

ORIGINAL ARTICLE

Automated dictionary generation for political eventcoding

Benjamin J. Radford*

LevelUp Research, LLC, Arlington, Virginia US

*Corresponding author. Email: benjamin.radford@gmail.com

(Received 30 September 2017; revised 5 June 2018; accepted 6 August 2018; first published online 27 March 2019)

Abstract

Event data provide high-resolution and high-volume information about political events and have supported a variety of research efforts across fields within and beyond political science. While these datasets are machine coded from vast amounts of raw text input, the necessary dictionaries require substantial prior knowledge and human effort to produce and update, effectively limiting the application of automated event-coding solutions to those domains for which dictionaries already exist. I introduce a novel method for generating dictionaries appropriate for event coding given only a small sample dictionary. This technique leverages recent advances in natural language processing and machine learning to reduce the prior knowledge and researcher-hours required to go from defining a new domain-of-interest to producing structured event data that describe that domain. I evaluate the method with the production of a novel event dataset on cybersecurity incidents.

Keywords: Mathematical modeling; measurement; text and content analysis

1. Introduction

Event data provide high-resolution and high-volume information about political events. Event datasets can be coded either by hand or with the aid of software, a process referred to here as “automated event coding.” While automated event coding promises reproducible, timely, and exhaustive data, several outstanding challenges limit its practical use to a subset of problems of interest for social scientists. Among these challenges is dictionary generation. Current automated event coding solutions require large dictionaries of actors, events, and event characteristics to be populated a priori such that pattern matching can be used to identify those dictionary entries in the raw text of news stories from which event data will be generated. The dictionaries are hand-coded and therefore suffer from many of the same limitations that hand-coded event datasets suffer from: they are costly to produce, require frequent updates, are not reproducible, and are vulnerable to the forgetfulness or oversight of human coders. This paper presents a novel method for generating dictionaries for event coding that ameliorates these problems. Automated dictionary generation (ADG) promises to allow researchers to rapidly generate novel datasets tailored to their research questions rather than adapting their research questions to fit existing event datasets.¹ By lowering the costs of dictionary generation, researchers will be able to adapt better existing event coding software to new domains and to iterate rapidly on their datasets.

¹While I will refer to this technique as automated, it might be better described as computer-assisted given that a minimal amount of human input is required at the outset in the form of seed terms or phrases.

This paper proceeds by first discussing existing methods for producing event data in political science. Next the ADG method itself is detailed. The paper then offers an example application of this method and introduces an event dataset on cybersecurity: CYLICON, the CYber LexICON event dataset.² This application consists of the generation and updating of verb, actor, agent, issue, and synset dictionaries. It is shown that ADG enables the expansion of automated event coding to new domains, and therefore new problem sets, with a minimal amount of researcher effort. The paper concludes with a brief discussion of directions for future research in automated event coding.

2. Event data in political science

Political event data are produced both by hand and via automated processes. Most datasets of political events are still coded manually. This process is costly, time consuming, and irreproducible. However, hand-coded event data is popular due to the perceived control it affords researchers in leveraging their expertise to code events precisely. Hand coding also allows researchers to collect information from multiple sources to construct event records with details that may not be available from any single source. Notable hand-coded event datasets include the Armed Conflict Location and Event Dataset, the International Crisis Behavior dataset, the Militarized Interstate Dispute dataset, and the Conflict and Peace Databank (Azar 1980; Brecher and Wilkenfeld 2000; Raleigh *et al.* 2010; Palmer *et al.* 2015; Brecher *et al.* 2016).

Since the mid 1990s, automated coding efforts for event datasets have grown in popularity (Schrodt 1998, 2011; Schrodt and Brackle 2013; Ward *et al.* 2013; Boschee *et al.* 2015; Caerus Associates 2015). In just the past several years, several event datasets have been introduced in political science: The Global Database of Events, Language, and Tone (GDELT), the Integrated Crisis Early Warning System (ICEWS) dataset, the Open Event Data Alliance's Phoenix dataset, and the Cline Center's Historical Phoenix Dataset (Leetaru and Schrodt 2013; Boschee *et al.* 2015; Open Event Data Alliance 2015b; Althaus *et al.* 2017).³ These datasets provide information on individual events, usually at the daily level, with specific details about the actors involved. They also often provide geographic information at a subnational level. These datasets are enormous, typically comprising millions of events.⁴

The event datasets listed above are built from streams of open-source news stories. The stories are processed through software that uses pre-defined dictionaries to infer the actors and actions they describe. Common software packages for this purpose include TABARI (Textual Analysis by Augmented Replacement Instructions) and PETRARCH (Python Engine for Text Resolution And Related Coding Hierarchy), both of which are successors to KEDS (Kansas Event Data System) (Schrodt 1998, 2011; Open Event Data Alliance 2015a).⁵ The Open Event Data Alliance, authors of PETRARCH, provide Figure 1 to illustrate their event-coding process. Raw stories are first collected from online sources. These are uploaded to a database and formatted to the specifications required by TABARI (or PETRARCH). The stories are then passed to

²The accompanying online appendix demonstrates an extension of ADG for actor-country classification. This demonstrates the effectiveness of the method in both generating and updating actor dictionaries in unsupervised and supervised settings. Performance is evaluated against existing "ground truth" data.

³UT Dallas hosts an event dataset portal at <http://eventdata.utdallas.edu/data.html>. This includes links to many variations of Phoenix including real-time and historical variants. The Open Event Data Alliance hosts an event data portal at <http://openeventdata.org/datasets.html>.

⁴GDELT, for instance, claims 103 million events as of February, 2016 (The GDELT Project 2016). ICEWS comprises nearly 15 million events. For a brief discussion of the validity of these datasets, see Wang *et al.* (2016).

⁵PETRARCH here refers to the original event-coding software to go by that name (sometimes referred to as PETRARCH 1) (Open Event Data Alliance 2015a). There are two additional event-coding software packages to go by the name PETRARCH: PETRARCH 2 and Universal Dependency PETRARCH (Norris *et al.* 2017; Open Event Data Alliance 2018). The former requires a heavily modified dictionary format; the latter relies on the modified dictionary format of PETRARCH 2 and is capable of producing event data in English, Spanish, and Arabic.

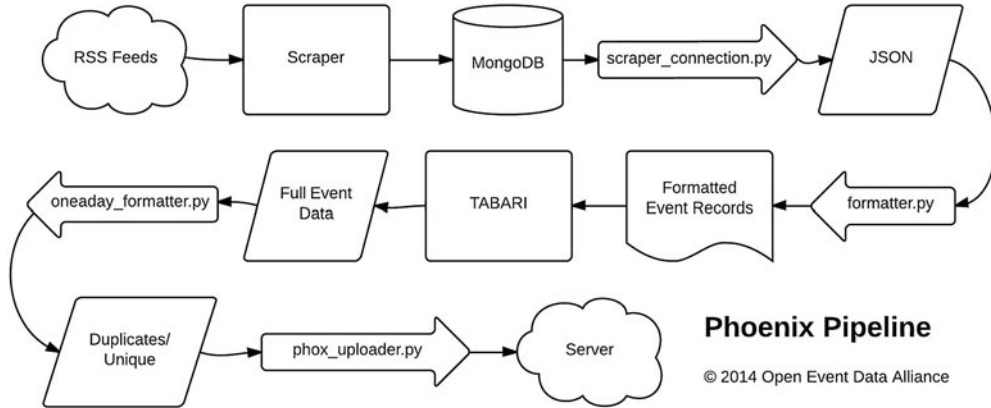


Figure 1. The Phoenix pipeline (Open Event Data Alliance 2015c).

TABARI (or PETRARCH) which uses the supplied dictionaries to produce structured data. The data are then de-duplicated using a one-a-day filter to remove multiple identically-coded event records from the same day. The resulting data are then uploaded to a server for distribution. Under ideal circumstances, human interaction is only required to select appropriate news sources, devise an ontology for the resulting structured data, and to populate the necessary dictionaries. However, this last step, dictionary creation, requires a substantial level of effort. The CAMEO verb dictionary used by PETRARCH and the Phoenix dataset is nearly 15,000 lines long and includes very specific phrases that would not necessarily be apparent to researchers a priori.⁶ The country actors dictionary, just one of multiple actor dictionaries utilized by Phoenix, is nearly 55,000 lines long. As of 2014, the ICEWS actor dictionary was over 102,000 lines long. Furthermore, as the relevant actors and language evolve, these dictionaries require regular updates to maintain up-to-date event data. Excerpts from the verb, country-actors, and synset dictionaries provided with PETRARCH are given in Table 1.

The purpose of event-coding dictionaries, like those used by TABARI and PETRARCH, is to provide an exhaustive list of the terms and phrases that map to a set of labels. In a fully automated event-coding solution, both the ontology and the dictionary could be produced without human intervention. The effort described here, however, focuses on the latter challenge: automating the process of synonym and near-synonym extraction and classification given a known ontology.

PETRARCH's dictionary structure includes a verb dictionary, three distinct actor dictionaries, an agents dictionary, an issues dictionary, and a discard dictionary. The verb dictionary categorizes verb phrases into the sets of predetermined actions described by event data. The three actor dictionaries categorize persons and named organizations by their affiliations (i.e. country, organization type) and their roles with respect to the domain of interest. These dictionaries also resolve multiple spellings or representations of an entity's name into a single canonical representation. The default PETRARCH coding scheme provides three actor dictionaries: country-affiliated actors, international actors, and non-state military actors. The agents dictionary describes how to classify unnamed entities. For example, the agents dictionary maps "thief" and "trafficker" to *criminal*. The issues dictionary identifies phrases common to the domain-of-interest to label news by topic. For example, the current Phoenix issues dictionary tags issues like *foreign aid*, *retaliation*, and *security services*. Finally, the discard dictionary identifies phrases that disqualify sentences or stories from being coded entirely. This helps to remove

⁶CAMEO, Conflict and Mediation Event Observations, is a common framework for event data and the basis for the three automated event datasets cited here (Schrodt *et al.* 2009).

Table 1. Excerpts from dictionaries supplied with PETRARCH.

CAMEO Verbs	Synsets	Country-Actors
— ABANDON [080] —	&STRONGHOLD	JOHN_FOSTER_DULLES_
ABANDON	+STRONGHOLD	[USAEI 19060101-530121]
- SAID + MUST * POLICY [100]	+BASTION	[USAGOV 530121-590422]
- * HEADQUARTERS [0874]	+CITADEL	CHRISTIAN_A_HERTER_
- * OUTPOST IN + [0874]	+BLOCKHOUSE	[USAEI 19130101-590422]
⋮	⋮	⋮
— WISH [—] —	&CEASEFIRE	HAMID_KARZAI_
WISH	+TRUCE	+KARZAI_
- * + RECOVERY [018]	+ARMISTICE	+PRESIDENT_KARZAI_

stories that might otherwise be erroneously coded. For example, sports reporting is omitted as it often uses the language of warfare to describe “victories,” “defeats,” and teams being “destroyed.”

The common CAMEO coding scheme is not a comprehensive description of public interactions between politically relevant actors and agents. For researchers interested in types of interaction that do not conform to the existing dictionary structure, the creation of new dictionaries is a necessary but costly step. The Phoenix verb dictionary contains many thousands of verbs and phrases parsed according to a particular format and organized within a predetermined ontology. Currently, not only must researchers do this parsing and organization by hand, but they must also begin with a comprehensive list of verbs and phrases that will comprise the dictionary. Historically, the work of identifying verb phrases and classifying them has been done by undergraduate or graduate research assistants. This is time-consuming, expensive, and difficult to reproduce. The coding decisions made by research assistants are supposed to follow prescribed rules but their actual judgments are not auditable. Tools adapted from machine learning and natural language processing can be leveraged to ameliorate these challenges of event data generation. The technique presented here relies primarily on a word embedding model called *word2vec*.

3. A method for automated dictionary generation

The ADG process consists of four steps. (1) First, techniques common to NLP tasks are used to pre-process the text corpus that is to be event-coded. This is a necessary step for both event coding by PETRARCH as well as the dictionary creation process. (2) Word2vec, a neural network language model (NNLM), is then used to learn a vector-space representation of the entire vocabulary. (3) Seed words and phrases, chosen according to a pre-defined ontology, are used to extract synonymous and near-synonymous words and phrases from the word2vec model that will populate the dictionaries. (4) Finally, a set of post-processing heuristics are applied to prune and format the dictionaries. While this entire process consists of multiple steps, the researcher is responsible only for supplying an ontology in the form of a small set of seed words and phrases. The process is diagrammed in [Figure 2](#) and described in detail below. While the examples provided are drawn from the application of ADG to cybersecurity, the process is domain agnostic and can be applied widely to a variety of event domains.

3.1. Step 1: Pre-processing

Every story in the corpus that is to be event-coded is parsed and part-of-speech tagged using a shift-reduce parser, the fastest parser available from Stanford’s CoreNLP (Bauer 2014).⁷ Additionally, CoreNLP’s named entity recognizer (NER) is used to tag named entities as one of *time*, *location*, *organization*, *person*, *money*, *percent*, and *date* (Finkel *et al.* 2005).

⁷The shift-reduce parser is chosen only for its speed and so other parsers may be substituted here as necessary.

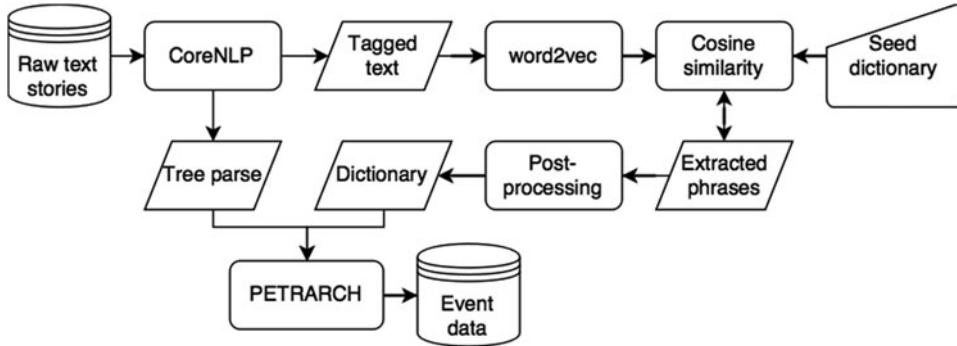


Figure 2. ADG pipeline.

Once the corpus has been parsed and named entities have been identified, two versions of the annotated text are saved. The first version is a representation of each sentence’s parse tree to be input into PETRARCH. The second version of the annotated corpus is formed by appending to each word both its entity-type tag and its part-of-speech tag. For example, the word “hackers” is transformed into “hackers:O:NNS” where “O” indicates that this word is not a named entity and “NNS” indicates a plural noun. “Snowden:PERSON:NNP” indicates that “Snowden” refers to a person and is a singular proper noun.⁸ POS and NER-tagging each word and phrase in the corpus is necessary to retain sufficient information about each term to post-process the resulting dictionary entries.

The NER and POS-tagged corpus is then processed to produce multi-word phrases. The method chosen here for deriving phrases from the corpus is recommended by Mikolov *et al.* (2013) and implemented in Rehurek and Sojka (2010). A robust literature on phrase detection exists but is out of scope for review here.⁹ Candidate bigrams (two-word phrases) are scored according to their frequency relative to the frequency of the constituent words being found independently:

$$score(w_1, w_2) = \frac{count(w_1, w_2) - \delta}{count(w_1) \times count(w_2)} \tag{1}$$

The words w_1 and w_2 are concatenated into a single multi-word term, $w_1_w_2$, if $score(w_1, w_2)$ surpasses a pre-defined threshold. δ is a discount factor that prevents spurious phrases from being formed by infrequently-occurring words. In order to produce phrases consisting of more than just two words, this algorithm is run iteratively. An example of this pre-processing is given in Figure 3.

3.2. Step 2: Vocabulary modeling

Once the text data have been tagged and phrases have been formed, a model is required to identify terms and phrases that are synonymous with the seed phrases. Word2vec is chosen for this purpose. The word2vec model is a single-hidden-layer, fully-connected, feed-forward neural network that has been shown to learn the meanings of words given their contexts in natural language texts. Word2vec produces word vectors, in the form of real-valued vectors, from raw text input in a process called embedding (Rehurek and Sojka 2010; Mikolov *et al.* 2013). These word vectors

⁸For more on the Penn Treebank POS tags, see Santorini (1990).

⁹For more, please see Dunning (1993).

These websites could contain specially crafted content that could exploit this vulnerability in Internet Explorer.

↓

These:0:DT websites:0:NNS could:0:MD contain:0:VB
specially:0:RB_crafted:0:VBN content:0:NN that:0:WDT could:0:MD
exploit:0:VB this:0:DT vulnerability:0:NN in:0:IN
Internet:MISC:NN_Explorer:MISC:NNP.

Figure 3. Example of pre-processing.

are low-dimensional numeric representations of a vocabulary that preserve the syntactic and semantic relationships between words. Word2vec learns the meaning of words from the contexts in which they are found in the text. The importance of a word's context is found in the distributional hypothesis, an assumption required by word2vec. Harris (1954), in describing the distributional hypothesis, explains that words more similar in meaning will occur among more similar contexts than will words that are dissimilar in meaning. Rubenstein and Goodenough (1965) demonstrate that “there is a positive relationship between the degree of synonymy (semantic similarity) existing between a pair of words and the degree to which their contexts are similar.”

Word2vec is actually a family of models that includes both a skipgram-based variant and a continuous bag of words (CBOW) variant.¹⁰ The skipgram model takes as input a one-hot-encoded (dummy variable) vector of length V , where V is the size of the vocabulary, in which all values are 0 except for the target word, w_i , which is coded 1. The skipgram model then attempts to predict the context words that are most likely to be found adjacent to the target word. Context words, $\{w_{i-k}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+k}\}$ are those words that fall within a window of size k on either side of the target word.¹¹ The skipgram model therefore estimates a function, $f(w_i)$, that maps target word w_i to its likely context words, $\{w_{i-k}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+k}\}$. The output of the skipgram model is a softmax-normalized vector of length V where elements represent the probabilities that each corresponding word will appear in the context window of the input word.¹² The CBOW variant is the reverse of the skipgram model and predicts a target word given its context. Both CBOW and skipgram models can be estimated with any of several software packages including the one used here, *gensim* (Rehurek and Sojka 2010).¹³

Word2vec consists of two weights matrices: an input weights matrix and an output weights matrix. By multiplying the input vector (shape $1 \times V$) with the input weights matrix (shape $V \times D$), a D -dimensional vector representation of the input word, its word vector, is formed.

¹⁰Word2vec builds on previous research into machine learning for natural language modeling, techniques for which include Latent Semantic Analysis and a variety of other NNLMs. Recently, Dhillon *et al.* (2015) use the singular value decomposition of a word-adjacency matrix to produce word embeddings. An extension of word2vec, called paragraph2vec (or doc2vec), estimates vector representations of groups of words in addition to the words themselves. These “documents” can be full sentences, paragraphs, or larger articles (Le and Mikolov 2014). Similar embedding models have explored character-level embedding and embedding based on global word co-occurrence counts (Pennington *et al.* 2014; Bojanowski *et al.* 2016).

¹¹The window is randomly sampled from between 1 and k such that words further from the target word are, on average, weighted less heavily than words immediately adjacent to the target word. Note that k is a researcher-specified hyperparameter while V is the overall size of the vocabulary; typically $V > k$.

¹²The softmax function, a multiclass generalization of the logistic function, is defined as $\sigma(x_j) = \exp(x_j) / \sum_{k=1}^K \exp(x_k)$. The softmax function maps a real-valued vector to a vector of values between zero and one that sums to one. It is therefore used to represent a probability distribution over discrete outcomes; in word2vec's skipgram case, those outcomes are context words.

¹³The entire *gensim* ecosystem of tools for natural language processing and topic modeling is available at <https://radimrehurek.com/gensim/>.

This vector representation is then multiplied by the output weights matrix (shape $D \times V$) to produce the model’s output layer.¹⁴ The softmax function (“activation”) is applied to this output layer. Because $D \ll V$, the hidden layer compresses the sparse input vectors into relatively small, dense vectors.¹⁵ These word vectors are of interest because they encode semantic and syntactic relationships between words and can be used to measure word similarities. Furthermore, algebraic operations on this vector space produce intuitive results. The canonical example of this is the analogy task, often demonstrated by showing that:

$$\vec{\text{king}} + \vec{\text{woman}} - \vec{\text{man}} \approx \vec{\text{queen}} \tag{2}$$

By adding the vector representation of “king” to the vector representation of “woman” and subtracting the vector representation of “man,” a well-trained word2vec model will produce a vector very near to the vector representation of “queen” (i.e. king:man::queen:woman).¹⁶ Why word vectors exhibit these linear relationships is the subject of active research (Pennington *et al.* 2014; Arora *et al.* 2016).¹⁷ English word embedding models are typically evaluated with a standard set of analogies like that offered by Mikolov *et al.* (2013) to test a model’s ability to represent 14 categories of semantic and syntactic relationships.

3.3. Step 3: Term and phrase extraction

Learning the corpus with word2vec allows us to easily identify synonyms or near-synonyms of our seed words and phrases. Given a seed phrase, a string search is performed on the model’s vocabulary and all words and phrases that contain the given seed word or phrase are selected. The word vectors associated with the resulting words and phrases are retrieved. These vectors are element-wise averaged to produce a single category-wide vector. The element-wise average is taken as $\| \sum_{\vec{w} \in C_i} \vec{w} \|_2$ where $\sum_{\vec{w} \in C_i} \vec{w}$ is the element-wise sum of all word vectors, \vec{w} , in category C_i . The resulting vector is l^2 normalized.¹⁸ Then, the top n_i most similar terms and phrases to each mean category vector are extracted from the word2vec model. Similar words and phrases are identified by first computing the cosine similarities of all word vectors with the category mean vector. Cosine similarity, defined as $(\vec{X} \cdot \vec{Y}) / (\|\vec{X}\| \times \|\vec{Y}\|)$, is a measure of the angle between two vectors and is particularly useful for comparing high-dimensional vectors. Cosine similarity is used to rank-ordered all terms and phrases in the word2vec model’s vocabulary by their similarity to the mean category vector in descending order. The top s

¹⁴Mikolov *et al.* (2013) introduce NEG, a negative sampling objective function, for optimizing word2vec. For a discussion of the skipgram negative sampling word2vec objective function, see Goldberg and Levy (2014). Word2vec can be optimized via stochastic gradient descent as described by Mikolov *et al.* (2013).

¹⁵ D is a parameter supplied by the researcher. Common values are 100 and 300.

¹⁶Note that the overhead arrow notation is here used to indicate a vector. For example: $\vec{\text{king}} \in R^n$.

¹⁷Word embedding models are imperfect approximations of language; failure cases may include instances where antonyms share very similar word vector representations because they occur in similar contexts (Nguyen *et al.* 2016). Additionally, infrequent words and phrases tend not to be represented as well by word2vec as frequently-occurring words and phrases. In fact, implementations of word2vec accept a minimum count parameter to filter out infrequent words. Readers interested in an alternative approach to keyword discovery that does not rely on word embedding should consult (King *et al.* 2017). These issues will impact the performance of ADG. For example, infrequently-referenced actors may fall below the minimum count threshold chosen for word2vec and therefore not appear in the final dictionaries. Those infrequently-referenced actors that make the cut-off might still not occur frequently enough to produce reliable word vectors. Additionally, research has shown that semantic relationships learned by word embedding models can mimic human biases. Caliskan *et al.* (2017) demonstrate that GloVe, when trained on standard texts, inherits biases measured in humans via the Implicit Association Test (Greenwald *et al.* 1998). GloVe is a word embedding model based on factorizing a global word co-occurrence matrix (Pennington *et al.* 2014).

¹⁸The l^2 norm for vector $\langle x_1, x_2, \dots, x_n \rangle$ is given by $\sqrt{\sum_{i=1}^n x_i^2}$.

most similar terms and phrases are chosen as candidates to populate the relevant category in the event-coding dictionary.¹⁹

3.4. Step 4: Post-processing

Extracted terms are then post-processed according to a set of rules associated with the dictionary they are meant to comprise. These post-processing steps can be automated. The set of post-processing rules can be found in the online appendix. The post-processing is necessary to coerce the extracted terms and phrases into the dictionary formats expected by PETRARCH. This involves, among other things, grouping verb phrases by their common verbs and tagging each dictionary entry with a category tag. A post-processing filter that removes phrases from the verb dictionary if they do not include at least one verb is also applied.

ADG represents a major step towards fully-automated event-data coding for novel domains. Because this process can be done largely without human interaction and the content of the dictionaries are a function of the raw data that are to be event-coded, the dictionaries can be updated in tandem with the event dataset itself; new verb phrases, actors, or agents can be learned by the underlying models as they enter the relevant domain's vocabulary. Additionally, because the process described herein relies on only a small amount of initial researcher input data and the raw text data itself, the process of event data generation is made more fully reproducible from start to finish.

4. CYLICON: a cyber event dataset

This method of ADG for event coding is now applied to a novel domain for event data: cybersecurity. First, a cybersecurity ontology is selected and seed phrases are chosen to represent each category of that ontology. Five dictionaries are generated: verbs, actors, agents, synsets, and issues.²⁰ Only one seed phrase is provided per category.²¹ Seed phrases are shown in Tables 2 and 3 and in the online appendix. For each seed term or phrase, the average vector of all terms and phrases containing the seed is computed and similar terms and phrases are identified according to the described ADG procedure. The extracted candidate terms and phrases are then post-processed and formatted into PETRARCH-styled dictionaries; no manual changes have been made to the dictionaries at any point after the input of the 26 seed phrases (one per category).

Ten categories of events are identified for the verb dictionary: defacements, DDOS events, infiltrations, leaks, infections, vulnerability discoveries, arrests, patches, phishing attacks, and censorship incidents.²² The ten seed phrases are representative examples of verb phrases for each category.²³ These are chosen by the researcher. The extracted verb dictionary contains 640 verbs and phrases after de-duplication and post-processing. The number of extracted phrases

¹⁹ s is a researcher-selected value that puts an upper limit on the number of terms and phrases that will constitute the dictionary. However, automated post-processing steps described in the appendix may result in the inclusion of fewer terms and phrases. Values for s used here are 300, 300, 50, and 25 for the verb, actor/agent, synset, and issue dictionaries, respectively.

²⁰The word2vec model was trained on a convenience sample of cybersecurity news data from a number of sources. The model is trained according to the *gensim* default parameters except `min_count 10`, `window 10`, and `vector size 300`.

²¹More than one seed word or phrase can be supplied per category. Only one seed phrase per category is used here to demonstrate the use of ADG with the minimum amount of researcher input.

²²These categories were chosen for exploratory purposes. For future iterations of CYLICON, existing cybersecurity ontologies, to include those developed by Herzog *et al.* (2007) and Swimmer (2008), will be considered as alternatives. These will require technical reports of cybersecurity events as opposed to the newswire-like corpus used here. They may also require adjustment to fit social science applications rather than their intended audience of cybersecurity experts and incident responders.

²³In fact, some seed terms are nouns rather than verb phrases. For example, the seed word for the vulnerability discovery category is simply `VULNERABILITY:O:NN`.

Table 2. Verb dictionary seeds

Category	Seed Phrase
DEFACED	DEFACED:O:VBD
PATCHED	PATCHED:O:VBD
INFILTRATED	BREACHED:O:VBD
LEAKED	LEAKED:O:VBD
PHISHED	PHISHED:O:VBD
DDOS	DISTRIBUTED:O:VBN_DENIAL-OF-S ...
INFECTED	INFECTED:O:VBD
VULNERABILITY	VULNERABILITY:O:NN
ARRESTED	ARRESTED:O:VBD
CENSORED	CENSORED:O:VBD

Table 3. Actor & agent dictionary seeds

Category	Seed Phrase
HACKER	HACKER:O:NN
RESEARCHER	RESEARCHER:O:NN
WHISTLEBLOWER	WHISTLEBLOWER:O:NN
USERS	USERS:O:NNS
ANTIVIRUS	ANTIVIRUS:O:NN

is due, in large part, to the minimum similarity threshold that is set by the researcher; terms and phrases must surpass this threshold with respect to the average category vector in order to be included in the final dictionaries. Here, a minimum cosine similarity of 0.6 has been chosen.

The new categories of actors and agents introduced in CYLICON include hackers, researchers, users, whistleblowers, and antivirus companies/organizations. These categories are appended to the existing actor and agent classifications already found in the default Phoenix dictionaries. New issue categories are appended to the issues already supplied with PETRARCH and include TOR, 0Day, hacktivism, DDOS, social engineering, and state-sponsorship. Synsets are produced for categories including hardware, virus, web asset, software, and computer.²⁴

The selected text corpus represents a convenience sample of 77,410 documents collected from online sources including cybersecurity-related blogs and news sites. Roughly 22,000 articles are sourced from the news section of www.softpedia.com. The remaining stories are largely sourced from blogs and technology-oriented news sites, the largest of which include feed aggregators, theregister.com, csoonline.com, circleid.com, and darkreading.com. There are 1,231 unique sources represented in the corpus. These sources are not a representative sample of cybersecurity events and were instead selected due to their relatively high concentration of relevant cybersecurity event stories. Collection occurred during 2014 and the latter part of 2015 and was inconsistent over time due to heterogeneity among sources with respect to the availability of archival text.

CYLICON includes 671 events in total. Arrests make up the largest category with 211 events, followed by infiltration (200), leaks (97), defacements (97), patches (19), infections (19), DDOS attacks (17), vulnerability discoveries (5), phishing attacks (5), and censorship incidents (1). Infiltration is a common category as many verb phrases from cybersecurity reporting accurately map to it. For example, phrases that include the words “breached” and “hacked” are often classified as infiltration by the ADG process. Additionally, when websites are defaced, it is common for reports to describe the websites as having been “breached and defaced,” indicating that the incident could be accurately assigned to either or both categories. Often, popular reporting on cybersecurity is not precise enough to distinguish the characteristic of a particular “hacking”

²⁴These categories are not mutually exclusive.

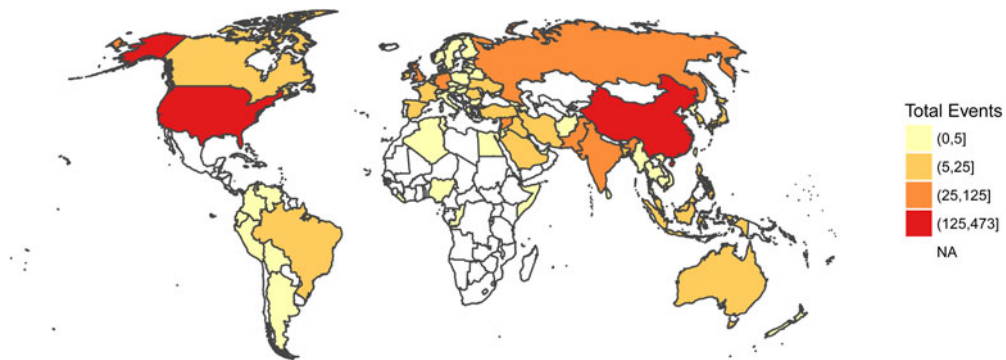


Figure 4. Spatial distribution of actors in CYLICON. White (NA) values indicate that no events in CYLICON identify an actor from a given country.

event in a single sentence. Because of this, a bias towards infiltration coding is induced. If the coded sentence explains that a target was “hacked” and a second sentence explains that the event resulted in the defacement of the target’s website, PETRARCH will fail to connect the defacement to the hacking event and will therefore code the event as an infiltration rather than a defacement. The discovery of vulnerabilities, issuance of patches, and phishing attempts, while very common, often go unreported in the news sources utilized here. They also tend not to conform to the source-action-target triple expected by PETRARCH. Of the 640 verb phrases in the CYLICON dictionaries, 157 of them account for all of the coded events. This is a 15-fold increase over the size of the verb seed dictionary.

The geographic distribution of actors involved in cyberspace according to CYLICON is shown in Figure 4. This map corresponds to conventional wisdom about the most active actors in cybersecurity-related events (The Economist 2012; Akamai 2015; Clapper 2015). However, this map is not representative of the entire CYLICON dataset; not all relevant actors are geo-coded. Of 1,338 total coded actors, 1,245 are assigned to specific countries. PETRARCH attempts to assign country codes to actors and agents when they can be inferred from the text; for example, the phrase “Syrian hackers” may be coded as SYRHAC. Actors affiliated with international organizations or otherwise unaffiliated with specific countries are, of course, not included in the map. Country associations for cybersecurity-based actors and agents have not been inferred for CYLICON.²⁵ The US is the most prominent country in CYLICON with 473 events followed by China (145), Great Britain (60), India (43), Pakistan (38), and Russia (38). 82 unique countries are represented in total.

Because event data from PETRARCH are dyadic, we can also examine country pair interactions. Figure 5 represents the most common dyadic pairs in CYLICON. Chord plots, common in network analysis applications, represent the volume of interaction between nodes or, in this case, countries. This particular chord plot is non-directed and does not include self-connections. The top 12 countries (by volume of events) are plotted and the remaining 70 are grouped into the category “other” for visual clarity. The larger edges conform to the expectations of Valeriano and Maness (2014); regional pairs and rivals are apparent in the graph. The US is most active with China and Russia. India and Pakistan account for the majority of one another’s cyber events. Iran interacts primarily with the US and Israel.

To better illustrate the successes and shortcomings of CYLICON, a selection of events are examined alongside their original text. Event codes are indicated by the triplet ACTOR1 ACTOR2 ACTION preceding each sentence. Selected sentences and their corresponding data

²⁵ Actors identified by the ADG process are assigned the country code XXX by default. The online appendix to this paper evaluates an ADG extension for actor geocoding.

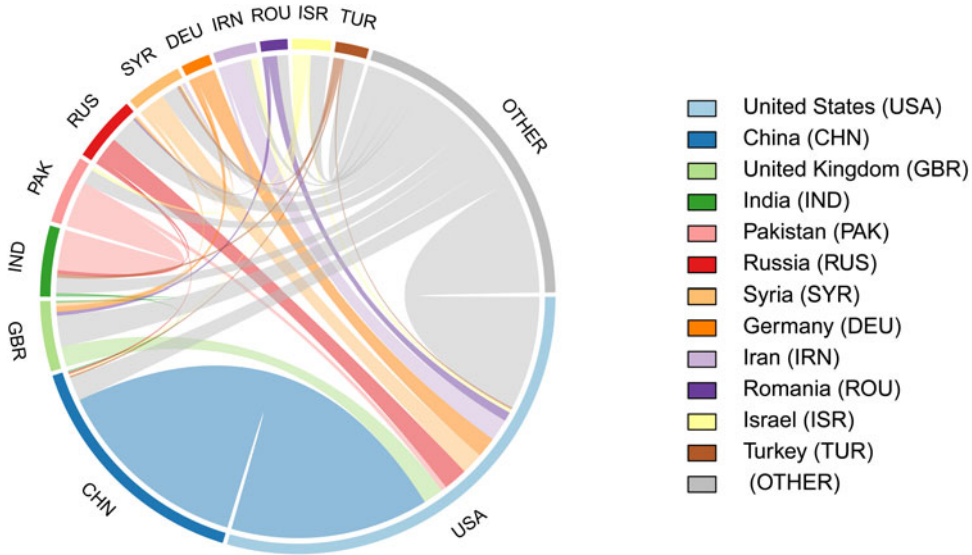


Figure 5. Top country dyads in CYLICON.

are enumerated in the list below, beginning with examples of accurate coding and ending with examples of inaccurate coding. Commentary follows.

1. ISR USAELIGOV INFILTRATED: “According to FBI, in the Year 2000 Israeli Mossad had penetrated secret communications throughout the Clinton administration, even Presidential phone lines.”²⁶
2. USACOP EST ARRESTED: “After the Estonian masterminds were apprehended by the FBI, the DNSChanger Working Group was established and the cleaning process began.” (Kovacs 2012b)
3. MYS PHLGOVMEDHAC DDOS: “After Anonymous Malaysia launched distributed denial-of-service (DDOS) attacks against several Philippines government websites, Filipino hackers went on the offensive, defacing a large number of commercial websites.” (Kovacs 2013)
4. USA USAMIL INFECTED: “US officials did not provide details on the status of the ‘corrupt’ software installed on DoD computers, but common sense points us to believe it was removed back in 2013.” (Cimpanu 2015)
5. BGD MED BGD DEFACED: “A Bangladeshi publisher of secular books has been hacked to death in the capital Dhaka in the second attack of its kind on Saturday, police say.” (BBC 2015)
6. IRNGOVGOVMIL USA INFILTRATED: “Head of Iran’s Civil Defense Organization Gholam Reza Jalali told the agency that the country never hacked financial institutions from the United States.” (Kovacs 2012a)

The first four examples are all accurately coded by PETRARCH. Item 1 is correctly identified as an instance of infiltration and the actors are accurate if imprecise (PETRARCH codes Mossad as ISR rather than ISRSPLY). In Item 2, the ADG process identified “were apprehended” as indicative of arrest. While Item 3 is correctly labeled a DDOS event, PETRARCH has mistakenly

²⁶Source unavailable due to dead hyperlink.

Table 4. Accuracy by event category

	INCORRECT	AMBIGUOUS	CORRECT
ARRESTED	45	2	164
CENSORED	0	0	1
DDOSED	4	4	9
DEFACED	32	14	51
INFECTED	8	1	10
INFILTRATED	44	11	145
LEAKED	45	1	51
PATCHED	14	0	5
PHISHED	5	0	0
VULNERABILITY	4	0	1
TOTAL	201	33	437

associated the term “hackers” with the target actor rather than with Anonymous Malaysia. Item 4 highlights the difficulty associated with coding infection events. The “corrupt software installed” indicates that a malware infection event has occurred. However, as is often the case with infection events, a source actor is not described. In this case the target actor is accurately identified but the source actor is coded as the US, which is not supported by the given text. Note that none of the verb phrases in Items 1, 2, or 4 were included in the seed terms.

Items 5 and 6 were incorrectly coded. The incorrect coding in item 5 resulted from the dual meaning of the verb “hacked.” It is possible that with a larger ontology, one that includes both computer infiltration and murder, “hacked_to_death” would be accurately coded. However, without a method for automatically pruning erroneously-coded phrases from the dictionaries, edge cases like this must be identified and removed by hand. No manual pruning has been performed on these dictionaries and so edge cases remain. Item 6 is incorrectly coded because the sentence itself is a denial of the action that was identified. An Iranian official denies that his country had hacked into financial institutions in the US but PETRARCH interpreted the sentence to mean that the event had, in fact, occurred.²⁷

All CYLICON events have been reviewed manually and scored to help quantify the efficacy of automatically-generated event data dictionaries. The text content associated with each event is inspected and event codes are manually assigned without any knowledge of the CYLICON-assigned codes. In the case that multiple events are explicitly described (e.g. “... have breached and defaced...”), all appropriate events are assigned. When only one event is described (e.g. “...have defaced...”), only that specific event is assigned. When the language is ambiguous, all reasonable assignments are made but the event is also labeled as “ambiguous.” Only the action or event type field is evaluated as only the verb dictionary was produced completely via ADG. The CYLICON actor, agent, and issue dictionaries are a combination of the Phoenix hand-coded dictionaries and automatically-generated dictionaries and are therefore not evaluated. Events are scored as correct if the associated action code from CYLICON is among the manually-identified event types for a given sentence. Events are scored as ambiguous if the associated action code from CYLICON is among the manually-identified event types but the text itself is ambiguous rather than explicit. For example, “Hackers have attacked servers...” is ambiguous because it could reasonably describe a DDOS event, an infiltration event, or a defacement. Events are considered incorrect if they fall into neither of the above two cases.

Table 4 presents the results of this review by event category. Overall accuracy, the number of correctly-coded events and ambiguous events divided by the total number of coded events, is 70 percent. If ambiguous events are instead considered inaccurate, the accuracy of coded events falls to 65 percent. These values are in line with or above the reported human coder performance on

²⁷The prevailing wisdom is that Iran was complicit in the attacks in question Volz and Finkle (2016). However, PETRARCH failed to code the event accurately given the supplied context.

top-level event categories. King and Lowe (2003) report that trained undergraduates can correctly classify events by their aggregate (top-level) event category between 39 and 62 percent of the time.²⁸ Schrodtt and Brackle (2013) report machine-coding accuracy percentages for TABARI on the ICEWS project in the low- to mid-70s. The false positive rate, the percentage of sentences incorrectly determined by PETRARCH to contain any event, is 16 percent.²⁹ This performance is achieved despite requiring only minimal researcher-hours and one seed phrase per category.

5. Conclusion


The ADG process described here allows researchers to quickly produce novel event datasets specific to their topics of interest. With minimal input from the researcher, ADG produces dictionaries of pre-categorized words and phrases for use with existing event coding software. In a demonstration of its application, ADG was used to populate and update a set of dictionaries for coding events in an entirely new domain for event data—that of cybersecurity.

While ADG takes a substantial step in the direction of a fully-automated event coding solution, work remains to be done in this area. Event coding software itself, like PETRARCH, remains largely heuristic-based. The stacking of multiple analysis techniques for sentence parsing, phrase-extraction, and named entity recognition, among others, compounds errors that lead to sub-optimal event coding. Future efforts should leverage advances in machine learning to minimize the application of heuristics and the stacking of text pre- and post-processing steps.³⁰

End-to-end event coding models may, for instance, facilitate the customization of event datasets through transfer learning.³¹ For example, a model may be trained to produce CAMEO-coded event data from news and then adapted, with the help of a relatively small training set, to produce cybersecurity event data instead. This would allow novel event datasets to be generated for user-specific purposes with only a small number of “gold standard” training samples. An extension to the ADG process presented here would replace the word2vec component with a bilingual embedding model like BilBOWA (Gouws *et al.* 2015). BilBOWA requires only a parallel bilingual corpus in order to align separate word embedding models in two different languages and could therefore be used in the ADG process to extract bilingual dictionaries.

ADG demonstrates that even unstructured text can be converted into structured data suitable for social science inquiry with minimal researcher input. As machine learning and neural network-based models continue to advance the state-of-the-art in data analysis across fields, their application to the social sciences promises to similarly revolutionize how we measure, interpret, and understand political phenomena.

Supplementary Material. The supplementary material for this article can be found at <https://doi.org/10.1017/psrm.2019.1>

Author ORCIDs.  Benjamin J. Radford, [0000-0002-8440-0655](https://orcid.org/0000-0002-8440-0655).

Acknowledgments. Benjamin J. Radford received his Ph.D. in political science from Duke University (benjamin.radford@gmail.com). The author thanks Michael D. Ward, Scott De Marchi, Kyle Beardsley, Mark J.C. Crescenzi, and three exceptional anonymous reviewers for their support and feedback on drafts of this work.

References

- Akamai (2015) State of the internet—security. 2, no. 4 (Q4).
 Althaus SL, Bajjalieh J, Carter JF, Peyton B and Shalmon DA (2017) Cline center historical event data. June 26. <https://www.clinecenter.illinois.edu/data/event/phoenix>.

²⁸The maximum accuracy value increases to 72 percent when event types are weighted by frequency.

²⁹King and Lowe (2003) note that their event coding software, *Reader*, suffers from a higher false positive rate than the human coders.

³⁰See IARPA (2018) for one such effort.

³¹For more on transfer learning, see Pan and Yang (2010).

- Arora S, Li Y, Liang Y, Ma T and Risteski A (2016) Rand-walk: a latent variable model approach to word embeddings. *arXiv:1502.03520v7* (July 22).
- Azar EE (1980) The conflict and peace data bank (copdab) project. *The Journal of Conflict Resolution* **24**, (April).
- Bauer J (2014) Shift-reduce constituency parser [in English]. The Stanford Natural Language Processing Group. Online. <http://nlp.stanford.edu/software/srparser.shtml>.
- BBC (2015) Bangladeshi secular publisher hacked to death. Online. <http://www.bbc.co.uk/news/world-asia-34688245> October 31.
- Bojanowski P, Grave E, Joulin A and Mikolov T (2016) Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Boschee E, Lautenschlager J, O'Brien S, Shellman S, Starz J and Ward M (2015) Icews coded event data. **V15**. <http://dx.doi.org/10.7910/DVN/28075>.
- Brecher M and Wilkenfeld J (2000) *A Study of Crisis*. Ann Arbor, Michigan. University of Michigan Press.
- Brecher M, Wilkenfeld J, Beardsley K, James P and Quinn D (2016) International crisis behavior data codebook, version 11. <http://sites.duke.edu/icbdata/data-collections>.
- Caerus Associates (2015) Phoenix event data set codebook 0.0.1b. https://s3.amazonaws.com/oeda/docs/phoenix_codebook.pdf.
- Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186. ISSN: 0036-8075. doi: 10.1126/science.aal4230. eprint: <http://science.sciencemag.org/content/356/6334/183.full.pdf> <http://science.sciencemag.org/content/356/6334/183>.
- Cimpanu C (2015) Military contractors that used Russian programmers for dod software get fined by US govt. <http://news.softpedia.com/news/military-contractors-that-used-russian-programmers-for-dod-software-get-fined-by-us-govt-495827.shtml>. *Softpedia Security News*. (November 6).
- Clapper JR (2015) *Worldwide threat assessment of the us intelligence community*. http://cdn.arstechnica.net/wp-content/uploads/2015/02/Clapper_02-26-15.pdf. Senate Armed Services Committee, February 26, 2015.
- Dhillon PS, Foster DP and Ungar LH (2015) Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research* **16**, 3035–3078. <http://www.pdhillon.com/dhillon15a.pdf>.
- Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19**, 61–74.
- Finkel JR, Grenager T and Manning C (2005) Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics*, pp. 363–370. <http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf>.
- Goldberg Y and Levy O (2014) word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv* (February).
- Gouws S, Bengio Y and Corrado G (2015) Bilbowa: fast bilingual distributed representations without word alignments. *Proceedings of the 32nd International Conference on Machine Learning* **37**, 748–756.
- Greenwald AG, McGhee DE and Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* **74**, 1464–80.
- Harris ZS (1954) Distributional structure. *WORD* **10**, 146–162.
- Herzog A, Shahmehri N and Duma C (2007) An ontology of information security. *International Journal of Information Security and Privacy* **1**, 1–23.
- IARPA (2018) Draft Broad Agency Announcement: Better Extraction from Text Towards Enhanced Retrieval (BETTER). IARPA-BAA-18-05.
- King G and Lowe W (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**, 617–642. <http://gking.harvard.edu/files/gking/files/infoex.pdf?m=1360039060>.
- King G, Lam P and Roberts ME (2017) Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science* **61**, 971–988.
- Kovacs E (2012a) Iranian official: we did not launch cyberattacks on American banks. *Softpedia Security News* (September 24). <http://news.softpedia.com/news/Iranian-Officials-We-Did-Not-Launch-Cyberattacks-on-American-Banks-294412.shtml>.
- Kovacs E (2012b) Tick tock: It's lights out for DNS changer-infected computers on July 9. *Softpedia Security News* (June 6). <http://news.softpedia.com/news/Tick-Tock-It-s-Lights-Out-for-DNSChanger-Infected-Computers-on-July-9-Video-279700.shtml>.
- Kovacs E (2013) Hundreds of sites hacked in conflict between Malaysia and Philippines hacktivists. *Softpedia Security News* (March 4). <http://news.softpedia.com/news/Hundreds-of-Sites-Hacked-in-Conflict-Between-Malaysia-and-Philippines-Hacktivists-334047.shtml>.
- Le Q and Mikolov T (2014) Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*.
- Leetaru K and Schrodt PA (2013) GDELT: Global Data on Events, Location and Tone, 1970–2012. Annual Meeting of the International Studies Association. <http://data.gdelproject.org/documentation/ISA.2013.GDELT.pdf>.

- Mikolov T, Chen K, Corrado G and Dean J (2013) Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*. <http://arxiv.org/pdf/1301.3781.pdf>.
- Mikolov T, Sutskever I, Chen K, Corrado G and Dean J (2013) Distributed representations of words and phrases and their compositionality. *arXiv* (October). <http://arxiv.org/abs/1310.4546>.
- Nguyen KA, im Walde SS and Vu NT (2016) Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 454–459.
- Norris C, Schrodt P and Beiler J (2017) PETRARCH2: Another event coding program. *The Journal of Open Source Software* 2, (January), 1–1. doi: 10.21105/joss.00133, <http://dx.doi.org/10.21105/joss.00133>.
- Open Event Data Alliance (2015a) *PETRARCH Python Engine for Text Resolution and Related Coding Hierarchy*. Online. <http://www.github.com/openeventdata/petrarch>.
- Open Event Data Alliance (2015b) Phoenix Data Project. Online. phoenixdata.org.
- Open Event Data Alliance (2015c) Phoenix Pipeline. Online. <http://phoenix-pipeline.readthedocs.org/en/latest>.
- Open Event Data Alliance (2018) Universal Dependency PETRARCH. <https://github.com/openeventdata/UniversalPetrarch>.
- Palmer G, D'Orazio V, Kenwich M and Lane M (2015) The MID4 dataset, 2002–2010: Procedures, coding rules and description. *Conflict Management and Peace Science* 32, 222–242.
- Pan SJ and Yang Q (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 1345–1359.
- Pennington J, Socher R and Manning CD (2014) GloVe: Global Vectors for Word Representation, In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
- Raleigh C, Linke A, Hegre H and Karlsen J (2010) Introducing ACLED – Armed Conflict Location and Event Data. *Journal of Peace Research* 47, 651–660.
- Rehurek R and Sojka P (2010) Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, May, pp. 45–50 <http://is.muni.cz/publication/884893/en>.
- Rubenstein H and Goodenough JB (1965) Contextual Correlates of Synonymy. In *Communications of the ACM* vol. 8(10), 627–633. ACM, New York, NY.
- Santorini B (1990) Part-of-speech tagging guidelines for the penn treebank project. <https://www.cis.upenn.edu/~treebank/>.
- Schrodt PA (1998) KEDS Kansas Event Data System version 1.0. <http://eventdata.parusanalytics.com/>.
- Schrodt PA (2011) TABARI: Textual Analysis by Augmented Replacement Instructions, Version 0.7.6. <http://eventdata.parusanalytics.com/tabari.dir/tabari.manual.0.7.6.pdf>.
- Schrodt PA and Brackle DV (2013) Automated Coding of Political Event Data. In Subrahmaniam VS (ed.) *Handbook of Computational Approaches to Counterterrorism*. Springer Science + Business Media. New York, NY.
- Schrodt PA, Gerner DJ and Yilmaz O (2009) Conflict and Mediation Event Observations (CAMEO): an event data framework for a post Cold War world. *International Conflict Mediation: New Approaches and Findings*.
- Swimmer M (2008) Towards an ontology of malware classes. <https://www.scribd.com/document/24058261/Towards-an-Ontology-of-Malware-Classes>.
- The Economist (2012) Hype and fear: America is leading the way in developing doctrines for cyber-warfare. other countries may follow, but the value of offensive cyber capabilities is overrated. *The Economist* (December 8).
- The GDELT Project (2016) The Datasets of GDELT as of February 2016. <https://blog.gdeltproject.org/the-datasets-of-gdelt-as-of-february-2016>. March 13.
- Valeriano B and Maness RC (2014) The dynamics of cyber conflict between rival antagonists, 2001–11. *Journal of Peace Research* 51(3), 347–360.
- Volz D and Finkle J (2016) US indicts iranians for hacking dozens of banks, New York dam. *Reuters*. <http://www.reuters.com/article/us-usa-iran-cyber-idUSKCN0WQ1JF>.
- Wang W, Kennedy R, Lazer D and Ramakrishnan N (2016) Growing pains for global monitoring of societal events. *Science Magazine Digital*, 1502–1503. (September 30).
- Ward MD, Beger A, Cutler J, Dickenson M, Dorff C and Radford B (2013) Comparing GDELT and ICEWS event data. http://mdwardlab.com/sites/default/files/GDELTICEWS_0.pdf.