Supervised Event Coding From Text Written in Spanish: Introducing Eventus ID with Extended Application

Javier Osorio*† & Alejandro Reyes‡

April 2015

Abstract

Recent advancements at the intersection of conflict research and computer science favored innovative approaches for generating event data using automated coding protocols. Unfortunately, these approaches almost exclusively rely on English-language sources, thus causing problems of coverage bias, which may lead to biased inferences. In an effort to attenuate the problems of Anglocentrism in computer-generated event databases, this paper introduces Eventus ID, a new software for supervised coding of event data from text written in Spanish. The innovations of this software include two event coding algorithms that better adapt to the complexities of Spanish language and a fully integrated event geo-referencing algorithm. These features allow the software to identify who did what to whom, when and where from news reports written in Spanish. Drawing on real press releases issued by the Mexican Army, this manuscript shows an application of Eventus ID by generating event data on military efforts to fight drug trafficking organizations in Mexico.

Keywords: Event data, automated coding, Spanish language, drug violence, Mexico.

Word count: 10,364 words

*This work was supported by the National Science Foundation [SES-1123572]; the Social Science Research Council-Open Society Foundations; the United States Institute of Peace; the Kellogg Institute at the University of Notre Dame; the Harry Frank Guggenheim Foundation; the Program on Order, Conflict and Violence at Yale University; and the Mario Einaudi Center for International Studies at Cornell University.

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

‡MS in Computer Science, Instituto Nacional de Astrofísica Óptica y Electrónica

Introduction

Recent advancements at the intersection of conflict research and computer science have favored the development of novel approaches for generating event data from news reports (Schrodt, Beieler and Idris 2014, Schrodt 2009, Leetaru and Schrodt 2013, Hanna 2014, Bond et al. 2003, Stepinski, Stoll and Subramanian 2006). The availability of news digital media paired with technological innovations opened the possibility of coding massive volumes of information and expanding our understanding about conflict processes (Schrodt 2012*b*, Subrahmanian 2013, O'Brien 2012, Bond et al. 2003). Despite their global scope, a common denominator of major automated event coding projects is the primarily reliance on information gathered from English-language news sources about events occurring in foreign locations. Most of these databases use news stories from the *New York Times*, the *Wall Street Journal*, the *BBC Summary of World Broadcasts*, and the *CIA Foreign Broadcast Information Service*. However, ignoring reports written in the native language of the locations under study is likely to induce problems of coverage bias, reduce the quality of information, delay timely news and ultimately degrade the accuracy of the event data outcome. This is particularly problematic if we consider that 91.8 percent of the world population are non–English speakers (Lewis, Simons and Fennig 2013). Inferences drawn from a biased, incomplete, distorted or inaccurately recorded sequences of events may generate misleading conclusions concerning the dynamics of conflict.

In an effort to attenuate excessive reliance on English-language sources in computer-generated data, this paper introduces Eventus ID, a software for supervised coding of event data extracted from unstructured text written in Spanish.[1] Eventus ID grows from a long tradition of automated event coding in conflict research (Schrodt, Davis and Weddle 1994, Leetaru and Schrodt 2013) and relates to recent efforts to expand natural language processing to non-English languages (The Stanford Natural Language Processing Group 2014).

[1]Note to the reviewer: if accepted, the manuscript will include a link to download the software, documentation, original press releases, data, Stata dofiles and all ancillary information mentioned in this paper.

The program is *supervised* as it requires humans for developing dictionaries used as search categories and for assessing the accuracy of the coding outcome. The software *codes event data* based on a pattern recognition system that identifies information on who did what to whom, when and where from news reports. Two event coding algorithms and an event locator protocol enable Eventus ID to code fine-grained event data at the subnational level. These features are an improvement over other supervised event coding protocols (Schrodt 2009). In addition, to the best of our knowledge, this is the first program specifically designed for coding events from *unstructured text written in Spanish*. This software aims to reduce the time and financial requirements for manually coding event data, which often represent an overwhelming burden (Baumgartner, Jones and MacLeod 1998). In this way, Eventus ID builds on the efforts of researchers manually coding conflict data in Spanish (e.g. Herkenrath and Knoll 2011), and seeks to facilitate the creation of customized databases in comparative politics and international relations relying on text written in Spanish.

There are several content analysis programs capable of conducting a variety of tasks such as counting words, concordance, lemmatization, classification or cluster analysis (Grimmer and Stewart 2013, Lowe N.d.). Eventus ID is a specific type of software designed for event coding from unstructured text. Eventus ID's direct predecessor is TABARI, the most popular software for event coding from news reports written in English (Schrodt 2009). Eventus ID shares with TABARI some basic characteristics of event coding such as the Deterministic Finite Automata approach for string matching and the use of sparse parsing of sentences rather than full syntactical analysis. However, Eventus ID and TABARI are quite different in several other respects. Eventus ID works with a set of flexible event coding algorithms that better adapt to the complexities of the Spanish language. In addition, Eventus ID is capable of identifying the geographic location of events when the information is provided in the source text. In order to facilitate coding irregular verbs in different tenses in Spanish, Eventus ID does not include a stemming application as the one used in TABARI. The lack of a stemming function reduces coding error but it requires the researcher to invest more effort

in developing detailed dictionaries. In general, the combination of these features contribute to improving the performance of Eventus ID for event coding from text written in Spanish.

The manuscript is structured in five sections. First, it discusses the challenges of event coding from text written in Spanish. Then, it offers an overview of the Eventus ID coding process. The third part addresses the technical innovations for coding and geo-locating events. The fourth illustrates the use of the software with an application on counter-narcotic efforts in Mexico. Finally, the paper addresses the limitations of Eventus ID discusses future challenges in computerized event coding.

Coding Event Data from Text Written in Spanish

An event is defined a set of categorical information providing a description of someone (*source*) doing something (*action*) to someone else (*target*) in a certain *location* on a particular *date*. The source-action-target arrangement parallels the basic grammatical structure of subject-verb-object. To detect the source and target, Eventus ID uses dictionaries of actors developed by the researcher. In addition, a dictionary of verbs serves for identifying the actions. While reading the text, Eventus ID uses the dictionaries of proper nouns and verbs provided as searching categories to recognize the actors and actions. Once these elements are detected, the program transforms the textual information into numeric format. Eventus ID also identifies the date and relies on a dictionary of locations to detect the place of occurrence.

There are two main challenges associated with coding event data from text written in Spanish. The first refers to the complexity of conjugating verb tenses in Spanish. To code events, Eventus ID uses a pattern recognition scheme similar to that implemented in TABARI. However, a key difference between these two protocols lies in the way they handle verb phrases. TABARI is designed for coding in English, whose grammatical rules allow for simple general rules of conjugating verb tenses. For example, the forms of most regular verbs in English vary only by adding "s," "ed," or "ing" to the end of the infinitive stem. Another

characteristic of English is that the gender and number of the noun do not affect the verb form (excepting only the third person singular present). This means that "you," "he," "she," "we" or "they" can be mostly combined with the different verb forms almost indistinctly. Based on these simple and general grammatical rules, TABARI incorporates a *stemming* algorithm to automatically identify all the different verb tenses from a stem (Schrodt 2009). Using the verb "to arrest" as an example, Table 1 shows that English allows the stem "arrest" to be easily conjugated for different tenses, gender and numbers by simply adding a suffix at the end of the stem. This grammatical characteristic of the English language allows TABARI to readily identify a variety of verb forms without major computational demands.

Although TABARI's stemming facility is convenient for coding in English, this feature can generate a substantial coding error in Spanish. Using an English-based stemming algorithm is not appropriate for event coding in Spanish because verb tenses in the latter do not end with "s," "ed" or "ing." Verb conjugation in Spanish is much more varied as the gender and number of the subject are integrated in the verb tense. Panel (b) in Table 1 shows that the ending part of the verb "*arrestar*" (to arrest) is very different across verb tenses, number and gender. The "shortcuts" that might be useful for coding in English would be counterproductive in Spanish as they would substantially increase the risk of coding error and the computational demands for conjugating all tenses.[2] Given the complexity of Spanish conjugation, Eventus ID does not include a stemming algorithm. Unfortunately, reducing the propensity of generating error through a stemming process comes at a cost. In order to improve the accuracy of the coding protocol, Eventus ID requires the user to develop large and detailed verb dictionaries considering a variety of verb tenses for different gender and number of subjects. Developing dictionaries is a labor intensive process that requires refinement, knowledge accumulation, and feedback from the validation process. The iteration of coding, verification and recoding allows fine-tuning the dictionaries.

[2]Panel (b) in Table 1 shows that conjugating in Spanish the verb stem "*arrest*" in only six tenses would require adding 36 different verb terminations (e.g. "*o*," "*as*," "*a*"). Notice that this only considers conjugating six of the 17 different verb tenses possible in Spanish (see Real Academia Española y Asociación de Academias de la Lengua Española 2012).

Table 1: Verb tenses in English and Spanish

Person	Indicative		Subjuntive		Gerund	Past passive voice
	Present	Past	Present	Imperfect		
(a) English						
I	arrest	arrested	arrest	arrested	arresting	was arrested
you	arrest	arrested	arrest	arrested	arresting	were arrested
he, she	arrests	arrested	arrests	arrested	arresting	were arrested
we	arrest	arrested	arrest	arrested	arresting	were arrested
you	arrest	arrested	arrest	arrested	arresting	were arrested
they	arrest	arrested	arrest	arrested	arresting	were arrested
(b) Spanish						
yo	arresto	arresté	arreste	arrestara o arrestase	arrestando	fui arrestado
tú	arrestas	arrestaste	arrestes	arrestaras o arrestases	arrestando	fuiste arrestado
ella, él, usted	arresta	arrestó	arreste	arrestara o arrestase	arrestando	fue arrestada o fue arrestado
nosotros	arrestamos	arrestamos	arrestemos	arrestáramos o arrestásemos	arrestando	fuimos arrestados
vosotros	arrestáis	arrestasteis	arrestéis	arrestarais o arrestaseis	arrestando	fuisteis arrestados
ellas, ellos, Uds.	arrestan	arrestaron	arresten	arrestaran o arrestasen	arrestando	fueron arrestadas o fueron arrestados
vos	arrestás	arrestaste	arrestés	arrestaras o arrestases	arrestando	fuiste arrestado

The second challenge refers to the frequent use of the present indicative in journalistic writing style. In general, the present indicative is used similarly in English and in Spanish. In English, the present progressive tense is a finite form of the verb that has the mood, tense, and person clearly defined. For example, in the sentence "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; thus the sentence would read "*arrestan a un criminal.*" As shown in Table 1, this conjugation of the verb "*arrestar*" (to arrest) corresponds to the third person of the present indicative.

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. What makes the analysis in Spanish more complex is that the conjugation already gives information about the person as part of the verb, and in consequence the subject of the action is often omitted from the sentence. To make things even more complex, the present indicative is often used for referring to events that occurred in the *past* (historical present). Thus while the sentence "*arrestan a un criminal*" literally refers to an action carried out in the present, it also refers figuratively to a past event. By omitting the subject from the sentence, the present indicative makes event coding more difficult. The reason is that the sentence would not contain a source, thus breaking the fully specified structure of an event as source-action-target. This is particularly problematic due to the pervasive use of the present indicative in journalistic narratives in Spanish media (Martínez, Miguel and Vázquez 2004, Alcoba Rueda 1983, Nadal Palazón 2009). For example, according to Guízar García (2004), 73 percent of news headlines in Mexico use the present indicative verb form. The ubiquity of these grammatical structures require the development of a coding protocol capable of adapting to the complexities of Spanish language.

Overview of Eventus ID Coding Process

There is a broad range of computerized textual annotation methods in political science (for reviews see Cardie and Wilkerson 2008, Grimmer and Stewart 2013, Schrodt and Gerner 2012). Fully automated methods of text analysis are capable of identifying the underlying features of textual content without the need of explicitly imposing ex-ante categories (e.g. Grimmer and King 2011). In contrast, supervised methods, as the one implemented in Eventus ID, require the researcher's intervention in several stages of the coding process, primarily in the development of dictionaries. This broad margin of intervention allows investigators to customize their research projects. As illustrated in Figure 1, the process of event identification of Eventus ID consists of six stages: information gathering, formatting the text corpus, event coding, event location, validation, and output generation.

Stage 1. Information gathering. The researcher identifies a set of documents relevant to the research project. In step 1.1, users can rely on a broad variety of RSS (Rich Site Summary) readers or web scrappers to automatically extract content form the web in a non-discriminant manner. Alternatively, researchers can rely on humans to select specific documents containing information relevant to the project and then download the selected documents. Manual selection is recommended for improving the validity and accuracy of the coding outcome. In this sense, there is usually a trade-off between the speed of non-discriminant information gathering and the direct relevance of its content to the project at hand. The output of this stage is a collection of news reports in individual files (step 1.2).

Stage 2. Corpus of text. Eventus ID comes with an ancillary software for comprising the collection individual files extracted from the web and compiling their information into a single file, named corpus, in a format readable by Eventus ID (step 2.1). The text corpus is the input to Eventus ID for event coding (step 2.2). The identification of event data is easier when conducted on small pieces of text rather than on long pieces of information. For this reason, the corpus breaks the content of each document into its component paragraphs and attaches an individual identifier per document and per paragraph.

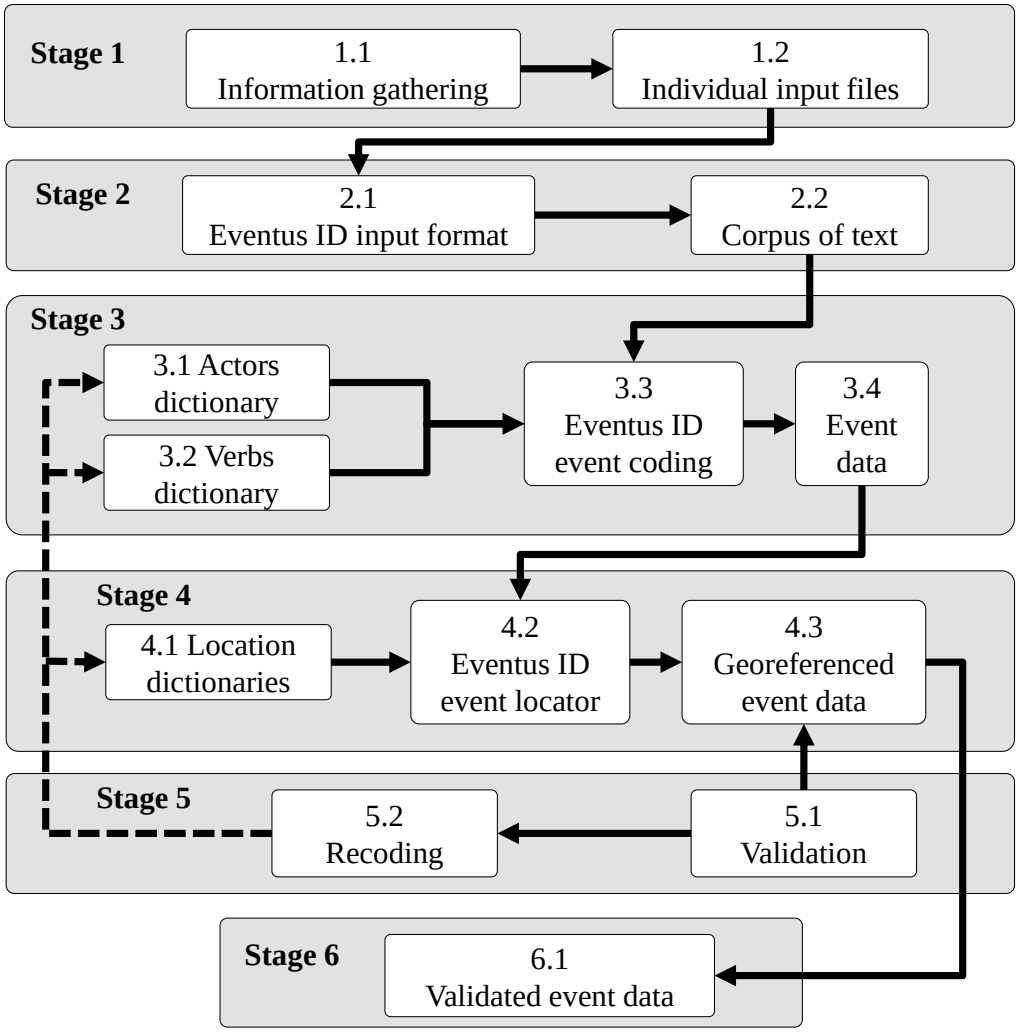


Figure 1: Eventus ID coding process

Stage 3. Event identification. The processes of event coding (stage 3) and event location (stage 4) are integrated into Eventus ID and are implemented automatically. However, for the sake of clarity, these two stages are discussed separately. In the third stage, the researcher runs Eventus ID to identify event data from the corpus. Identifying events requires researchers to develop dictionaries of actors and verbs (steps 3.1 and 3.2) that serve as search categories to find the source, action and target in the corpus. As discussed in the next section, Eventus ID identifies event data based on two event coding algorithms that better accommodate to the characteristics of the Spanish language (step 3.3). The output is a database containing event data (step 4.3).

The dictionary development assumes that the researcher is familiar with the subject of study and knows the main actors and the actions they conduct. Building specialized dictionaries is the key for customizing projects according to the needs of the investigator. The dictionaries can take into account regionalisms in Spanish language, which tend to vary substantially across–and often within–Latin American countries. However, if the researcher fails to include a relevant actor or verb in the dictionaries, then Eventus ID will not be able to identify such element in the text due to the lack of a search criteria.

To assist in the identification of actors, researchers can use Name Entity Recognition (NER) software (e.g. The Stanford Natural Language Processing Group 2014). NER is useful for automatically discovering in the corpus possible actor names that were unknown to the researcher, and for updating the actors dictionary as new entities appear. However, NER is not designed for detecting verb sentences. In this sense, researchers still have to devote substantial efforts to developing detailed actors and verbs dictionaries.

Stage 4. Event location. This stage requires the development of location dictionaries containing the names of states and municipalities to be used as search categories, along with some filters to enable disambiguation of specific cases (step 4.1). The event location algorithm uses locality names to inspect the corpus of text in order to identify the places where the events occurred (step 4.2). The output is a database of georeferenced event data (step 4.3).

Stage 5. Validation. The accuracy of the coding protocol is evaluated by comparing a sample of manually coded data with the computer-generated events. Discrepancies between manual and computer outcomes provide insights for modifying the actor, verb or location dictionaries. A series of coding-validation-recoding iterations (steps 5.1 and 5.2) are often necessary to improve the accuracy and the validity of the resulting data.

Stage 6. Output. The final output of the coding process is a validated database of georeferenced event data (step 6.1).

Coding Innovations in Eventus ID

Event Coding Algorithms

In order to code events from the text corpus, Eventus ID uses two pattern recognition algorithms: the *general sequence algorithm* codes events that comply with the full source–action–target structure, and the *partial sequence algorithm* codes incomplete events that omit the source and present a verb–target structure. The later algorithm is particularly helpful for identifying grammatical structures using the present indicative verb form. Both algorithms use 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 identify only the relevant parts of the text that correspond to an event, while the rest of the text is ignored for coding purposes.

Both coding algorithms begin by identifying the date of the event (`date`), followed by the document name and paragraph identification label from the text corpus (`FileName_P1_P2`). Each coding scheme then uses its own recognition sequence to identify events contained in the corpus. Table 2 shows the steps undertaken by each event coding algorithm. The general sequence algorithm starts coding an event by searching for the source (coded as `actor1`), then the action (coded as `verb`) and finally for the target of the action (coded as `actor2`). The partial sequence algorithm skips the source and starts coding the event by searching for the action (`verb`) and then looks for the target (`actor2`). Both searching criteria are implemented simultaneously on each paragraph of the corpus. Eventus ID then saves the outcome of the coding process in a plain text file. Each line in the outcome file contains the set of components corresponding to the coded event, separating its elements by tabs (denoted by the → symbol). As a result, the general and partial sequence algorithms respectively generate the following outcomes:

`date` → `FileName_P1_P2` → `actor1` → `verb` → `actor2`

`date` → `FileName_P1_P2` → → `verb` → `actor2`

Table 2: Eventus ID coding algorithms

General sequence algorithm	Partial sequence algorithm
Step 1) Search for the actor	Step 1) Search for the verb
- Load the actor dictionary - Start reading the text - Search for the longest actor first - When an actor is found, store as `actor1` - Pause the search where the actor is found	- Load the verb dictionary - Start reading the text - Search for the longest verb first - When a verb is found, store as `verb` - Pause the search where the verb is found
Step 2) Search for the verb	Step 2) Search for the actor
- Load the verb dictionary - Resume reading from previous pause - Keep reading the text - Search for the longest verb first - When a verb is found, store as `verb` - Pause the search where the verb is found - If no verb is found, go to Step 4	- Load the actor dictionary - Resume reading from previous pause - Keep reading the text - Search for the longest actor first - When an actor is found, store as `actor2` - Pause the search where the actor is found - If no actor is found, go to Step 3
Step 3) Search for the actor	Step 3) Save the event
- Reload the actor dictionary - Resume reading from previous pause - Keep reading the text - Search for the longest actor first - When an actor is found, store as `actor2` - Pause the search where the actor is found	- Save `[◇] [verb] [actor2]` in the database - If no actor is found, save the event as `[◇] [verb] [◇]` in the database - Start again from Step 1
Step 4) Save the event	
- Save `[actor1] [verb] [actor2]` in database - If no verb is found, save the event as `[actor1] [◇] [◇]` in the database - Start again from Step 1	
Note: the ◇ symbol represents a blank space when no element is coded.	

To illustrate how the general sequence algorithm works, consider the following sentence in both English and Spanish:[3]

`Army troops arrested a member of a criminal group.`

`Tropas del ejercito arrestaron a miembro de un grupo criminal.`

In this example, all three elements of the event (source-action-target) are present in the sentence in the required order. In consequence, the general sequence algorithm identifies

[3]Spanish text examples in this article deliberately omit accents.

"`Army troops`" as the source, "`arrested`" as the action and "`member of a criminal group`" as the target.

The general sequence algorithm can also help to identify sentences in passive voice, a commonly used verb tense in Spanish journalism. Instead of having a regular subject-verb-object grammatical structure, passive voice inverts the order of the subject and object in the sentence and yield to an object-verb-subject structure as illustrated below:

`A member of a criminal group was arrested by Army troops.`

`Miembro de un grupo criminal fue arrestado por tropas del Ejercito.`

As part of the verb dictionary development, researchers can develop special instructions for Eventus ID to identify verbs in passive voice (e.g. "`was arrested`," "*`fue arrestado`*"). Once identified, the researcher can reverse the object-verb-subject structure of passive voice sentences into a regular subject-verb-object arrangement.

Another feature of the general sequence algorithm is that it does not require the three elements of source-action-target to be present in the sentence for Eventus ID to register the event. This feature is useful when the reports include lists of nouns. For example:

`The Army seized an arsenal containing: AK-47 rifles, R-15 assault rifles, grenade launchers, ammunition and communication devices.`

`El Ejercito decomiso un arsenal que contenia: rifles AK-47, rifles de asalto R-15, lanza granadas, municiones y equipo de comunicacion.`

Eventus ID would code the source ("`Army`"), the action ("`seized`"), and the target ("`arsenal`") as one event, and then a list of nouns ("`AK-47 rifles`," "`R-15 assault rifles`," "`grenade launchers`," "`ammunition`," and "`communication devices`") in separate event lines.

The following sentence illustrates a grammatical structure that fits the coding scheme of the partial sequence algorithm. Since the translation of present indicative from Spanish into English obscures the nuances of the present indicative verb form, the next example is only presented in Spanish:

`Arrestan a un criminal.`

In this sentence, the subject of the action is omitted because of the conjugation of the verb in present indicative tense. In the absence of a first actor, Eventus ID uses the partial sequence algorithm to identify "*arrestan*" (to arrest) as the action, and "*a un criminal*" (a criminal) as the target of the event. In this way, the general and partial sequence event coding algorithms of Eventus ID adapt better to the grammatical complexities of Spanish than more rigid coding approaches.

Geo-location Algorithm

Geolocating event data is often a difficult task (Chojnacki et al. 2012). To address this challenge in the coding process, Eventus ID is capable of identifying the geographic location of event data at two sub-national levels of analysis (e.g. state and municipal) when the information is provided in the text corpus. This allows the software to generate disaggregated event data indicating who did what to whom, and referencing each specific occurrence to a spatial location. In contrast to TABARI, the geo-location feature is already integrated into Eventus ID's coding protocol. The software uses lists of states and municipalities in order to identify the locality mentioned in the original source. To develop the location dictionaries, researchers can rely on lists of toponyms provided by government agencies. Alternatively, researchers can use NER software to identify possible locations when the names of such places are unknown to the researcher (The Stanford Natural Language Processing Group 2014). In addition, due to the complexities of geo-referencing data in automated event coding (Hammond and Weidmann 2014), the event location protocol includes a filter dictionary for disambiguation. The filters prevent certain words or phrases from being erroneously classified as geographic locations. This is particularly useful to avoid confusing names of individuals (e.g. "Paz Vega") or organizations (e.g. "Asociacion por la Paz") with geographic locations (e.g. "La Paz, Baja California").

Table 3 outlines the event location procedure. Eventus ID uses the event database generated by the event coding algorithms and identifies the source paragraph from which each

event was extracted. It then reads the entire text corpus in order to identify the specific paragraph containing that event. Once the paragraph is identified, the algorithm uses the information provided by the location dictionaries to search for the name of a state or municipality in the paragraph. If a location is identified, the protocol uses the filters dictionary to verify whether the location should be kept or discarded. If the location is not filtered, the algorithm saves the geographic code next to the corresponding event codes in the database. If no location is found in the paragraph, the algorithm expands the search to the rest of the news report starting from the first paragraph of the document. If a state or municipality is recognized in the corpus, the protocol checks whether it should be filtered or not. If it passes the filter, the algorithm saves the code of the location next to its corresponding event code in the event database. The coding protocol records in the corresponding event line as many localities as they are mentioned in the paragraph. If no location is identified in the document, the protocol stops searching for the location of this event and moves to the next episode in the event coding database.

To illustrate how the event location algorithm works, consider the following paragraph:

> `In a press release issued yesterday, the Mexican government announced that troops deployed in the municipality of San Luis Rio Colorado, Son. seized 227 packages of marijuana with a total weight of two tonnes and 250 kilograms while patrolling rural roads in the area.`
>
> *En un comunicado de prensa emitido el dia de ayer, el gobierno mexicano informo que tropas destacamentadas en el municipio de San Luis Rio Colorado, Son. decomisaron paquetes de mariguana con un peso total de dos toneladas y 250 gramos, mientras patrullaban caminos rurales del area.*

Given the right set of actors, verbs and location dictionaries, Eventus ID recognizes the key components of the event as "`troops`" (source), "`seized`" (action) and "`packages of marijuana`" (target). In addition, the location algorithm identifies the municipality "`San`

Table 3: Eventus ID event location algorithm

1) Identify an event
- Load the event database. - Select an event from a new line. - Identify the paragraph (`FileName_P1_P2`) from which the event was extracted. - Use the entire paragraph name as searching criteria.
2) Identify the paragraph in the text corpus.
- Load the text corpus. - Search for the paragraph from which the event was extracted.
3) Search for the location of the event in the paragraph.
- Load the location dictionaries (states and municipalities). - Use the items of the location dictionaries as searching criteria. - Start searching for the location in the source paragraph. - If the location is found, store the code. - Keep searching for locations and storing them until the end of the paragraph. - If there are no more locations in the paragraph, then go to Step 5. - If no location is found in the paragraph, go to Step 4.
4) Expand the search to the rest of the document.
- Select the remaining paragraphs belonging to the same document (`FileName`). - Search for the location in all paragraphs of the document. - Begin searching in the first paragraph. - If the location is found in the document, store the location code and go to Step 5. - If the location is not found in the document, stop searching and go to Step 1.
5) Filter the location.
- Load the filters dictionary. - Verify that the location identified does not match any item in the filters dictionary. - If the location matches a filter, go back to Step 3. - If the location does not match a filter, go to Step 6.
6) Save the location.
- Save the location at the end of the coded event line in the event database. - Start again from Step 1.

`Luis Rio Colorado`" in the state of "`Sonora`" (Son.) as the location where the event took place. In this way, Eventus ID produces a full account of event data indicating who, did what, to whom, when and where.

Application

Event data

Eventus ID is a generic supervised event coding protocol developed for assisting researchers in generating customized event data on their own topics of interest. To illustrate the use of the software, this section presents a particular application for generating a database of counter-narcotic efforts conducted by the Mexican Army (Secretaría de la Defensa Nacional, SEDENA). The application relies on military press releases issued between December 2012 and January 2013. The decision to use documents from the Mexican Army is based on the availability of press releases that allow replicating this example without inflicting copyrights. Schrodt (2014) warns that a central challenge for the transparency and validity of automated coding pertains to the intellectual property rights that limit researchers from sharing copyrighted materials. The lack of original reports inhibits the possibility of testing computerized coding protocols using real corpuses. This situation often forces scholars to use a fake corpus or not to release any text corpus at all. To address this issue, this application has the permission of SEDENA to use real press releases as corpus of text.[4]

The development of encompassing and detailed dictionaries is crucial for improving the quality of the coding outcome. In this sense, supervised event coding is not a panacea. It requires researchers to devote time and attention for developing customized dictionaries for their own research (Best, Carpino and Crescenzi 2013). Depending on the nature of their project, researchers might find it useful to rely on pre-existing dictionaries such as the Conflict and Mediation Event Observations (CAMEO) framework (Gerner et al. 2002, Schrodt 2012*a*). In other cases, researchers might require the assistance of NER software to identify possible names of actors (The Stanford Natural Language Processing Group 2014). In this particular application, the actors dictionary consists of a list of 140 proper nouns. This dictionary is used for identifying both the source and target in the text corpus. Actors are organized in six

[4]Note to the reviewer: if accepted, the article will comply with the requirements of the Mexican Army to include an Appendix providing the citations of the press releases used in the corpus of text.

categories including: military personnel; unidentified individuals; criminals; drugs; weapons; criminal assets (e.g. vehicles, real estate); and others (e.g. stolen gasoline barrels). The verbs dictionary contains a list of 26 verb phrases organized in four action categories: arrests; seizures; destruction; and rescue. The location dictionaries comprise the official names of all states and municipalities (Instituto Nacional de Estadística y Geografía 2011*a*) and a list of filters for geographic disambiguation. Eventus ID uses these different dictionaries to identify actions conducted by the Mexican Army against drug trafficking organizations.

Coding accuracy is a central concern in computerized textual annotation (Schrodt and Gerner 1994, King and Lowe 2003, Grimmer and Stewart 2013). To validate the precision of the computer output, a team of human coders followed Eventus ID's coding rules to code the entire corpus and identified a total of 273 geo-referenced events. Human coding thus serves as the "gold standard" to cross-check the correctness of machine data, a regular practice in validating textual annotation (Cardie and Wilkerson 2008). After an iterative processes of event coding, assessment, elimination of duplicates, disambiguation and recoding,[5] the final computer outcome identified a total of 281 geo-referenced events at the daily municipal level.

Panel (a) in Table 4 shows the extent of congruence between the manual and computerized event data. Both protocols agree in 255 events. In contrast, the automated procedure failed to identify 18 events that were detected in the manual coding process. In addition, the computerized outcome generated 26 events that were not previously identified by human coders. A careful inspection of the latter reveals that most of these events were duplicated locations caused by news stories mentioning more than one municipality. Further refinement of the location dictionaries could help to improve the accuracy of the data.

Following the standards proposed by Cardie and Wilkerson (2008), Panel (b) in Table 4 reports the performance evaluation metrics of the computerized textual annotation. The criteria for correctness uses an astringent "necessary and sufficient" condition (Goertz 2005), in which an event is considered correctly coded if all the elements of the computer-generated

[5]This process relates to stage 5 in Figure 1 and feedback loops in stages 3 and 4.

Table 4: Validity assessment of the automated coding protocol

(a) Number of events

		Eventus ID		
		True	False	*Total*
Human	True	255	18	*273*
	False	26		
	Total	*281*		

(b) Precision assessment

Assessment	Percentage
Accuracy	90.7%
Recall	93.4%
False positives	9.3%
False negatives	6.6%

event match the source, action, target, location (at both the state and municipal levels), and the date identified by human coders. *Accuracy* represents the percentage of events proposed by the system that are correctly identified in comparison to the manually coded database. The result shows that the accuracy of Eventus ID is up to 90.7 percent when compared to human coders. *Recall* is the percentage of the manually classified events that are correctly identified by the system, which amounts to 93.4 percent.[6] The category of *false positives* corresponds to 9.3 percent of events proposed by the system that are not identified by humans. Finally, *false negatives* refer to the percentage of events classified by humans that are not detected by the system, which is only 6.6 percent.

As indicated by Cardie and Wilkerson (2008), there is no consensus in the standards of computerized coding accuracy. At minimum, automated coding should perform better than random chance at 50 percent of precision. At maximum, computer-generated events should match human coders at 100 percent. This application of Eventus ID is close to the upper end of the accuracy scale, thus suggesting that the program is almost as precise as humans for coding events. In addition, the application presented in this manuscript performs better than the first database generated with Eventus ID (beta version), which reports 82 percent of accuracy (Anonymous 2013*a*). Moreover, this application of Eventus ID has a higher level of precision than the 75-85 percent reported in databases using TABARI (Schrodt and Gerner 1994, Schrodt 2009, Best, Carpino and Crescenzi 2013) and is as accurate as The Reader

[6]Consider H as the total number of events identified by humans and M as the total number of machine-generated events. Denote C as the total number of events correctly identified by the computer when compared to human coding. The accuracy (A) metric is $A = \frac{C}{M}$ and recall (R) is $R = \frac{C}{H}$.

(King and Lowe 2003). Besides its precision, it is important to mention that Eventus ID coded the entire corpus of 102 kilobytes in only 25.16 seconds, just a fraction of the several hours that took humans to manually code the corpus.[7]

As expected from the nature of the press releases, the Army is the only source actor in all the events. Focusing on the types of actions conducted by the Army, Figure 2 reveals that the large majority of the activity pertains to seizures (90.2 percent). The rest of the actions relate to arrests (7.9 percent), the destruction of drug plantations (1.4 percent) and one event where the Army rescued a hostage (0.5 percent). It is relevant to notice that none of the press releases provide information of the Army using lethal force. Probably, the Mexican Army did not engage in any violent confrontation with DTOs during the period under study. However, due to allegations of human rights violations perpetrated by the Army (Daly, Heinle and Shirk 2012) and political incentives to distort official information (Andreas 2010), the objectivity of these press releases might be called into question. In any case, that is a concern of bias in the information source, not of accuracy in the computerized coding procedure.

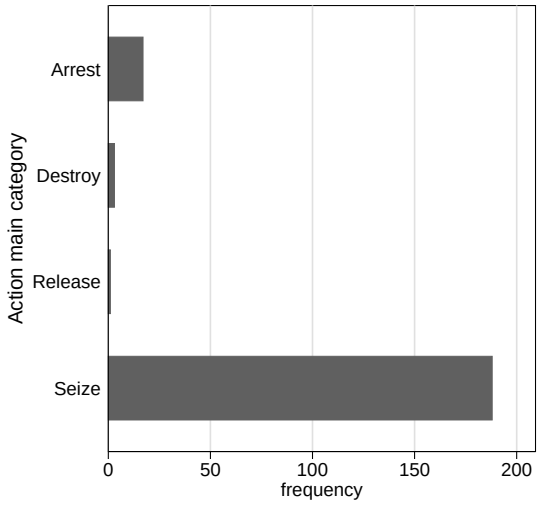


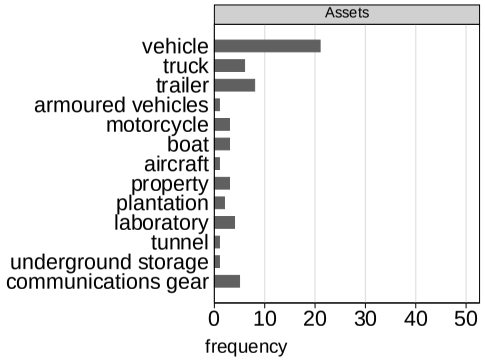
Figure 2: Type of actions conducted by the Army

Eventus ID is capable of identifying even further details about the actions conducted by the Army. For example, Figure 3 shows the specific items seized in military operations.

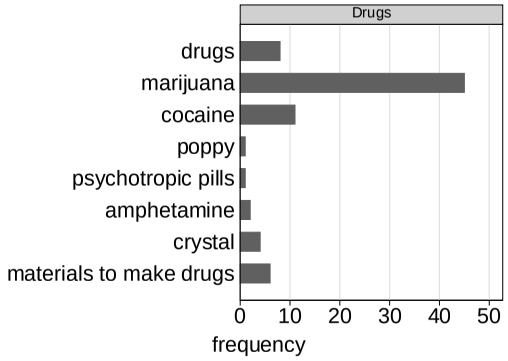
[7]This comparison does not consider the considerable time invested in dictionary development, a task necessary for both human and computerized coding.

Panel (a) indicates that the majority of confiscated assets correspond to a broad variety of terrestrial, aerial, and maritime vehicles (72.9 percent). The remaining assets refer to drug plantations, laboratories, and other types of properties (27.12 percent). The category of drug interdiction, presented in Panel (b), shows that marijuana is the most frequently confiscated drug (57.7 percent), followed by cocaine (14.1 percent). Some interdiction efforts manage to stop the transit of other types of drugs (poppy, psychotropic pills, amphetamines, crystal and generic drugs) and materials used for producing them (28.2 percent). Panel (c) reveals that criminal organizations have at their disposal a variety of weapons including small guns and assault weapons (47.1 percent), as well as ammunition of different calibers (33.3 percent). Criminal weaponry also includes grenades and grenade launchers (17.6 percent). Finally, the category "Other" in Panel (d) refers to a few episodes where authorities confiscated miscellaneous materials and stolen gasoline barrels.

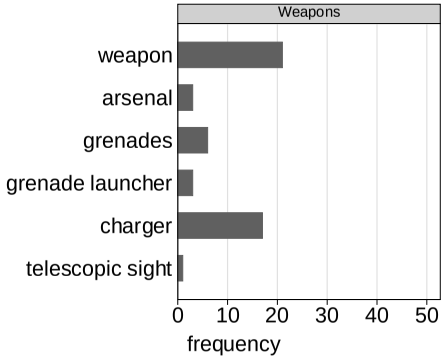
Finally, the geo-referencing algorithm identified the locus of events at two sub-national levels (states and municipalities), thus providing a high degree of location accuracy. The spatial analysis indicates that the Army conducted operations to disrupt criminal organizations in fifteen states during the period of analysis. About half of the events occurred in three states: Tamaulipas (30.7 percent), Veracruz (10.7 percent) and Sinaloa (9.8 percent). The remaining 48.8 percent of the episodes were located in 12 other states. Panel (e) in Figure 3 provides a finer-grained location of events at the municipal level. The data shows that about a third of all anti-drug activities concentrated in three municipalities: Reynosa (15.8 percent), Durango (8.37 percent), and Culiacán (6.5 percent). The rest of the events are distributed across the other 34 municipalities.



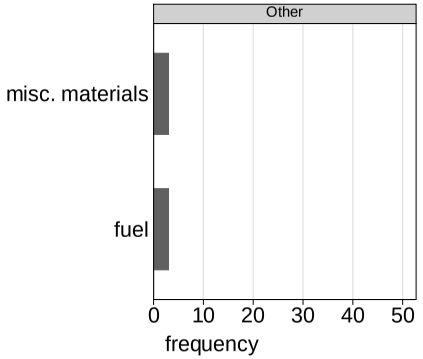
(a) Seizure of assets



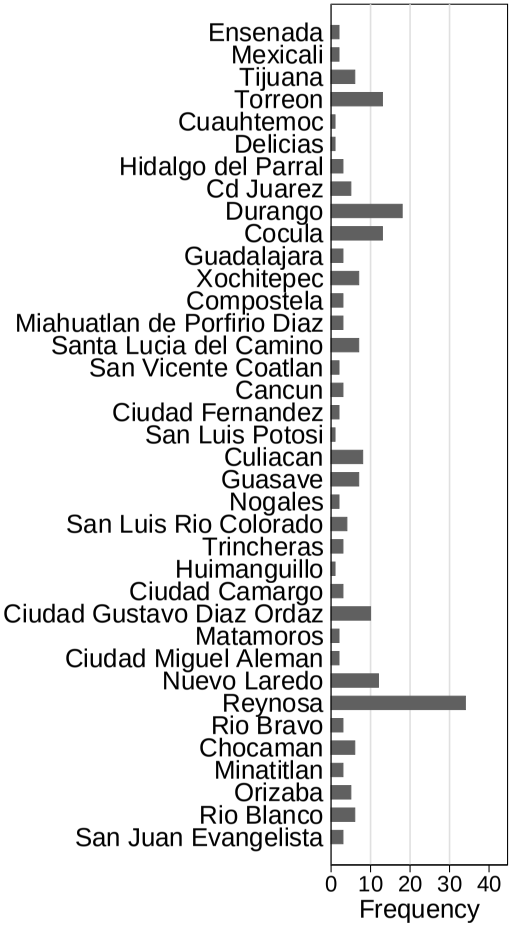
(b) Seizure of drugs



(c) Seizure of weapons



(d) Seizure of other items



(e) Event location

Figure 3: Events by type of seizure and municipal location

Statistical analysis

In recent years, the escalation of drug violence in Mexico has attracted the attention of conflict scholars (for a review see Shirk and Wallman 2014). This emerging area of research has focused primarily on analysing the levels of drug violence (Guerrero 2011, Shirk 2010, Donnelly and Shirk 2010, Snyder and Duran-Martinez 2009) by assessing the effects of law enforcement on the escalation of criminal conflict (Calderón et al. 2012, Duran-Martinez 2013, Rios 2012, Anonymous 2013*b*; 2014, Dell 2011) and the influence of international factors (Dube, Dube and Garcia-Ponce 2013, Castillo, Mejia and Restrepo 2014). Despite their contributions to understanding drug violence, this stream of research has devoted scant attention to identify the causes of law enforcement, particularly the behavior of the military in domestic drug law enforcement.

To complement this emerging literature, the database generated in this application can serve as an initial step towards understanding the determinants of counter-narcotic efforts. However, due to the limited time period analyzed here, results should not be consider as an encompassing effort to explain the full scope of law enforcement operations. Such an ambitious project escapes the objective of illustrating the use of Eventus ID.

In contrast to the vast academic literature on police behavior (for reviews see Sherman 1980, Riksheim and Chermak 1993, Klahm and Tillyer 2010), studies on the involvement of the military in law enforcement activities remain largely underdeveloped (Kraska 2007, Weiss 2011). Due to the Posse Comitatus Act of 1878 that prevents the US military from enforcing domestic laws without the approval of the President or Congress, the military has rarely been used for law enforcement in the U.S. (Balko 2013). Probably, this is one of the reasons for why the academic literature on military engagement in policing activities still is understudied. However, research on police behavior provides valuable insights. Extant explanations of law enforcement are grouped in four approaches: situational, organizational, community and individual. Situational explanations refer to the interaction between offenders and law enforcers. This relates to specific aspects of criminal behavior such as the type of offense,

characteristics of the offender, and factors related to police officers (Reiss 1968, Friedrich 1980). The organizational approach analyzes the influence of organizational structures (e.g. police department) on the behavior of individual law enforcement agents (Wilson 1968*b*, Benson, Rasmussen and Sollars 1995, Chappell, MacDonald and Manz 2006). The community level refers to the contextual characteristics (physical, economic and demographic) of the community where the interaction takes place (Sherman 1980). Finally, the individual perspective explains police actions based on the attitudes and personality of individual law enforcement agents (Niederhoffer 1969, Wilson 1968*a*).

To analyze the data, the events generated by Eventus ID are reformatted into a panel database comprising all Mexican municipalities (N=2,457) between December 1st, 2012 and January 31st, 2013 (T=62 days). Events are aggregated at the daily-municipal level according to the source–action–target–date–location structure. There are seven event categories referring to the number of *arrests*, seizures of *assets*, *drugs*, *weapons*, and *other items*, the *destruction of assets*, and the *liberation of hostages*. The analysis also considers a measure of *composite events*. This variable counts the concurrence of multiple of the above mentioned event categories that occur in the same day and location. This helps to distinguish manifold episodes where, for example, authorities discover a drug lab, seize drugs and weapons, and arrest a person, from simpler incidents such as an isolated drug seizure. Finally, the *event index* is an additive measure aggregating the number of events across all categories that occurred in a municipality-day. This variable can be interpreted as the intensity of law enforcement. These metrics of counter-narcotic efforts are used as dependent variables. Table 5 reports their descriptive statistics.

In line with extant explanations of law enforcement behavior, the statistical assessment examines several independent variables. To capture situational characteristics, the analysis considers a variety of criminal offenses such as the number of *property crimes*, *homicides*, *injuries*, *kidnapping*, *theft*, *rape*, and *other crimes* at the state-month level (Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública 2014). To assess community charac-

Table 5: Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Arrests	152,272	0.00014	0.01538	0	3
Seizure of assets	152,272	0.00041	0.03221	0	4
Seizure of drugs	152,272	0.00070	0.04780	0	8
Seizure of weapons	152,272	0.00052	0.05119	0	7
Destruction of assets	152,272	0.00003	0.00628	0	2
Seizure of other items	152,272	0.00005	0.00850	0	2
Hostage release	152,272	0.00001	0.00256	0	1
Composite events	152,272	0.00079	0.04988	0	5
Event index	152,272	0.00185	0.12886	0	20

teristics, the analysis includes socioeconomic factors such as the quarterly *unemployment* rate, *population* size (INEGI 2011*b*) and levels of *poverty* (CONEVAL 2012) at the municipal level. The general criminal environment is assessed through the *number of DTOs*, a variable representing the density of criminal groups at the municipal level in 2010 (Anonymous 2014).

The community analysis also includes geographic factors that might attract law enforcement activities such as *drug production* areas; territories favorable for the reception of offshore drugs shipments along the *Gulf* and *Pacific* coast; and the string of municipalities located along the *North* border (between Mexico and the U.S.), which is favorable for international drug distribution (Anonymous 2013*a*). In addition, *local drug markets* measures the number of hospitalization cases caused by the use of illicit narcotics and serves as a proxy for municipal drug consumption (Secretaría de Salud 2012). Finally, to account for organizational characteristics, the analysis uses a measure of *corruption* that reflects the percentage of the state population who reported paying a bribe to avoid being arrested by the police (Transparencia Mexicana 2012). Since this variable measures police corruption rather than the integrity of military personnel, this factor should be considered cautiously. Unfortunately, there is no specific data at the individual level to measure the personality and attitudes of military personnel participating in counter-narcotic operations in Mexico. For this reason, the statistical analysis includes no measures related to the individual approach.

Table 6: Determinants of counter-narcotic military efforts

		Seizures of			Composite	Event
	Arrests	Assets	Drugs	Weapons	Events	Index
	(Model 1)	(Model 2)	(Model 3)	(Model 4)	(Model 5)	(Model 6)
Property crimes	0.00	0.00*	0.00	0.00	0.00*	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Homicides	-0.01	-0.01	-0.01	0.00	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Injuries	0.00**	0.00***	0.00**	0.00	0.00**	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Kidnapping	0.14***	0.14***	0.15**	0.18***	0.10***	0.23***
	(0.04)	(0.04)	(0.07)	(0.05)	(0.03)	(0.08)
Theft	-0.00**	-0.00**	-0.00	-0.00**	-0.00**	-0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Rape	-0.02	-0.02	-0.03	-0.03	-0.01	-0.03
	(0.02)	(0.03)	(0.05)	(0.03)	(0.02)	(0.05)
Other crimes	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Unemployment	0.02	0.02	0.15	0.28	-0.05	0.07
	(0.25)	(0.25)	(0.32)	(0.21)	(0.21)	(0.37)
Poverty	-0.59	-0.96*	0.52	-1.29***	-0.59	-1.97**
	(0.42)	(0.52)	(0.46)	(0.45)	(0.46)	(1.00)
Population (log)	0.93***	0.89***	1.53***	0.71***	0.86***	0.99***
	(0.18)	(0.21)	(0.35)	(0.22)	(0.22)	(0.28)
Number of DTOs	0.06	0.05	0.28***	0.04	0.01	0.02
	(0.08)	(0.09)	(0.10)	(0.11)	(0.09)	(0.13)
Drug production area	0.43	0.55	-0.12	0.28	0.60	0.80*
	(0.31)	(0.36)	(0.46)	(0.40)	(0.40)	(0.45)
Gulf area	-0.31	-0.13	-0.58	0.05	-0.10	-0.32
	(0.58)	(0.72)	(0.94)	(0.92)	(0.58)	(1.48)
North area	2.42***	2.57***	2.00*	1.73**	3.09***	2.59*
	(0.81)	(0.86)	(1.06)	(0.72)	(0.91)	(1.37)
Pacific area	0.27	0.64	0.78	1.14*	0.45	0.40
	(0.62)	(0.57)	(0.79)	(0.65)	(0.68)	(0.76)
Local drug markets	-0.00	-0.00*	-0.01	-0.00	-0.00*	-0.00
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Corruption	0.05	0.06	0.01	0.03	0.06	0.11
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.09)
Constant	-35.42***	-35.03***	-43.14***	-36.24***	-34.69***	-40.13***
	(2.36)	(2.70)	(4.53)	(2.44)	(2.67)	(3.85)
Observations	144,832	144,832	144,832	144,832	144,832	144,832

*p<0.1, ** p<0.05, *** p<0.01

Table 6 reports the results of the negative binomial regression model for panel data analyzing military efforts to fight DTOs during the period of study. The dependent variables in these models are the number of arrests (Model 1), seizures of assets (Model 2), drugs (Model 3), and weapons (Model 4), as well as the composite events measure (Model 5) and the event index of enforcement intensity (Model 6). The analysis does not consider measures for seizures of other items, destruction of assets, nor liberation of hostages as dependent variables because these factors only had one or two positive outcomes in the period under study. In general, results show that military actions seem to be explained by a combination of elements including criminal behavior, the strategic value of some territories for drug-related activities, and socioeconomic factors. Models 1-6 indicate that situational factors influence counter-narcotic operations. Kidnapping is positively associated with military actions across a broad range of law enforcement tactics. Injuries have a positive–yet marginal–effect on all military efforts, with the exception of gun seizures in Model 4 where it fails to reach statistical significance. Property crimes seem to play a marginal role in increasing seizures of assets (Model 2) and composite events (Model 5). The number of thefts shows a counter-intuitive relationship with respect to military operations in almost all models, but its magnitude is very small. Surprisingly, homicides do not seem to affect military behavior.

The results show that community factors have heterogeneous effects on military behavior. All models indicate that military actions are more frequent in highly populated areas. But economic factors show inconsistent results. Confiscation of criminal assets and weapons seems to be lower in poor municipalities than in wealthy locations (Models 2 and 4). This relationship also applies for the index of event intensity (Model 6). Unemployment does not seem to affect military operations. The criminal environment measured by the density of DTOs only seems to influence the number of drug seizures in Model 3, but no other type of law enforcement tactics. Estimates across all models indicate that counter-narcotic efforts concentrate in municipalities located along the Northern border. Areas of drug production only contribute to explain the aggregate intensity of military activities (Model 6), but do not

reach statistical significance in disaggregated measures. The number of gun seizures tends to be higher in municipalities located along the Pacific shore that in inland territories (Model 4). Contrary to the expected relationship, local drug markets seem to decrease seizures of assets (Model 2) and the composite measure of events (Model 5), yet their contribution barely larger than zero. Finally, at the organizational level, the measure of corruption shows no relationship with any counter-narcotic tactics implemented by the Army.

The disaggregated nature of the event data processed with Eventus ID allows to identify main trends in military activity, as well as the effect of some situational and community elements influencing the implementation of specific tactics. In general, military counter-narcotic efforts seems to intensify in contexts of high kidnapping and injuries rates, but with low incidence of theft, taking place in poor and densely populated communities located along the U.S.-Mexico border.

Discussion

The massive availability of global news digital media and the development of technology for automated event coding opened the possibility for understanding a broad range of conflict processes throughout the world. Accurately depicting the characteristics of specific events that conform aggregated trends of conflict is of paramount relevance for both academic and policy communities. Scholars need precise data to derive valid inferences about political phenomena, as much as the policy sector needs reliable information for decision making. Unfortunately, most efforts to understand conflict in foreign countries using automated event coding techniques have the limitation of relying almost exclusively on English-written news reports (Weller and McCubbins 2014). Neglecting valuable information produced in the native language of foreign locations is likely degrade the quality of the metrics used to analyze conflict processes and, ultimately, affect the conclusions drawn from the data. Eventus ID

contributes to ameliorate the problem of Anglo-centrism in computer generated event data by allowing researchers the possibility to code events from text written in Spanish.

As this application illustrates, Eventus ID is capable of generating an accurate categorization and geo-location of different types of events based on news reports written in Spanish. This example provides detailed event data about a variety of counter-narcotic tactics conducted by the Mexican Army against drug trafficking organizations at the daily municipal unit of analysis. Such detailed data allows making substantive contributions to an emerging area of research by identifying a set of situational and community determinants of military behavior. Moving beyond this initial application, the software offers researchers the possibility to generate their own event data on a broad variety of topics in Latin America using native information sources. This technological innovation thus broadens the opportunities to advance our understanding of political phenomena in the region.

Future developments in computerized event coding protocols should consider three main areas of opportunity. First, event coding software should be further adapted to code in a broader variety of languages. In this respect, Eventus ID would require minor adaptations of the character recognition structure to process text in Portuguese, Italian or French. After all, these Romance languages have comparable grammatical structures and primarily rely on the Latin alphabet for their written expression. Adapting Eventus ID for processing text in other non-Latin scripts such as Cyrillic or Indo-European (e.g. Greek or Indic) characters would require more advanced tuning. Finally, developing coding protocols for Arabic or Altaic (e.g. Mongolic or Japonic) languages remains an even more challenging task.

Second, computerized event coding protocols should follow the direction set by PETRARCH for the use of full parsing for event coding instead of sparse parsing textual annotation (Schrodt, Beieler and Idris 2014). In addition, event coding efforts should take further advantage of the recent innovations of the The Stanford Natural Language Processing Group (2014) for automatically identifying actors and verbs in English and in foreign languages, which would facilitate the construction of coding dictionaries.

Finally, some scholars might benefit from using ontologies such as CAMEO (Gerner et al. 2002) and IDEA (Bond et al. 2003) that comprise a large list of actors and verbs on a broad range of political topics. This "one size fits all" approach might facilitate event coding by taking an entire stream of information and processing it with a single set of dictionaries. However, due to the substantial complexity of processing unstructured news stories on a large range of topics, this approach might generate coding error and undermine the validity of the event coding output. An alternative prospective could consider a twofold task. First, develop supervised or automated protocols for categorizing the general stream of information into topical categories before the event coding process. Then, rely on experts to develop specialized dictionaries for coding events on each particular topic. Unfortunately, pre-categorizing the information and building specialized dictionaries introduces additional steps in the event coding process and might hinder the prospects of conducting real-time automated event coding and forecasting (Bond et al. 2003, Schrodt, Beieler and Idris 2014, Brandt, Freeman and Schrodt 2011). In any case, future research on machine-generated event data should balance the speed of processing massive amounts of information across languages with the accuracy of the coding output.

References

Alcoba Rueda, Santiago. 1983. "El presente de los titulares de prensa: no deíctico, protiempo anafófico.".
URL: *http://dfe.uab.es/dfeblog/salcoba/files/2008/10/presente_titulares_tiempo_anafora.pdf*

Andreas, Peter. 2010. *Sex, Drugs, and Body Counts: The Politics of Numbers in Global Crime and Conflict*. Ithaca: Cornell University Press.

Anonymous. 2013*a*.

Anonymous. 2013*b*.

Anonymous. 2014.

Balko, Radley. 2013. *Rise of the Warrior Cop*. New York: Public Affairs.

Baumgartner, Frank, Bryan Jones and Michael MacLeod. 1998. "Lessons from the Trenches: Ensuring Quality, Reliability, and Usability in the Creation of a New Data Source." *The Political Methodologist* 8(2):1–10.

Benson, Bruce L., David W. Rasmussen and David L. Sollars. 1995. "Police bureaucracies, their incentives, and the war on drugs." *Public Choice* 83(1/2):21–45.

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

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.

Brandt, Patrick, John R. Freeman and Philip A. Schrodt. 2011. "Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict." *Conflict Management and Peace Science* 28(1):41–64.

Calderón, Gabriela, Alberto Díaz-Cayeros, Beatriz Magaloni, Gustavo Robles and Jorge Olarte. 2012. "The Temporal and Spatial Dynamics of Violence in Mexico." *Working paper* .
URL: *https://www.dropbox.com/s/ce0lhfs9unsuqa3/paper_policies_violence_Aug30.pdf*

Cardie, Claire and John Wilkerson. 2008. "Text Annotation for Political Science Research." *Journal of Information Technology and Politics* 5(1):1–6.

Castillo, Juan Camilo, Daniel Mejia and Pascual Restrepo. 2014. "Scarcity without Leviathan: The Violent Effects of Cocaine Supply Shortages in the Mexican Drug War." *Working paper* .

Chappell, Allison T., John M. MacDonald and Parick W. Manz. 2006. "The Organizational Determinants of Police Arrest Decisions." *Crime and Delinquency* 52(2):287–306.

Chojnacki, Sven, Christian Ickler, Michael Spies and John Wiesel. 2012. "Event Data on Armed Conflict and Security: New Perspectives, Old Challenges, and Some Solutions." *International Interactions* 38(4):382–401.

Consejo Nacional de Evaluación de la Política de Desarrollo Social. 2012. "Medición de la Pobreza por Municipio.".
URL: *http://web.coneval.gob.mx*

Daly, By Catherine, Kimberly Heinle and David A Shirk. 2012. Armed with Impunity. Technical report Trans-Border Institute San Diego, CA.: .

Dell, Melissa. 2011. "Trafficking Networks and the Mexican Drug War." *Working paper* .

Donnelly, Robert A. and David A. Shirk. 2010. *Police and Public Security in Mexico*. San Diego, CA.: Trans-border Institute.

Dube, Arindrajit, Oeindrila Dube and Omar Garcia-Ponce. 2013. "Cross-Border Spillover: U.S. Gun Laws and Violence in Mexico." *American Political Science Review* 107(3):397–417.

Duran-Martinez, Angelica. 2013. *Criminals, Cops, and Politicians: Dynamics of Drug Violence in Colombia and Mexico*. Providence: PhD Dissertation in Political Science, Brown University.

Friedrich, Robert J. 1980. "Police Use of Force: Individuals, Situations, and Organizations." *Annals of the American Academy of Political and Social Science* 452(1):82–97.

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*

Goertz, Gary. 2005. *Social Science Concepts: A User's Guide*. Princeton, New Jersey: Princeton University Press.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* Published.

Grimmer, Justin and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *Procedings of the National Academy of Sciences of the United States of America* 108(7):2643–2650.
URL: *http://www.pnas.org/content/108/7/2643.full.pdf+html*

Guerrero, Eduardo. 2011. "La Raíz de la Violencia." *Nexos* .
URL: *http://www.nexos.com.mx/?P=leerarticulo&Article=2099328*

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: .

Hammond, Jesse and Nils B. Weidmann. 2014. "Using machine-coded event data for the micro-level study of political violence." *Research and Politics* 1(2).

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*

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.

Instituto Nacional de Estadística y Geografía. 2011*a*. "Marco Geoestadístico Nacional.".
URL: *http://www.inegi.org.mx/geo/contenidos/geoestadistica/default.aspx*

Instituto Nacional de Estadística y Geografía. 2011*b*. "Sistema Estatal y Municipal de Bases de Datos.".
URL: *http://sc.inegi.org.mx/sistemas/cobdem/*

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57(03):617–642.

Klahm, Charles F. IV and Rob Tillyer. 2010. "Understanding Police Use of Force: A Review of the Evidence." *Southwest Journal of Criminal Justice* 72(2):214–239.

Kraska, Peter B. 2007. "Militarization and policing–Its relevance to 21st century police." *Policing* 1(4):501–513.
URL: *http://policing.oxfordjournals.org/cgi/doi/10.1093/police/pam065*

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*

Lewis, M. Paul, Gary F. Simons and Charles D. Fennig. 2013. "Summary by language size.".
URL: *http://www.ethnologue.com/statistics/size*

Lowe, Will. N.d. "Software for Content Analysis – A Review.".
URL: *http://dl.conjugateprior.org/preprints/content-review.pdf*

Martínez, Francisco, Lucas Miguel and Cristian Vázquez. 2004. "La titulación en la prensa gráfica.".
URL: *http://www.perio.unlp.edu.ar/grafica1/htmls/apuntescatedra/apunte_titulacion.pdf*

Nadal Palazón, Juan. 2009. *El Discurso Ajeno en los Titulares de la Prensas Mexicana*. Mexico City: Universidad Nacional Autónoma de México.

Niederhoffer, Arthur. 1969. *Behind the Shield: The Police in Urban Society*. Garden City, NY: Anchor Books.

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.

Real Academia Española y Asociación de Academias de la Lengua Española. 2012. *Nueva gramática de la lengua española. Morfología y sintaxis.* Madrid: Espasa.

Reiss, Albert J. 1968. "Police brutality-answers to key questions." *Trans-action* 5(8):10–19.

Riksheim, Eric C. and Steven M. Chermak. 1993. "Causes of Police Behavior Revisited." *Journal of Criminal Justice* 21(4):353–382.

Rios, Viridiana. 2012. "How Government Structure Encourages Criminal Violence. The causes of Mexico's Drug War.".
URL: *http://www.gov.harvard.edu/files/Rios_PhDDissertation.pdf*

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

Schrodt, Philip A. 2012*a*. "CAMEO: Conflict and Mediation Event Observations. Event and Actor Codebook.".
URL: *http://eventdata.parusanalytics.com/cameo.dir/CAMEO.Manual.1.1b3.pdf*

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

Schrodt, Philip A. 2014. "The legal status of event data.".
URL: *http://asecondmouse.wordpress.com/2014/02/14/the-legal-status-of-event-data/*

Schrodt, Philip A. and Deborah Gerner. 1994. "Validity Assessment of a Machine- Coded Event Data Set for the Middle East, 1982-1992." *American Journal of Political Science* 38:825–854.

Schrodt, Philip A. and Deborah Gerner. 2012. Fundamentals of Machine Coding. In *Analyzing International Event Data: A Handbook of Computer-Based Techniques*.
URL: *http://parusanalytics.com/eventdata/papers.dir/AIED.Preface.pdf*

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*

Schrodt, Philip A., Shannon G. Davis and Judith L. Weddle. 1994. "Political Science: KEDS - A program for the machine coding of event data." *Social Science Computer Review* 12(4):561–587.

Secretaría de Salud. 2012. "Base de datos sobre egresos hospitalarios.".
URL: *http://www.sinais.salud.gob.mx/basesdedatos/std_ egresoshospitalarios.html*

Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública. 2014. "Incidencia Delictiva. Estadística y Herramientas de Análisis. Análisis de Información por Año, Entidad Federativa, Delito y Mes.".
URL: *http://www.estadisticadelictiva.secretariadoejecutivo.gob.mx/mondrian/testpage.jsp?query=anio_*

Sherman, Lawrence W. 1980. "Causes of Police Behavior: the Current State of Quantitative Research." *Journal of Research in Crime and Delinquency* 17(1):69–100.

Shirk, David A. 2010. Drug Violence in Mexico. Data and Analysis from 2001-2009. Technical report Trans-Border Institute, Joan B. Kroc School of Peace Studies, University of San Diego San Diego, CA.: .

Shirk, David A. and Joel Wallman. N.d. "Understanding Mexico's Drug Violence." *Working paper*. Forthcoming.

Snyder, Richard and Angelica Duran-Martinez. 2009. "Does illegality breed violence? Drug trafficking and state-sponsored protection rackets." *Crime, Law, and Social Change* 52:253–273.

Stepinski, Adam, Richard J. Stoll and Devika Subramanian. 2006. "Automated Event Coding Using Conditional Random Fields.".
URL: *http://www.cs.rice.edu/ devika/conflict/papers/draft1.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*

Transparencia Mexicana. 2012. "Índice Nacional de Corrupción y Buen Gobierno.".
URL: *http://www.tm.org.mx/indice-nacional-de-corrupcion-y-buen-gobierno-incbg/*

Weiss, T. 2011. "The blurring border between the police and the military: A debate without foundations." *Cooperation and Conflict* 46(3):396–405.

Weller, Nicholas and Kenneth McCubbins. 2014. "Raining on the Parade: Some Cautions Regarding the Global Database of Events, Language and Tone Dataset.".
URL: *http://politicalviolenceataglance.org/2014/02/20/raining-on-the-parade-some-cautions-regarding-the-global-database-of-events-language-and-tone-dataset/*

Wilson, James Q. 1968*a*. "Dilemmas of Police Administration." *Public Administration Review* 28(5):407–417.

Wilson, James Q. 1968*b*. *Varieties of police behavior*. New York: Atheneum.