentence. The grammatical structure for conjugating present indicative verbs in Spanish is the same in other Romance languages such as French, Portuguese, and Italian.

Since the present indicative has no explicit source (the person is already implicit in the verb conjugation), the syntactical structure does not follow the traditional source-action-target structure. In consequence, a strict SAT algorithm would not identify a complete triplet, and would yield no coding output. Due to the ubiquity of this grammatical structure in media outlets, it is necessary to develop a coding protocol capable of adapting to the complexities of non-English sources.

The ontological and syntactical examples discussed above challenge the traditional SAT approach and illustrate the necessity of developing coding algorithms that relax the strict source-action-target requirement, and allow for more flexible event configurations.

Coding Event Data using Eventus ID

Technical innovations

Eventus ID uses dictionaries of actors, verbs, and locations to identify the key elements of an event from unstructured text. The dictionaries of actors contain lists of nouns and proper names that are used for identifying the source and target actors in the corpus. The verbs dictionary is used to detect the actions being conducted as described in the news reports. To identify these elements, Eventus ID relies on the principles of the *sparse parsing* technique originally implemented in TABARI. The sparse parsing method uses the actor and verb dictionaries as searching criteria to classify only the relevant parts of the text that correspond to an event, while the rest of the text is ignored for coding purposes.[1] This shallow analysis does not rely on full syntactic trees, but on local matching to identify actors and verbs. Once these elements are detected in the corpus, Eventus ID transforms the textual information

[1] In contrast to Eventus ID and Tabari, Petrarch relies on full parsing and part-of-speech tagging.

into numeric format and stores the codes in a database. The program also identifies the date and the location of events.

As discussed by Osorio and Reyes (2016), earlier versions of Eventus ID included two algorithms for event coding and an embedded geo-location protocol. The *general sequence algorithm* was designed for coding events that strictly comply with the traditional source–action–target structure. In addition, the *partial sequence algorithm* was designed for coding incomplete events that omit the source or the target, thus providing a more flexible alternative to the traditional SAT requirement implemented in Tabari and Petrarch. The geo-location algorithm is capable of detecting toponyms at two sub-national levels (e.g. state and municipal).

The coding algorithms in Eventus ID 3.0 maintain the capability for identifying events according to the traditional source-action-target triplet using the general sequence algorithm, and allow coding incomplete events using the partial sequence algorithm. In addition, the key technological innovation in the latest version of Eventus ID includes a *multiple actors algorithm* that allows coding both general and partial event structures while considering several actors. This development enables coding more complex grammatical structures.

To illustrate how the multiple actors algorithm works consider the following paragraph:
`‘‘Elements of the Federal Police and members of the Army, acting in coordination with the Office of the Attorney General, arrested the leader of a criminal group in a joint operation in Tijuana, Baja California. In addition, 13 kilograms of cocaine were seized in the operation.’’`
In this case, Eventus ID will identify “`Federal Police`,” “`Army`,” and “`Office of the Attorney General`” as the sources of the event, then it will classify `‘‘arrested’’` as the action, and finally it will detect “`leader of a criminal group`” as the target of the event. This will be the result of the joint application of the general sequence algorithm and the multiple actor algorithm. The geo-location algorithm will identify “`Baja California`” as the state, and “`Tijuana`” as the city mentioned in the first sentence. In addition, the coding

protocol will detect the period separating the sentences in the paragraph, thus preventing the elements of the first sentence to affect the coding of the second sentence by resetting the event coding algorithms. In the second sentence, the partial sequence algorithm will identify "`cocaine`" as an actor and "`seized`" as a verb. Since the state and city are mentioned earlier in the same paragraph, the geo-location algorithm will impute these locations in the event extracted from the second sentence. Once these pieces of information are detected in the corpus, the program restructures the events to generate the following output:[2]

1. `date` → `Federal Police` → `arrested` → `leader of a criminal group` → `Baja California` → `Tijuana`
2. `date` → `Army` → `arrested` → `leader of a criminal group` → `Baja California` → `Tijuana`
3. `date` → `Office of the Attorney General` → `arrested` → `leader of a criminal group` → `Baja California` → `Tijuana`
4. `date` → `cocaine` → `was seized` → `.` → `Baja California` → `Tijuana`

In this way, the combination of algorithms embedded in Eventus ID 3.0 are capable of extracting fine grained information from complex sentences, thus providing accurate information on who did what to whom, when and where. Researchers can then use this detailed information to better accommodate their particular research objectives as part of the post-coding procedure. For example, researchers interested in analyzing state behavior using the standard "one per day" approach could aggregate the "`Federal Police`," "`Army`," and "`Office of the Attorney General`" into a single category of "Government Authorities," and compress events 1-3 into a single incident. This form of aggregation would consider the state as a monolithic entity and would regard two incidents of arrest as duplicates that can be easily eliminated. Alternatively, studies focused on disaggregating the state could be interested analyzing the coordination between different security forces, and thus could prefer using fine-grained information. Researchers could also easily develop additional re-coding

[2]The → sign indicates a tabular space in the output file.

rules to re-arrange the position of the source in event 4 and turning it into a target while imputing "Government Authorities" as the source of the event.

Coding in Multiple Languages

In addition to the multiple actors algorithm, Eventus ID 3.0 includes an expanded capability for processing text written in a variety of languages using Latin script including Spanish, French, Portuguese, and English.[3] The program also enables the possibility of processing text written in Cyrillic script, thus opening the possibility to code in Russian. The sparse parsing technique implemented in Eventus ID makes the software agnostic to the specific grammatical rules of different languages. All the researcher has to do is to develop dictionaries that contain lists of nouns and verbs to be identified in the corpus. Although effective, this approach has an implicit trade-off as it requires researchers to develop exhaustive lexicon to effectively classify the behavior of interest. The effort of dictionary development substantially increases as the scope and complexity of behavior of interest expands, and as the text in the corpus becomes less structured.

To code in multiple languages using Latin script, Eventus ID implements by default a "brute force" approach by removing from the corpus the tildes from specific letters used in distinct languages. For example, the program substitutes the acute accent " á " with a regular "a," the grave accent " ò " with a regular "o," the cedilla "ç" with a regular "c," and the letter " ñ " with "nn." These simple but effective modifications make the corpus language agnostic and allow the software to process text in a variety of languages that use Latin script as conventional writing form.

Although useful, the tilde substitution approach has its limitations as it blunts some nuanced language characteristics that might be relevant for some researchers. For example, by eliminating the acute accent, this approach would not be able to distinguish between the Spanish words "*mato*" (which translates as "I kill," the first person in present tense in English)

[3]The software has been tested in additional languages such as Italian, Tagalog, Montenegrin. However, the results of those coding protocols are not reported in this paper.

and "*mató* " (which translates as "he/she killed," the third person singular in preterite). If the researcher is primarily interested in identifying an event in which "someone killed someone else" regardless of the nuanced conjugation of the verb, then this Eventus ID feature will facilitate the event data generation in multiple languages using Latin script. If the researcher is interested in capturing specific grammatical nuances based on special characters, then the program allows modifying the default setting so the tildes are not removed from the corpus.

In addition to processing text written in Latin scrip, Eventus ID 3.0 is capable of coding text written in Cyrillic alphabet, thus opening the possibility of coding in Russian and other languages that use this script as written form of expression. This technological development contributes to recent efforts in natural language processing for information extraction in Russian (Pivorvarova, Du and Yangarber 2013, Du et al. 2013). The program allows coding in Cyrillic by simply processing input files (e.g. corpus and dictionaries) in UTF-8 format. Future developments could easily expand the capability of the software to process text written in other segmental alphabets such as Greek. Enhancing event coding capabilities to process logographic writing systems such as Chinese or partially phonemic writing systems like Arabic might be more difficult.

The multi-language characteristic of Eventus ID allows processing the following sentence written in different languages:

English: "`Students protested in front of the Presidential House`"

Spanish: "`Los estudiantes protestaron frente a la Casa Presidencial`"

French:[4] "`Les etudiants ont manifeste devant la maison presidentielle`"

Portuguese: "`Os estudantes protestaram em frente da casa presidencial`"

Russian: "`Студенты протестовали перед президентском доме`"

and generate the same event coding output identifying "students" as the source, "protested" as the action being conducted, and "the presidential home" as the target.

[4]This sentence deliberately omits accute accents in "étudiants," "manifesté," and "présidentielle."

Application for Generating Protest Data

To demonstrate the use of Eventus ID 3.0, this paper presents an application for coding event data on protest movements in Mexico, Brazil, France, and Ukraine using newspapers in their native languages, and comparing them with English-based news reports. The information gathering considered news reports mentioning protest movements taking place in these four countries between January 1st, 2013 and December 31st, 2014. The news stories came from local newspapers including *El Universal* from Mexico, which reports information in Spanish; *Jornal do Brasil* from Brazil, which reports information in Portuguese; *Le Monde* from France, which reports information in French; and *Golos* from Ukraine, which reports information in Russian. In addition, the application included reports from the New York Times noting protest movements taking place in the above mentioned foreign locations. The coding protocol considered detailed dictionaries of actors, actions, and locations for each country in English and in its respective native language. In general, the application reveals that English-based information sources suffer from considerable problems of under-reporting, over-aggregation and inaccuracy when compared to event data derived from local sources in their native language.

Table 1 shows that The New York Times largely under reports the number of news stories related to protest in foreign locations when compared to the number of news reports that local newspapers generate. The average under-coverage of news reports from NYT is about 70%, but it ranges from the relatively best covered country, Ukraine, with 44% under-coverage, to the least covered country, France, that has 91% under-coverage. Right at the outset of the analysis, it is clear that relying exclusively on U.S.-based information sources to analyze conflict in foreign locations generates severe problems of under-reporting.

[Insert Table 1]

The immediate consequence of limited news coverage of the NYT is that English-based news contain fewer events than local information sources. Figure 1 reports the number of

events across countries and information sources. The data is aggregated by three types of actions: dissent, repression, and conflict. Dissent refers to material acts of protest conducted by challengers (e.g. rallies, protests, or strikes). Repression refers to material acts of coercion conducted by government forces (e.g. arrests, use of tear gas). Finally, conflict refers to acts of violence that can be perpetrated by protesters or government authorities (e.g. attacks, beatings). As the top row of panels in Figure 1 shows, the NYT generates substantially fewer events than the number of events coded from local newspapers as reflected in the lower row of panels. Moreover, there are discrepancies not only in the number of events, but also in the types of actions (dissent, repression, and conflict).

[Insert Figure 1]

Disaggregating the types of events by specific actions reveals further discrepancies between event data generated using English-based information and local sources in their native language. Figure 2 reports the tactical repertoire by country and information source. The top row shows the specific tactics gathered from the NYT, and the bottom row shows the repertoire reported in local media. In general, the NYT reports a narrower tactical menu than the types of actions mentioned in local newspapers. The tactical imbalance across information sources is more pronounced in Mexico and in France, while there are less sharp differences in Brazil and Ukraine.

[Insert Figure 2]

The problem of over-aggregation in English-based news stories also affects the precision of the information with respect to the geographic location of the events. Figure 3 reports the number of geo-referenced events at the municipal level. Results report considerable discrepancies in the geographic accuracy of the event data generated from the NYT when compared to the events generated from local newspapers. This problem is particularly acute in Mexico where the NTY exclusively reports events in Mexico City, whereas the local newspaper indicates that protests spread throughout the country. Data from Ukraine presents a similar,

yet less marked, pattern of geographic truncation. Events from Brazil and France seem to have fewer problems of geographic inaccuracy.

[Insert Figure 3]

Finally, Figure 4 disaggregates the types of interactions between the source and target by reporting sub-actor level information. In each country, the top panel presents the different event categories derived from local newspapers and indicates the frequency of interactions between the source (labels at the left) and the target (labels at the bottom). The bottom panel presents the frequency of interactions between source and target extracted from the NYT. In general, results show that local newspapers provide more fine-grained information about the specific perpetrators and targets of different action categories than the information generated from English-based sources.

[Insert Figure 4]

Discussion

Gathering, processing, and analyzing accurate data is of paramount importance for both academic and policy communities, especially when related to conflict processes across the world. The quality of information determines the extent to which scholars are capable of analyzing the characteristics and etiology of conflict, and helps to inform policy decision makers on how to monitor, prevent, resolve, or mitigate the effects of political turmoil in foreign locations. Unfortunately, most large-scale efforts of generating event data at a global scale using machine coding protocols exclusively rely on English language media, thus neglecting massive and highly detailed information from local sources in their native languages. As this study shows, such methodological decisions are highly consequential for the quality of information generated in the coding effort. Results show that generating event data on conflict processes in foreign locations using English-language sources is likely to generate considerable

problems of under-reporting, over-aggregation, and inaccuracy when compared to event data generated from local information sources in their native language. These problems affect different dimensions of accuracy including the precision of detecting perpetrators and targets, the fidelity of the tactical repertoire used by the actors, and the accuracy of the geographic locations identified in the data.

Warnings about the lack of accuracy of international databases on conflict generated using English-language media are not new (Herkenrath and Knoll 2011). However, such critiques lacked a methodological solution capable of overcoming the limitations and costs of manually coding large volumes of information. As this article shows, Eventus ID offers a technical tool capable of accurately coding event data from unstructured news reports written in multiple languages including Spanish, French, Portuguese, Russian and English. This software opens the possibility of taking advantage of the massive availability of local sources in different languages, and using their information for launching scalable projects of computerized event coding capable of generating high quality data.

Future technological innovations in natural language processing are likely to expand the possibilities of coding event data in an increasing number of languages. As technology evolves, researchers should follow the direction set by Petrarch for the use of full syntactical analysis instead of sparse parsing (Schrodt, Beieler and Idris 2014). Such developments would allow researchers to take full advantage of highly detailed information contained in local news reports, while overcoming the grammatical complexities of managing multiple languages. However, the challenges of expanding the technological capabilities of processing event data in multiple languages are not circumscribed to their technical implementation. Future research in multi-language event coding will have to face the challenge of expanding the ontologies of actors and actions beyond already vast dictionaries such as CAMEO (Gerner et al. 2002). Such ontological expansion will be crucial for capturing the enormous complexity of conflict behavior depicted in highly detailed local narratives.

References

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

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Cole, Peter. 1987. "Null Objects in Universal Grammar." *Linguistic Inquiry* 18(4):597–612.

Du, Mian, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva and Roman Yangarber. 2013. Building Support Tools for Russian-Language Information Extraction. In *4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. Sofia, Bulgaria, August 8-9: pp. 380–387.

Gerner, Deborah, Philip A. Schrodt, Omur Yilmaz and Rajaa Abu-Jabr. 2002. The Creation of CAMEO (Conflict and Mediation Event Observations): An Event Data Framdwork for a Post Cold War World. In *Prepared for delivery at the 2002 Annual Meeting of the Americal Political Science Association*. 01/10/2013.
URL: *http://web.ku.edu/ keds/papers.dir/Gerner.APSA.02.pdf*

Guízar García, Elizabeth. 2004. El uso de los verbos en los titulares de cinco diarios de la ciudad de México: análisis sintáctico PhD thesis Universidad Nacional Autónoma de México Mexico: .

Hanna, Alexander. 2014. "Developing a System for the Automated Coding of Protest Event Data.".
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425232*

Haspelmath, Martin. 2001. The European Linguistic Area: Standard Average European. In *Language Typology and Language Universals. An International Handbook. Vol 2.*, ed. Martin Haspelmath, Ekkehard Konig, Wulf Oesterreicher and Wolfgang Raible. New York: pp. 1492–1510.

Herkenrath, M. and a. Knoll. 2011. "Protest events in international press coverage: An empirical critique of cross-national conflict databases." *International Journal of Comparative Sociology* 52(3):163–180.

Huang, C.T. James. 1984. "On the distribution and reference of empty pronouns." *Linguistic Inquiry* 10(1):531–574.

Leetaru, Kalev and Philip A. Schrodt. 2013. "GDELT: Global Data on Events, Location and Tone, 1979-2012.".
URL: *http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf*

O'Brien, Sean P. 2012. A multi-method approach for near real time conflict and crisis early warning. In *Handbook of Computational Approaches to Counterterrorism*, ed. V. S. Subrahmanian. New York: Springer pp. 401–418.

Osorio, Javier and Alejandro Reyes. 2016. "Supervised Event Coding From Text Written in Spanish: Introducing Eventus ID." *Social Science Computer Review* p. online first.
URL: *http://ssc.sagepub.com/content/early/2016/01/07/0894439315625475.abstract*

Pivorvarova, Lidia, Mian Du and Roman Yangarber. 2013. Adapting the PULS Event Extraction Framework to Analyze Russian Text. In *4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. Sofia, Bulgaria, August 8-9: pp. 100–109.

Schrodt, Philip A. 1994. "The statistical characteristics of event data." *International Interactions* 20(1-2):35–53.

Schrodt, Philip A. 2009. "TABARI. Textual Analysis by Augmented Replacement Instructions.".
URL: *http://eventdata.parusanalytics.com/software.dir/tabari.html*

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

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. In *International Studies Association*. Toronto: .
URL: *http://parusanalytics.com/eventdata/papers.dir/Schrodt-Beieler-Idris-ISA14.pdf*

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

The Stanford Natural Language Processing Group. 2014. "Stanford Named Entity Recognizer.".
URL: *http://nlp.stanford.edu/software/CRF-NER.shtml*

Table 1: Number of news stories by country and source

	Mexico	Brazil	France	Ukraine
Local newspaper	136	60	340	207
The New York Times	16	25	32	116
Under coverage	88%	58%	91%	44%

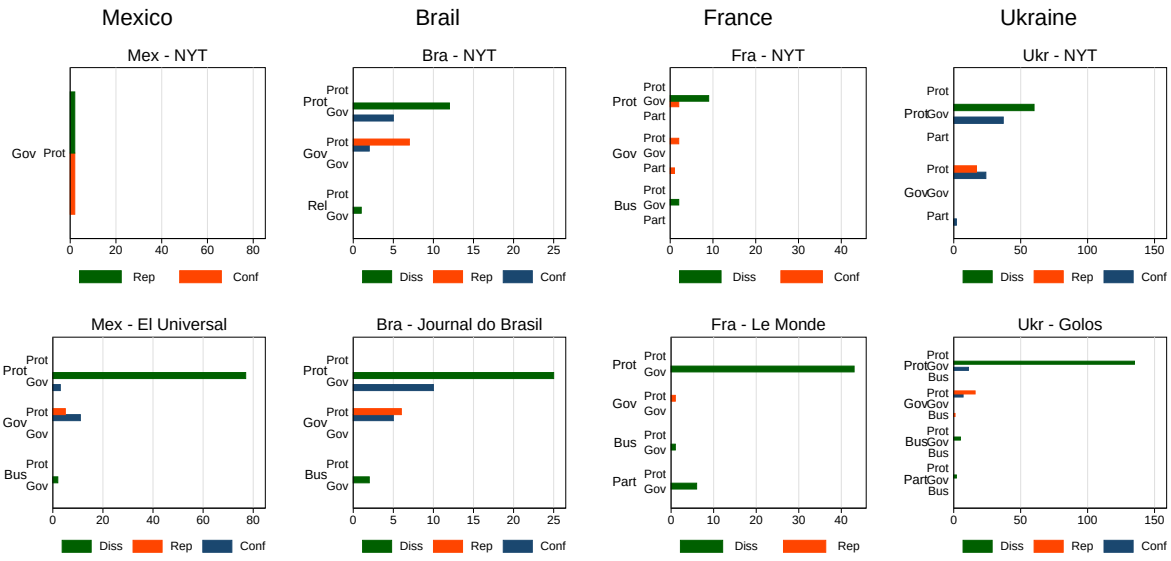

Figure 1: Number of events per country and newspaper

Note: The terms "Diss," "Rep," and "Conf" refer to action types of dissent, repression, and conflict, respectively. The left-most actor on the vertical axis of each panel represents the source of the action, and the right-most actor indicates the target of the event. To indicate the actors, terms "Gov," "Prot," "Bus," "Part," relate to government, protesters, business, and parties, respectively.

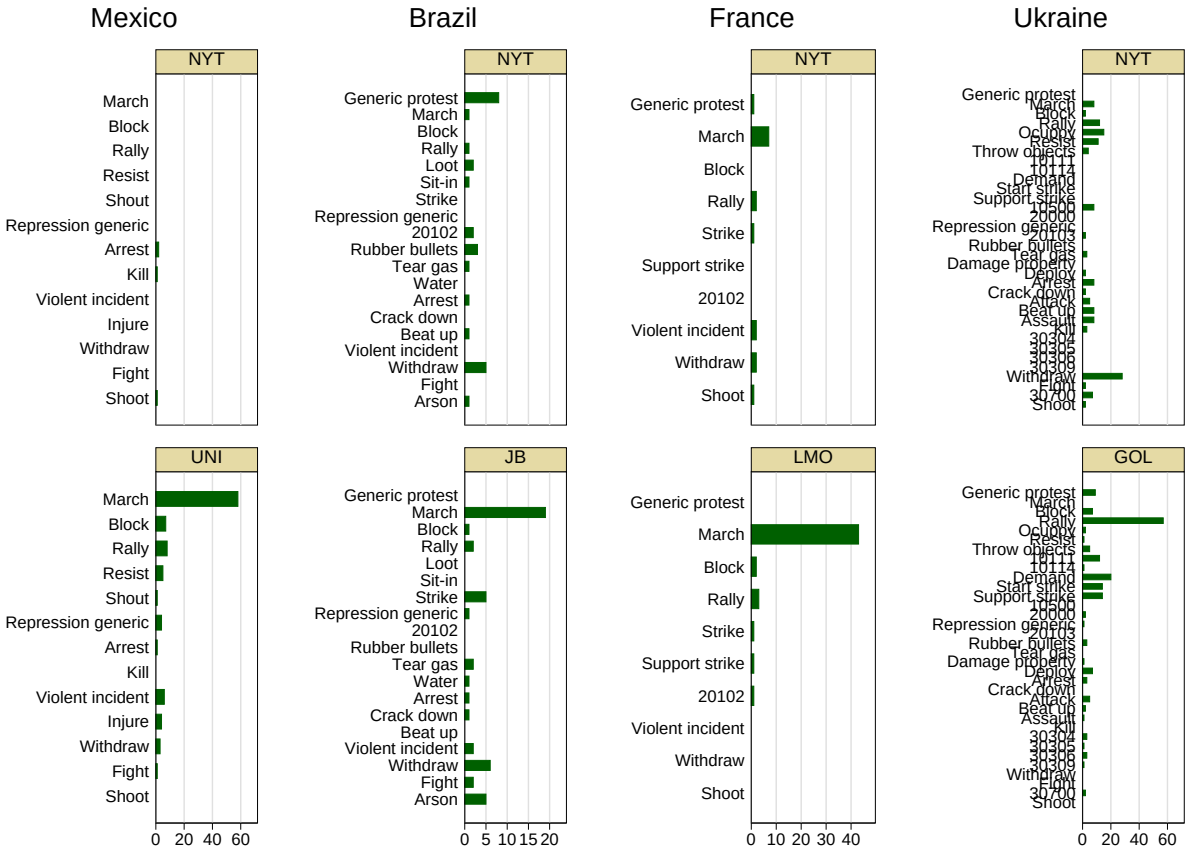

Figure 2: Tactical repertoire per country and newspaper

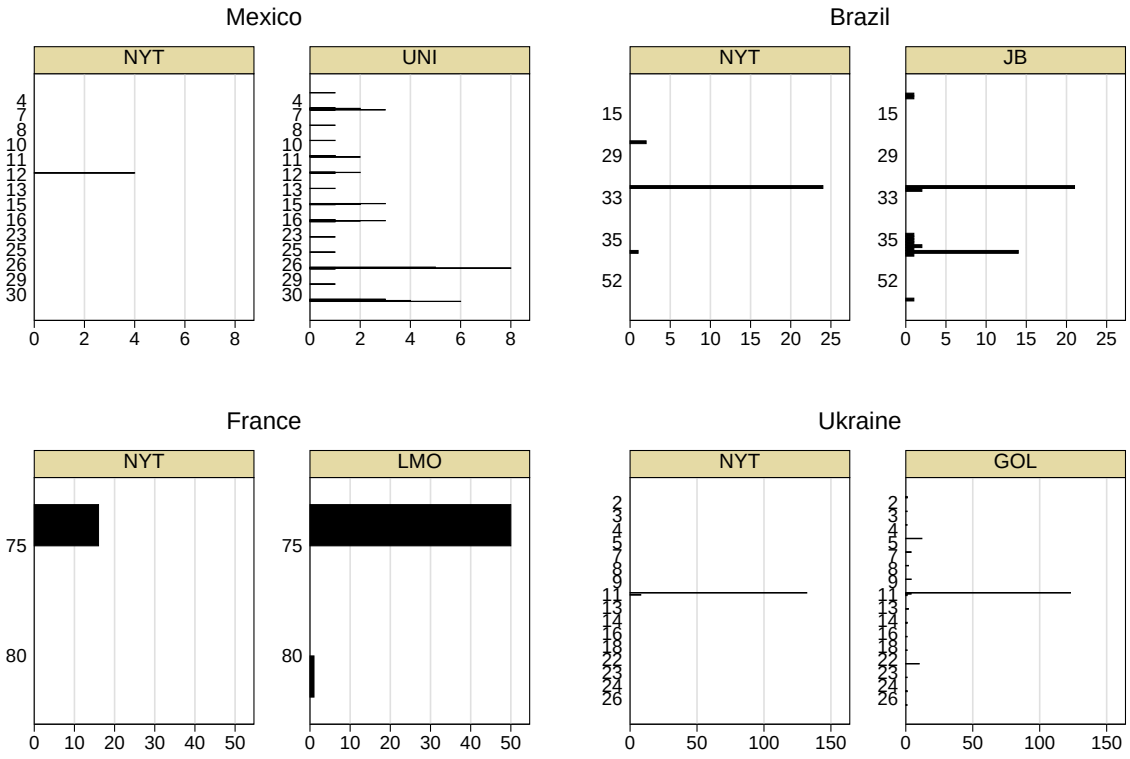

Figure 3: Municipal location of events per country and newspaper

Note: Labels represent the numeric code of the states by country. Municipal level labels are omitted from the graphs. Future versions of the paper will present maps with hot-spots instead of bar plots.

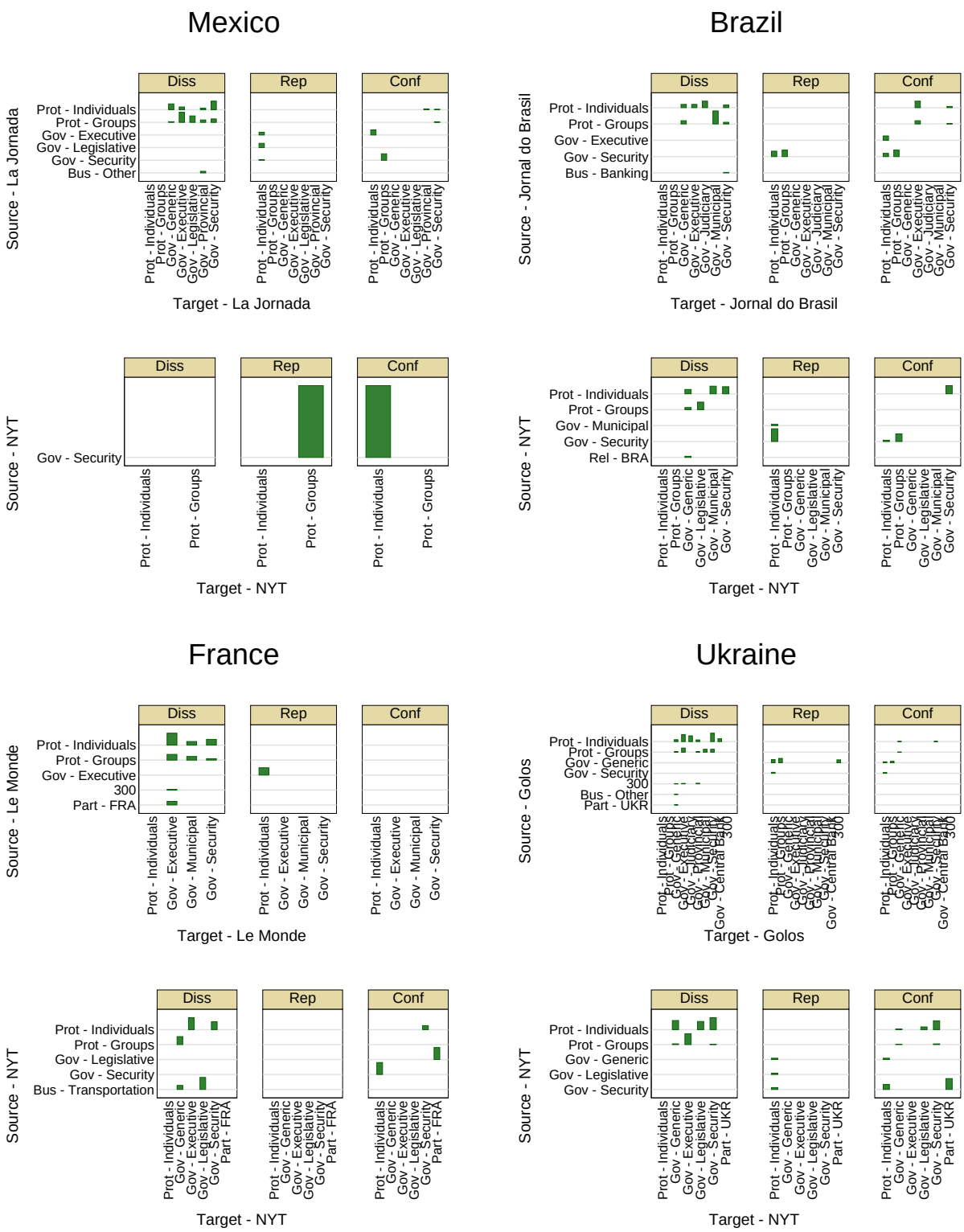

Figure 4: Subactor interactions per country and newspaper