The Dimensionality of Political News Reports *

Benjamin E. Bagozzi Philip A. Schrodt

beb196@psu.edu schrodt@psu.edu

Version 1.0 : June 12, 2012

*Paper presented at the European Political Science Association meetings, Berlin, June 2012. The authors wish to acknowledge the valuable suggestions and feedback that they received from Burt Monroe, Doug Rice and James Honaker, as well as high performance computing support from the Penn State Research Computing and Cyberinfrastructure Group. This project was funded in part by National Science Foundation grant SES-1004414, and by a Fulbright-Hays Research Fellowship to Schrodt for work at the Peace Research Institute, Oslo (`http://www.prio.no`). Address for both authors: Department of Political Science, Pennsylvania State University, University Park, PA 16802 USA

Abstract

Latent Dirichlet Allocation (LDA) models are a machine learning method for finding sets of words that characterize latent dimensions in texts. In this paper, we apply LDA to a very large corpus of international newswire texts for 2000 to 2011 covering 61 countries in Europe and the Middle East to determine the extent to which these latent dimensions correspond to the categories found in existing event data ontologies such WEIS, CAMEO and IDEA. We analyze the texts after removing common stop words, proper nouns and meronyms based on the CountryCodes file developed by the CAMEO project. The LDA analysis produces very plausible latent topics—which is not always true of LDA—and most of these topics are general rather than country-specific, though as expected, some country-specific crisis and topic areas (e.g. specific international and ethnic conflicts) can be found as well. Clustering analysis supports a dominant cooperation-conflict dimension in most but not all of the cases, and shows very strongly that when compared across countries, the latent topics cluster substantively rather than by country, and those clusters contain representatives of the topics derived from the set of all of the stories. However, LDA also reveals a number of additional issues, particularly economic and institutional, that are generally not covered well in the existing event data coding schemes, in large part because of their traditional emphasis on violent conflict. This suggests that automated coding methods could be used to substantially expand the scope of event data, particularly in areas relevant to political economy and routine institutional behavior such as elections. It may be possible to use these methods to increase the effectiveness of report filters and to identify stories containing sentences that should produce events but are missed by automated coding programs due to incomplete dictionaries.

1 Introduction

Political event data have had a long presence in the quantitative study of international politics, dating back to the early efforts of Edward Azar's COPDAB [Azar, 1980] and Charles McClelland's WEIS [McClelland, 1976] as well as a variety of more specialized efforts such as Leng's BCOW [Leng, 1987]. By the late 1980s, the NSF-funded *Data Development in International Relations* project [Merritt et al., 1993] had identified event data as the second most common form of data—behind the various Correlates of War data sets—used in quantitative studies [McGowan et al., 1988]. The 1990s saw the development of two practical automated event data coding systems, the NSF-funded KEDS [Gerner et al., 1994, Schrodt and Gerner, 1994] and the proprietary VRA-Reader (`http://vranet.com`; King and Lowe 2004) and in the 2000s, the development of two new political event coding taxonomies—CAMEO [Gerner et al., 2009] and IDEA [Bond et al., 2003]—designed for implementation in automated coding systems. By the 2000s, with the decline of inter-state war, most event data studies shifted to the study of internal conflict, with the major project during this period being the $37-million U.S. Defense Advanced Research Projects Agency (DARPA) Integrated Conflict Early Warning System (ICEWS; O'Brien 2010). While the original objective of ICEWS was conflict forecasting in Asia, the dependence of its most successful forecasting models on event data caused the program to morph into the production of a global event data set for 1996 to 2012, coded with the CAMEO event scheme and a customized sub-state actor scheme, which should be released sometime in the second half of 2012 and may be updated on a regular basis after that point.

All event data sets use news stories as the basis of the data, with the automated systems generally coding individual sentences except for the coreferencing of pronouns across sentences. In contemporary automated systems, the focus is generally the major international news wires—in particular Reuters, *Agence France Presse* and BBC—sometimes supplemented with elite English-language regional sources. Usually only the first four to six sentences are coded, since these usually are sufficient to summarize the major interactions involved in the story, and sentences further down are more likely to include historical details that could be confused for contemporary events.

The IDEA and CAMEO event coding systems are both based on the original WEIS system. WEIS, in turn, was developed in a somewhat ad hoc manner, and McClelland [1983] viewed it only as a "first draft" of a system, presumably not something intended for widespread use, at least in derivative form, half a century later. Furthermore, WEIS was developed in a radically different political and media environment than that found in the twenty-first-

century post-Cold War world of Factiva, Lexis-Nexis, and machine-readable international news wires.

The objective of this paper is therefore to determine the extent to which the content of politically salient news articles from international news wire services corresponds to the coding categories found in these major event data coding schemes. We use standard contemporary natural language processing (NLP) methods and software, and latent Dirichlet allocation (LDA) models to determine the most common latent topics found in reports covering a wide range of European and Middle Eastern countries. We first informally look at the correspondence between those categories and the event data systems, and then use hierarchical clustering to determine both the distinctiveness of those latent topic clusters and the extent to which these are general rather than nation-specific. Finally, we will consider some of the implications of these results for future enhancements of automated event data coding.

2 Method

Latent Dirichlet allocation models were introduced by Blei et al. [2003] and briefly described in the abstract of that article as:

> LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

In the typical LDA application to document classification, each document is assumed to be a mixture of multiple, overlapping *latent topics*, each with a characteristic set of words. Classification is done by associating words in a document with the topics most likely to have generated the observed distribution of words in the document. The purpose of LDA is to determine those latent topics from patterns in the data.

The latent topics are useful for two purposes. First, to the extent that the words associated with a topic suggest a plausible category, they are intrinsically interesting in determining the most common sets of issues found in the set of documents. For example, one of the sample data sets in the *R* `lda` package [Chang, 2010] determines the set of issues discussed in a series of political blogs. Second, the topics can be used with other classification algorithms such as logistic regression, support vector machines or discriminant analysis to classify new documents. The full mathematical details of LDA estimation can be obtained from Blei

et al. [2003] or the other usual suspects on the web and will not be repeated here, as we are simply using these methods off-the-shelf (or off-the-`CRAN`, as the case may be.)

The existence of at least some topic clustering in news reports can be taken for granted: all major newspapers, for example, organize themselves by major topics: for instance the major headings in the Web page of the *New York Times* begin with `World, U.S. Politics, New York, Business Dealbook, Technology, Sports, Science, Health, Arts, Style, Opinion` followed by an additional 24 less-highlighted categories. Consistent with the LDA approach, these categories are over-lapping, and for example a report on a new breakthrough in laser manufacturing might appears in the `Business, Technology` and `Science` sections. Aside from the practice of placing the same article in multiple categories, this organizational scheme pre-dates the web, and can also be found in the indexing systems of the major newswires.[1]

Most of the stories relevant to event data analysis, however, are found in some variation of the `World/International Politics` category. Hence, other than their usefulness in eliminating irrelevant content such as sports stories, these categories provide little-to-no differentiation of political interactions. Consequently, the questions we are addressing are (1) whether we can find—using an objective methodology that does not depend on the human understanding of the story—latent dimensions within those stories, and (2) the extent to which these correspond to the categories assumed to exist by the event data coding schemes.

In a sense, with LDA we are doing the reverse of traditional event data coding. The latter approach takes a story and categorizes it into one of a finite set of categories specified *a priori*. We are taking the stories and determining the categories.

The importance of latent dimensions in event data is due to issues of measurement. News reports are only a tiny, tiny fraction of all of the events that occur daily, and are non-randomly selected by reporters and editors. The event coding schemes are very generic and bin together events that may not always belong together in all contexts. Because of these issues, the ability to determine latent dimensions in the news stories themselves is important in the overall scientific exercise of improving instrumentation for conflict forecasting, and these may not be self-evident (or purely derivable from theory) because of measurement factors. We do not have a "god's-eye view" of political interactions—we have the highly (and non-randomly) selected view provided by the international media.

[1]Based on some conversations over the years with various data providers, this indexing is still largely done by humans, though surely these firms must be at least considering automated indexing. In addition, many of the categories are pre-determined: except in the very rare instances where a politically relevant event such as a terrorist attack occurs, an individual covering a sports event knows the stories will be classified as `Sports`, a reporter covering a fashion show knows it will be in `Style` and so forth.

3 Data

3.1 Sample

This sample used for this study was drawn from a collection of politically-relevant English-language international news texts; these were largely from the major international news wire sources. Using filtering from the data providers and the news wires internal indexing of the content, stories without political content—sports, style, routine business reports—were supposed to have been eliminated and this appears to have been fairly effective though not perfect.[2] The texts record only the first six sentences of every news story, and are stored in TABARI-readable format [Schrodt, 2011] with unique story-identifiers. The reports covered 73 autonomous or semi-autonomous nation-states in Europe and the Middle East during the time period January 1st 2001 to July 2011. A subset of these had very limited coverage over our time period of interest, and were too sparse for LDA analysis. We therefore removed the 12 nation-states with populations under 1-million from our sample before our primary LDA analysis.[3] The 61 remaining nation-states in our sample are listed in Table 7 in the Appendix.

For more detailed processing in this paper, we selected four countries in Europe: France, Greece, Norway and Poland, and four in the Middle East: Egypt, Israel, Jordan and Turkey. In addition, we initially experimented with three relatively small cases—Albania, Armenia, and Portugal—and have also included these below. In addition we created a combined corpus of stories from all 61 nation-states using a random sample of 5,000 news-stories from each nation-state-corpus, and then pooled these news story samples into one single, "combined-nation" corpus, which we refer to as the "ALL" case.

3.2 Preprocessing

For each nation-state, we processed the texts to remove proper nouns, meronyms[4] or stop-words.[5] First, the texts were individually read into R using `readLines()` and converted

[2]Which, as we have bemoaned in other venues, e.g. Schrodt et al. [2008], does not work consistently in all search engines.

[3]The nation-states dropped from the sample were: Andorra, Cyprus, Faroe Islands, Gibraltar, Greenland, Holy See, Iceland, Liechtenstein, Luxembourg, Malta, Monaco, San Marino.

[4]A meronym—the technical term used in *WordNet*—refers to a part of something, so "Norwegian" and "Oslo" are meronyms of "Norway."

[5]Since we did not know whether this would be removing too much information, we also created data sets containing all words, stop word removal only, and proper noun removal only. Because we are getting good

into character vectors with individual entries at the news story level. Using the *tm* package [Feinerer, 2007b, Feinerer et al., 2008], each character vector was then converted to an object of the class $Dataframe Source$—representing a data frame interpreting each news-story row as a document—using `data.frame()` and `DataframeSource()` and then to an object of class *Corpus* using `Corpus()` and `Corpus(VectorSource())`. Once the texts were converted to objects of class *Corpus*, the `tm_map()` command was applied to each corpus for for punctuation removal (`removePunctuation`), whitespace removal (`stripWhitespace`), lower-case conversion (`tolower`), and word-stemming (`stemDocument`).

The text were then preprocessed for stopword removal. Stopwords, as defined here, are the set of all common function words—such as "on", "the", "at", or "which"—listed within the `tm` package's internal stopword dictionaries. These were removed using `tm_map()` along with the commands `removeWords()`, `stopwords("english")`, and `stopwords("spanish")` to remove all English and Spanish stopwords from each document. Whitespace removal was then re-applied to the resultant corpora using (`stripWhitespace`).

We attempted to remove all proper nouns from each corpus. Proper nouns are defined as nouns representing unique entities such as "Italy" or "Silvio Berlusconi," rather than nouns referring to general classes of entities such as "country" or "politician". Our proper noun removal strategy combined two complementary approaches. First, we used the `openNLP` and `openNLPmodels.en` packages [Feinerer, 2007a] to apply the `tagPOS(,language="en")` command to each news-story in each text corpus. This command "tagged" each word in each corpus for its part-of-speech; with proper nouns tagged with either the `/NNP` or `/NNPS` identifier. We then used `grep()` and `sub()` to create unique word-vectors that stored only those words that were tagged as proper nouns by `tagPOS` (though we then removed the physical `/NNP` and `/NNPS` from the words in these vectors), for each nation-state. These nation-state specific proper-noun vectors were then used as "dictionaries" to remove all proper nouns from each nation-state corpus with the commands `tm_map()` and `removeWords()`.

Unfortunately, this strategy had limited effect on proper noun removal for our corpora, and many of the proper nouns in our corpora were not tagged as such by `tagPOS(,language="en")`.[6] In part, this was likely due to our reliance on the English-language version of `openNLP`, and it should be noted that alternate part-of-speech language dictionaries are currently in development [Feinerer and Hornik, 2010].

results with the more complete processing, we have not looked at these other sets in any detail at this point.

[6]This approach is also exceptionally slow. For example, to apply the proper-noun removal strategy outlined above to our "no proper nouns or stop words" Russia corpus, we had to split the Russia corpus into 47 different sub-corpora, and then run 47 separate proper-noun removal scripts on these sub-corpora, with each script taking roughly 24 hours to complete.

However, a second likely problem is that many of the proper nouns appearing in international news stories are simply too obscure to be identified and included within any general `openNLP` dictionary. Our texts—while all in English—make repeated references to politically-relevant names and geographic locations arising from over 60 nations and dozens of languages. Hence, many of the proper nouns in our texts are uncommon, non-English based, and spelled differently depending on the source. We therefore applied a second proper-noun-removal approach to our corpora in an effort to achieve greater proper noun removal using the actor and location names found in `CountryInfo.txt` to create a second proper noun dictionary.

`CountryInfo.txt`[7] is a general purpose file intended to facilitate natural language processing of news reports and political texts. In addition to a number of other pieces of relevant information, `CountryInfo.txt` records the country names, adjectival forms and synonyms (including some non-English spellings) of country names, capital cities, cities with populations over 1-million, regions and geographic features, leaders (1960-2008) and other members of government (2003-2010) for roughly 240 countries and administrative units. We included every `CountryInfo.txt` actor or location-name in the CountrInfo-derived proper nouns dictionary, and then augmented this proper noun collection with a number of highly salient non-state-actor proper nouns (e.g. "Osama Bin Laden", "Euro", "WTO"). We then applied this (second) proper noun word-vector to each of our preprocessed nation-state corpora using `tm_map()` and `removeWords()`.

As evident in the various top-word tables, this still doesn't completely work, and at various points one finds character strings that are in all likelihood the names of media sources and party abbreviations. These do not appear to be causing any serious problems, but in a later analysis we may try to simply remove every string that begins with an upper-case letter: this has the disadvantage of removing words that are capitalized at the beginning of the sentence, which sometimes will be relevant nouns (e.g., "Police confronted demonstrators..."). However the amount of text we are dealing with is sufficiently large that this would probably be less of a problem than the current system and, because it could easily be done in a separate Python or perl program, rather than inside R, it would be considerably faster.

4 Analysis

We estimated LDA using a series of R scripts. The `lexicalize()` command from the *lda*() package [Chang, 2011], was first applied to convert our nation-state and combined-

[7] Version 11.12.14. `http://eventdata.psu.edu/software.dir/dictionaries.html`.

nation corpora into LDA-readable corpora and vocabulary-vectors. Computation of LDA latent categories was done using the `lda.collapsed.gibbs.sampler()` where we first set the number of topics to be estimated to 10, and then to 20. The topics extracted under the 20-topic setting seems to produce classes that can be interpreted, and hence we focus on the 20-topic LDA models in our analysis below.

For these LDA models, the number of iterations used in LDA's Gibbs sampling were evaluated at values of 5, 10, 20, 50, 100, and 200. Values of 100 appeared to produce the most stable results across the ALL combined-country corpus while still completing within a reasonable (i.e. 24 hour) time period, and hence we favored iterations of 100 for the LDA results reported below. The remaining free parameters within the LDA models were set to default or otherwise reasonable values (e.g. alpha and eta set equal to 0.1). After estimation, the topics matrices from each LDA model was stored for subsequent analysis. These topics matrices include K rows for each extracted LDA topic (in our case, 20) and V columns for each unique word found within a corpus. The cell entries of the topic matrix then indicate the number of times a word (column) was assigned to a row-topic [Chang, 2011]. The analysis below, however, will focus on the top twenty words, which are generally sufficient to identify a clear general subject for the vector of words.

5 Results

Tables 1 through 6 show the top twenty words for the twenty latent topics for the combined set of all countries, for France and for Turkey. Comparable tables for the remaining cases that we have analyzed in detail are in the Appendix. This analysis uses data with the stop words and [most of] the proper nouns and meronyms removed through the process described above. The labels at the top of each column have been assigned by the authors to give a general sense of how we would characterize the cluster of words, and of course one could make differing interpretations. We did the analysis on the stemmed version of the words and so only those stems are reported. Categories which refer to specific countries are guesses based on the cluster of common words: the proper nouns themselves were usually not present.

Categories where it was impossible to determine a common theme were labeled "Comment": Remarkably for this type of analysis, there are only six of these in the 240 topics.[8] The non-political "Accident" category also occurs six times, and there are ten topics that refer to the media themselves: these may reflect clusters of "stories" that are dealing with secondary

[8]That or we are a little too good at seeing common topics where they don't exist...

coverage—probably announcements of related material—which could be removed with better filtering, or may be due in part to insufficient proper-noun removal.

Figure 1 shows the distribution of the assigned latent topic categories which occur three or more times.[9] The correspondence between latent topics and the event data categories is mixed. Several of the categories—meetings and other forms of negotiation and diplomacy, economic cooperation, and uses of violence—correspond clearly to high-frequency categories found in event data. The one exception is the [infamous] "Comment" category, which typically accounts for 20% to 30% of coded event data, but is relatively infrequent as a latent topic, particularly if the "media" categories are simply the result of filtering errors. This may be due to the fact that the content of comments varies widely, and consequently cannot be consolidated into distinct, high-frequency latent dimensions, but it may also indicate that some of the "Comments" in event data could be productively classified into more specific categories.

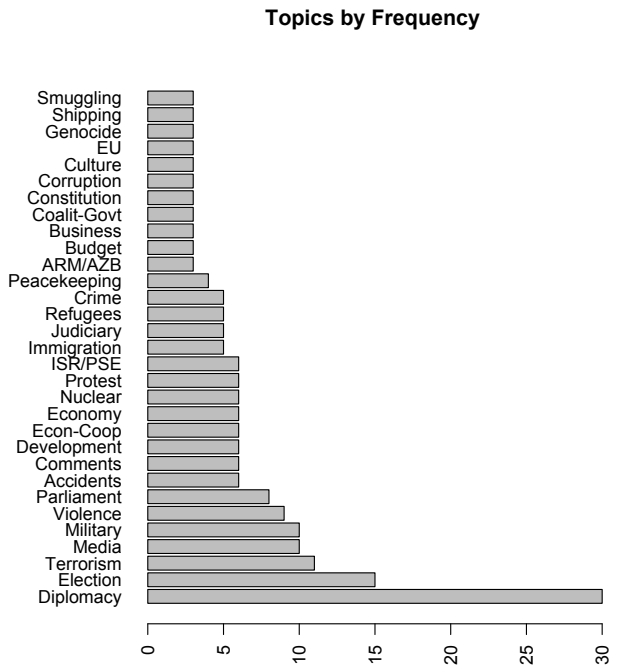

Figure 1: Frequency of Latent Topics

[9]Topics occurring two times: ceremony, Cyprus (for Greece and Turkey), drugs (though the "Smuggling" category frequently had a drug component), energy, Gaza (Israel and Egypt), international organizations, domestic meetings, military cooperation, rebellion, royalty (Jordan and Norway). Topics occurring only once: agriculture, cartoons (Norwegian coverage of anti-Islamic cartoons), democratization, disaster, history, hostages, human rights, humanitarian aid, missile attacks (Israel), Myanmar, negotiations, Nobel Prize (Norway), oil (Norway), police, settlements (Israel), Sri Lanka, travel, war crimes, and whaling (Norway)

At least as interesting, however, are the topics which appear to have wide and distinctive coverage but do *not* correspond to major categories in the existing systems, in large part because of the origins of WEIS in the study of international conflict. The most obvious of these are various institutional issues involving elections, parliamentary coalition formation and debate, budgeting and other aspects of parliamentary and legislative behavior, and judicial processes, particularly dealing with trans-national criminal behavior such as drugs and smuggling. None of these topics should come as any surprise to individuals who regularly read newspapers, but we are currently not coding them in any detail.

The latent topics produced by the analysis of the sample of reports for all of the countries combined—Tables 1 and 2—generally reflect the topics found with high frequency in the individual countries, in particular diplomacy, various aspects of elections, protests and parliaments, and various forms of violence, including terrorism, crime and accidents. These are consistent with ones intuitive sense of what is considered "newsworthy." The "nuclear" category is *probably* primarily dealing with nuclear proliferation issues involving North Korea, Iran and possibly Israel. Coverage of ceremonial and religious events gets a fairly distinct category, and the "smuggling" category probably encompasses coverage of both illegal drugs and various forms of human trafficking, including illegal immigration. The only incoherent topic is the one labelled "Comment": this possibly deals with "lifestyle" feature stories but, incongruously, has "world" and "war" as the top two words.

France and Turkey seem to be fairly typical for the country-level lists, and are a combination of the high-frequency general topics on diplomacy and violence which are found in almost every country we examined, and individual topics specific to the country. In the case of France, the EU and international organizations are distinct topics, although these are relatively unusual for the cases in general. The "culture" topic almost stereotypically says "France!"—film, culture, art, love (!), honor (!!)—but honest, that's what the LDA produced!—we didn't rig this. The "Travel" topic is again evidence of failed filtering and simply reflects the fact that France is one of the top tourist destinations for English-speaking travelers. The remaining topics are typical for the samples as a whole.

Turkey shows a similar mix. Here the idiosyncratic topics include what we infer to be stories about the on-going dispute in Cyprus (there is a comparable category in Greece) and what is probably the on-going controversy over the Ottoman genocide of Armenians in 1915 (there is a comparable category in Armenia). The "disaster" category may be combining natural disasters and—unexpectedly—the NGO flotilla which attempted to break the Israeli blockade of Gaza in May 2010. The EU predictably emerges as a category, and less predictably "energy" emerges as a separate category, only here and for Norway.

As a committed reader can discern for his or herself from Tables 9 through 26 in Appendix 2, this pattern of about a 3:1 ratio of general to specific topics is characteristic of almost all of the cases. These almost universally contain topics in the "Diplomacy", "Military/Violence", "Elections/Parliament" and one or two economic categories. Idiosyncratic topics, for the most part, seem to make sense given the issues relevant to the countries in the first decade of the twenty-first century.

Three countries stand out in our initial review in terms of their idiosyncratic topics. Completely expected is Israel, which has an unusually high number of topics—about half, depending on how one counts—specific to the Israeli-Palestinian conflict, though many of these are also found in Egypt and Jordan. There is also a "missile attacks" category reflecting attacks from both Hezbollah in 2006 and Gaza during much of this period, and the "nuclear" category probably reflects a combination of comments about Israel's own nuclear weapons, and Israel's concern about the Iranian nuclear program. Norway unexpectedly has nine topics which occur only in two or fewer other countries: anti-Islamic cartoons, energy, Myanmar, Nobel prize, oil, royalty, shipping, Sri Lanka, and whaling. This suggests that news stories specific to Norway are generally *quite* specific to Norway, particularly its oil and natural gas production, and controversial whaling program,[10] or mention Norway in the context of the Nobel Prize and various peacekeeping and mediation efforts.

Finally Egypt—which had a tightly controlled media during the period covered by our data—is interesting in terms of how boring and generic most of the categories are: five are the non-political"Media" (3 out of only 10 such instances in the sample), "Accident", or "Comment" and another three are the generic "Diplomacy" which appears to just cover routine meetings. This finding may generalize and be common in countries with heavy restrictions on the media, where reporters may be forced to rely primarily on secondary "he said, she said" reports (leaks, NGO reports, other media outlets) instead of on-the-ground reporting. The top words in the "media" category for ALL might also be evidence of this, as is to some extent Jordan.

[10]Schrodt is writing this in Oslo and notes that it is, in fact, a bit disconcerting to look at a restaurant menu and see "whale" among the main courses.

Table 1: Top Words for ALL (Topics 1-10)

Negotiation	Comments	Protest	Media	Accidents	Military	Econ-Coop	Diplomacy	Nuclear	Democracy
talk	world	protest	report	hospit	militari	cooper	agenc	nuclear	polit
meet	war	demonstr	newspap	kill	defenc	visit	news	weapon	reform
peac	global	violenc	conflict	plane	forc	meet	report	sanction	polici
summit	newspap	peopl	tv	crash	troop	trade	sourc	energi	countri
negoti	women	opposit	resolut	peopl	defens	develop	meet	secur	democrat
visit	via	attack	interview	die	command	econom	visit	resolut	democraci
leader	look	condemn	territori	injur	arm	discuss	excerpt	power	chang
discuss	write	ralli	publish	polic	oper	agreement	web	atom	futur
minist	polit	right	issu	accid	air	invest	websit	program	posit
middl	life	polic	statement	fire	mission	bilater	minist	plant	govern
issu	home	terror	polit	passeng	peacekeep	deleg	tv	uranium	societi
diplomat	campaign	call	independ	children	deploy	project	site	enrich	presid
offici	littl	leader	websit	victim	soldier	energi	omit	destruct	develop
presid	daili	govern	situat	death	armi	sign	passag	diplomat	econom
agreement	time	activist	settlement	flight	train	relat	radio	talk	institut
foreign	live	demand	intern	car	secur	countri	deleg	inspector	process
agre	play	war	region	medic	exercis	industri	presid	programm	issu
process	stori	human	posit	caus	base	busi	ambassador	world	integr
plan	appear	support	news	road	missil	minist	discuss	threat	stress
confer	week	forc	peac	airport	staff	educ	present	arm	access

Table 2: Top Words for ALL (Topics 11-20)

Parliament	Diplomacy	Violence	Crime	Terrorism	Election	Parliament	Ceremony	Economy	Smuggling
elect	cooper	kill	court	attack	vote	law	ceremoni	bank	border
parti	relat	attack	crime	terrorist	percent	right	anniversari	percent	polic
democrat	meet	forc	charg	arrest	elect	agreement	visit	fund	drug
vote	countri	soldier	prison	suspect	parti	visa	cultur	budget	ship
opposit	develop	troop	trial	terror	referendum	amend	church	compani	illeg
candid	bilater	rebel	war	polic	parliament	human	celebr	oil	guard
parliamentari	region	fire	sentenc	bomb	poll	refuge	peopl	tax	cross
coalit	visit	bomb	prosecutor	kill	govern	draft	attend	market	traffick
presidenti	discuss	wound	former	secur	constitut	citizen	award	economi	smuggl
parliament	issu	ethnic	investig	investig	coalit	legal	event	financ	arrest
poll	secur	milit	tribun	intellig	opposit	document	mark	financi	seiz
deputi	stabil	villag	arrest	releas	join	bill	presid	price	port
polit	summit	militari	alleg	hostag	reform	adopt	day	gas	detain
chairman	econom	border	accus	report	bloc	legisl	memori	econom	custom
leader	integr	armi	judg	kidnap	treati	committe	famili	product	water
presid	presid	provinc	lawyer	milit	minist	issu	citi	rate	boat
seat	talk	citi	murder	offici	voter	govern	mosqu	export	oper
voter	support	civilian	crimin	embassi	polit	sign	ambassador	money	vessel
socialist	express	polic	suspect	alleg	candid	approv	spa	growth	agenc
rule	intern	northern	convict	prison	democrat	regul	histori	invest	immigr

Table 3: Top Words for France (Topics 1-10)

Travel	Immigration	IOs	Diplomacy	Terrorism	Nuclear	Culture	Violence	EU	Business
plane	immigr	resolut	meet	attack	nuclear	film	polic	constitut	compani
flight	right	war	visit	kill	sanction	ceremoni	kill	treati	industri
air	human	secur	talk	hostag	weapon	wife	fire	propos	project
crash	law	militari	discuss	terror	atom	cultur	car	vote	energi
ship	report	weapon	summit	kidnap	enrich	celebr	injur	bloc	oil
passeng	illeg	action	minist	releas	secur	book	bomb	referendum	invest
airlin	ban	nation	presid	bomb	uranium	world	peopl	summit	contract
train	freedom	inspector	leader	milit	program	award	attack	presid	market
tunnel	journalist	disarm	offici	terrorist	resolut	life	explos	parliament	bank
airport	border	forc	foreign	rebel	energi	art	citi	reject	busi
board	terror	peac	attend	militari	negoti	histori	town	reform	deal
crew	terrorist	draft	counterpart	soldier	programm	anniversari	die	enlarg	ead
aircraft	countri	diplomat	issu	forc	talk	festiv	hospit	negoti	fund
coast	threat	intern	invit	word	diplomat	presid	death	membership	gas
transport	govern	council	deleg	arm	inspector	famili	dead	draft	trade
port	newspap	alli	trip	leader	power	memori	wound	bill	firm
boat	media	crisi	arriv	dead	bomb	centuri	near	talk	product
water	citizen	veto	bilater	suicid	inspect	love	build	leader	sign
pilot	letter	vote	cooper	condemn	activ	honour	vehicl	approv	agreement
land	inform	support	met	guerrilla	meet	former	accid	debat	sector

Table 4: Top Words for France (Topics 11-20)

Election	Peacekeeping	Protest	Judiciary	Economy	Military	Terrorism	Election	Development	Diplomacy
elect	rebel	protest	court	percent	militari	polic	minist	cooper	radio
vote	forc	polic	trial	growth	forc	arrest	elect	develop	international
poll	troop	riot	investig	budget	troop	suspect	presid	countri	report
parti	peac	demonstr	former	economi	defenc	investig	polit	summit	peac
presidenti	armi	youth	judg	econom	defens	children	parti	global	sourc
percent	war	violenc	sentenc	deficit	deploy	attack	prime	meet	excerpt
candid	govern	street	alleg	financ	oper	charg	socialist	econom	present
socialist	fight	suburb	lawyer	market	soldier	detain	govern	world	broadcast
round	ceasefir	school	charg	tax	peacekeep	terrorist	presidenti	intern	presid
voter	town	student	prison	rate	command	alleg	former	relat	ambassador
polit	soldier	union	genocid	financi	air	prison	resign	confer	process
campaign	militari	peopl	accus	bank	mission	murder	job	bilater	news
victori	former	citi	crime	cut	missil	sourc	conserv	visit	meet
win	conflict	strike	prosecutor	trade	aircraft	prosecutor	candid	issu	yesterday
segolen	countri	march	arrest	price	arm	judici	campaign	cultur	minist
conserv	coloni	law	jail	product	secur	hospit	polici	discuss	agenc
seat	coup	unrest	human	reform	armi	drug	reform	leader	talk
won	capit	thousand	extradit	global	base	court	cabinet	polit	visit
elector	civilian	night	appeal	zone	exercis	sentenc	leader	partnership	polit
leader	leader	immigr	justic	minist	helicopt	terror	appoint	region	websit

Table 5: Top Words for Turkey (Topics 1-10)

Accidents	Terrorism	Parliament	Diplomacy	Disaster	Election	EU	Military	Diplomacy	Cyprus
plane	polic	parti	talk	aid	elect	membership	troop	cooper	island
hospit	attack	parliament	visit	attack	parti	negoti	war	agreement	northern
crash	kill	elect	diplomat	ship	vote	access	militari	sign	solut
peopl	bomb	deputi	relat	earthquak	polit	talk	forc	visit	denkta
kill	explos	chp	minist	peopl	rule	reform	deploy	deleg	republ
fire	terrorist	law	foreign	quak	govern	process	secur	defens	plan
injur	injur	justic	meet	kill	parliament	summit	soldier	defenc	talk
passeng	peopl	ak	negoti	condemn	secular	bloc	weapon	agenc	leader
accid	soldier	leader	tie	humanitarian	poll	entri	send	meet	negoti
rescu	blast	ecevit	offici	flotilla	opposit	join	alli	countri	divid
hostag	wound	republican	nuclear	exercis	justic	right	defens	report	settlement
ship	citi	articl	countri	air	percent	enlarg	northern	news	issu
citi	provinc	amend	issu	raid	constitut	criteria	resolut	sourc	peac
provinc	clash	mhp	discuss	plane	presid	start	plan	discuss	reunif
town	diyarbakir	vote	peac	intern	candid	decis	govern	minist	communiti
die	secur	govern	presid	condol	ak	progress	command	relat	north
train	southeastern	resign	gl	activist	ban	countri	peacekeep	tv	meet
villag	demonstr	democrat	daili	disast	democrat	date	base	protocol	agreement
bodi	agenc	decis	counterpart	statement	reform	chapter	oper	particip	tasso
miss	town	dsp	process	water	parliamentari	bid	decis	inform	referendum

Table 6: Top Words for Turkey (Topics 11-20)

Business	Diplomacy	Terrorism	Smuggling	Diplomacy	Ceremony	Judiciary	Genocide	Development	Energy
econom	meet	terrorist	illeg	relat	visit	court	genocid	trade	market
economi	visit	rebel	polic	peac	gen	prison	newspap	cooper	oil
meet	attend	northern	provinc	countri	staff	prosecutor	world	countri	percent
businessmen	deleg	terror	border	region	command	charg	peopl	relat	nuclear
countri	talk	attack	migrant	note	patriarch	trial	daili	energi	compani
invest	minist	worker	oper	develop	meet	sentenc	publish	econom	export
sector	hold	militari	forc	stabil	met	arrest	articl	project	gas
note	recep	fight	secur	cooper	ceremoni	alleg	women	tuzmen	bank
develop	confer	border	immigr	support	ambassador	right	word	invest	energi
program	met	oper	detain	import	press	suspect	histori	visit	pipelin
busi	receiv	region	arrest	stress	church	human	countri	bilater	price
industri	expect	forc	agenc	confer	close	investig	live	agreement	suppli
project	held	organ	town	process	attend	former	empir	region	fund
social	summit	arm	heroin	issu	religi	lawyer	polit	transport	lira
organ	host	separatist	captur	contribut	militari	polic	call	develop	economi
financi	arriv	troop	drug	peopl	chairman	jail	public	volum	plant
tourism	offici	soldier	deport	world	consul	stori	critic	sign	sanction
crisi	ambassador	crossbord	kg	intern	forc	accus	book	meet	deal
chamber	discuss	base	sourc	effort	mustafa	ocalan	religi	talk	financi
reform	news	armi	seiz	state	celebr	crime	kill	businessmen	rate

6 Hierarchical Analysis of the Clusters

The second part of our analysis will look at whether there is any structure to the latent topics both within countries and, more importantly, across countries. The latter issue is critical to the question of whether topics are truly general—the implicit assumption of event data—or whether they are distinctive to individual cases. We will use hierarchical clustering—the *R* routine `hclust()`—for this purpose.

This clustering requires measuring the similarity between the sets of words in the vectors. One approach would be to use an existing dataset of word-distances—for example the *WordNet* similarity measures available at `http://wn-similarity.sourceforge.net/`. However, these involve multiple word senses, which we have not determined in our preprocessing, and in many instances, the stemming we have done makes it impossible even to associate a string with a unique word.

As an alternative, we have derived a measure of word similarity—or more accurately, character-string similarity, but we will call them "words" for convenience—from the data itself. The latent topic models should be co-associating words that are dealing with the same general topic. We therefore first measured the number of times a word-pair occurred within any one of the 240 topic vectors in the cases we are examining closely, and then use that number as the measure of the similarity of the two words in the pair. Table 8 in the Appendix shows about 125 of these pairs with the highest weights.

We then define the distance between two vectors as a constant minus the sum of the similarity weights[11] for the word-pairs generated by taking all combinations of the words in the two vectors. For example, if one vector contains "meet" and the other contains "leader", we subtract 15 from the distance score; if the other vector contains "discuss", we subtract 29.[12]

Figures 2 through 6 show the results of this clustering in the form of dendograms. Figures 2 through 4, for the combined sample, France and Turkey, are essentially a validity check on the distance measure itself, though they also indicate some clustering of the vectors. For the most part, these are quite plausible. In the combined sample, Figure 2 , most of the topics fall into unambiguous clusters: from left to right, diplomacy (3), an unclear cluster (3), parliament/elections (2), ceremonial (2), probably economics (3), crime/terrorism (3), and violence (4). The main anomaly here is the two "Parliament" topics are split, possibly

[11] Arbitrarily set at 2000: we need this operation to convert the *similarity* measures of the word-pairs to the *distances* used by the clustering algorithm

[12] If the same word occurs in both vectors, we subtract 30: this is an arbitrary constant and was set at the level of the most similar words.

because one deals with party politics (first column in Table 2) and the other with the topics of debate (seventh column). Similarly, the first of the two "Diplomacy" vectors corresponds to the second column in Table 2 and seems involve substantive diplomacy, whereas the second is column eight of Table 1 and seems to mostly involve media coverage. The clusters generally reflect a single cooperation-conflict dimension.

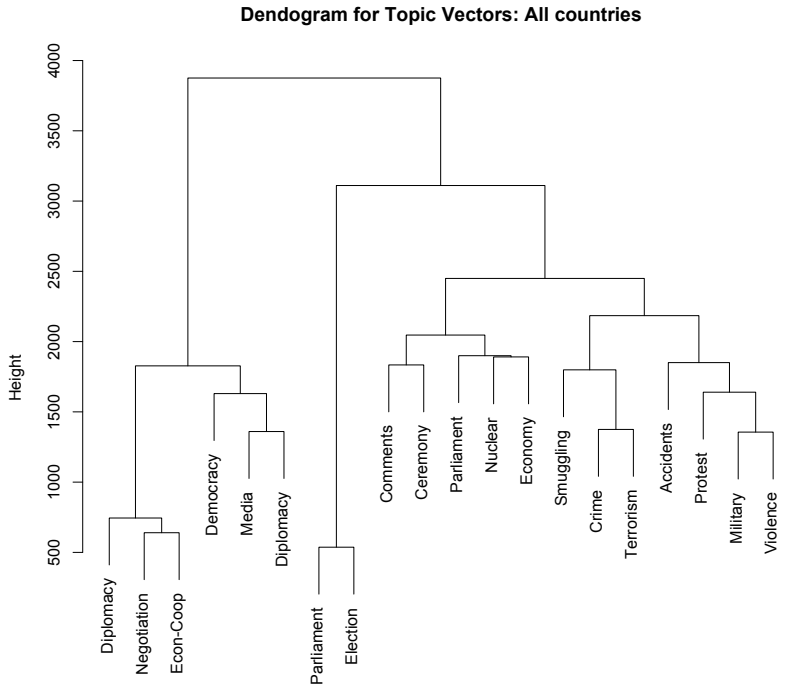

Figure 2: Clustering of topic vectors: All countries

The clusters for France and Turkey are even more straightforward. In the case of France, the clusters from left to right are elections (2), business (2), EU/diplomacy (5), peacekeeping (3), legal (2), violence (2), protest/immigration (2) and travel/culture (2). The clustering seems to differentiate different types of "diplomacy" topics—development, EU and peacekeeping—as well as separating the prosecution of terrorism from coverage of acts of terrorism. Unlike the all-countries case, or Turkey, a cooperation-conflict dimension is less evident here. Turkey, in contrast, shows a very strong cooperation-conflict dimension, with diplomacy on the left end, terrorism, criminal activity and disasters on the right, and the legislative and business activity in the middle.

Finally, Figures 5 and 6 show clustering for the four European cases and the four Middle Eastern, plus the combined sample (`ALL`). The country identification and number of the vector is given in addition to its assigned topic: for example `GRC-5` refers the fifth column of

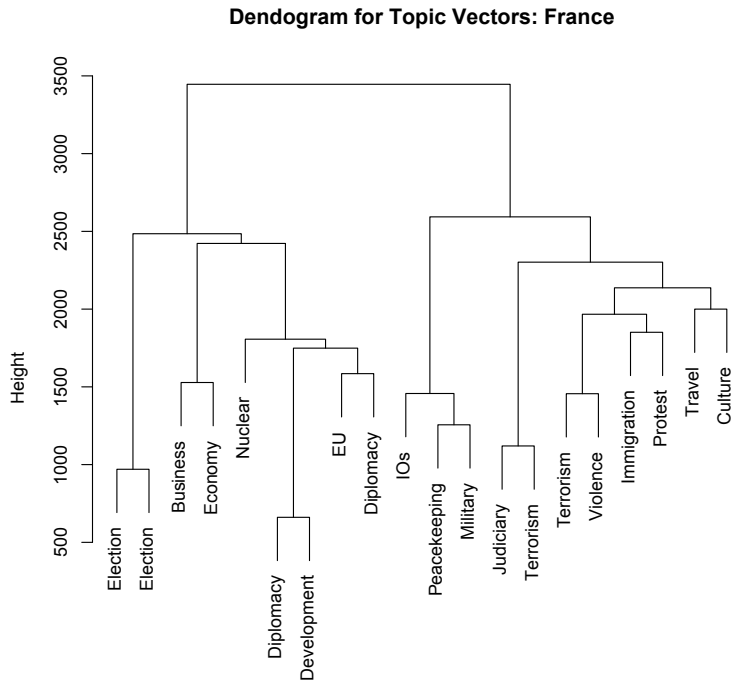

Figure 3: Clustering of topic vectors: France

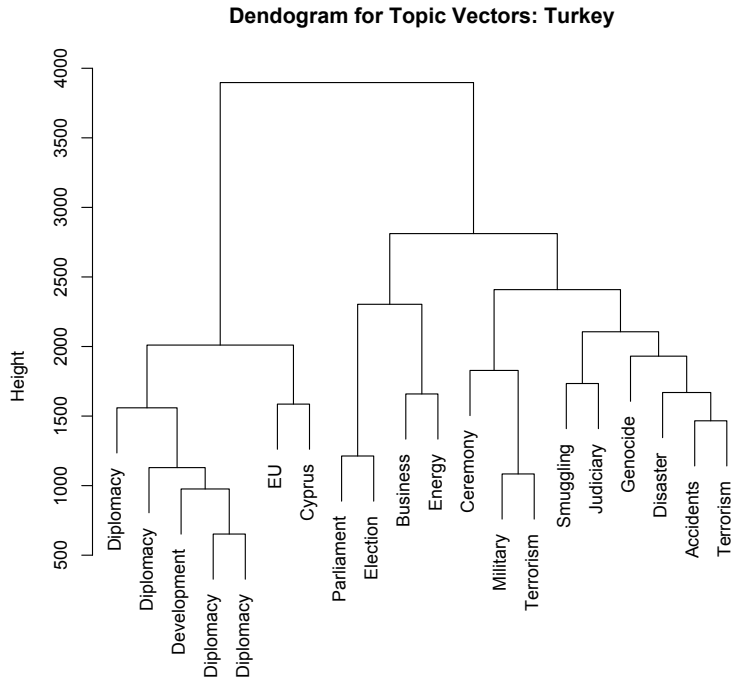

Figure 4: Clustering of topic vectors: Turkey

Table 9. The large-scale categories listed across the bottom of each graph were determined by the authors—though in most cases, these are quite unambiguous—and the splits between these (vertical dashed lines) always correspond to major splits within the graph itself.

The key question we had was whether the topic clusters would generalize or whether they would be specific to each country, and whether the combined clusters from the `ALL` data would correspond to a structure across the various countries. Figures 5 and 6 provide very strong support for general structures: there is virtually no evidence for country-specific clustering except for some of the "Governance" categories (e.g. Norway) for Europe. The European cluster also generally shows a cooperation-conflict dimension, though less strongly than the `ALL` and Turkey cases; somewhat surprisingly, this does not seem as evident in the Middle Eastern case. Furthermore, without exception, all of these larger clusters contain at least one of the `ALL` topics—a very strong result given that the `ALL` topics are only 20% of the cases—which suggests that these do in fact provide archetypes for the topics found in the individual states.

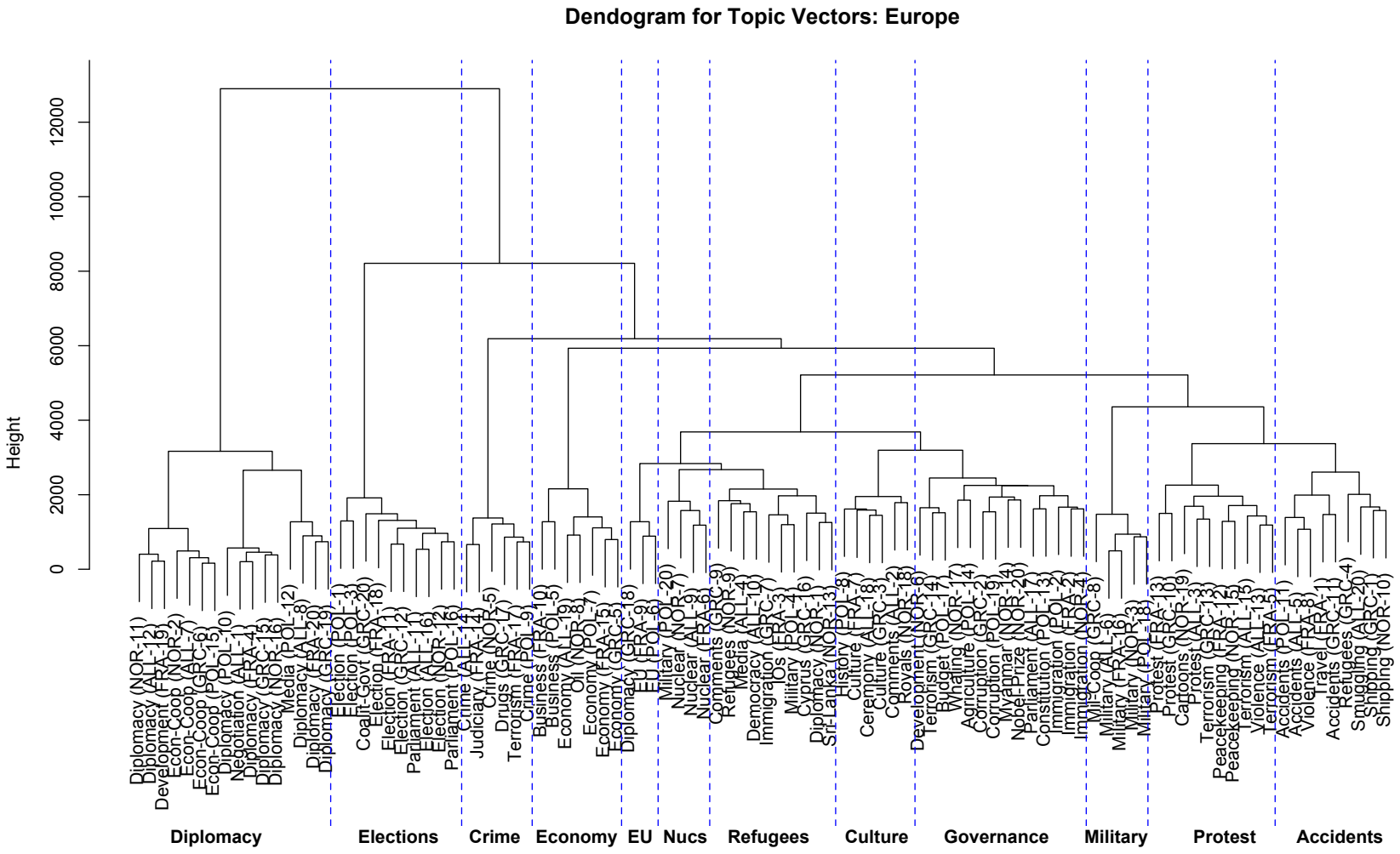

Figure 5: Clustering of topic vectors: Europe

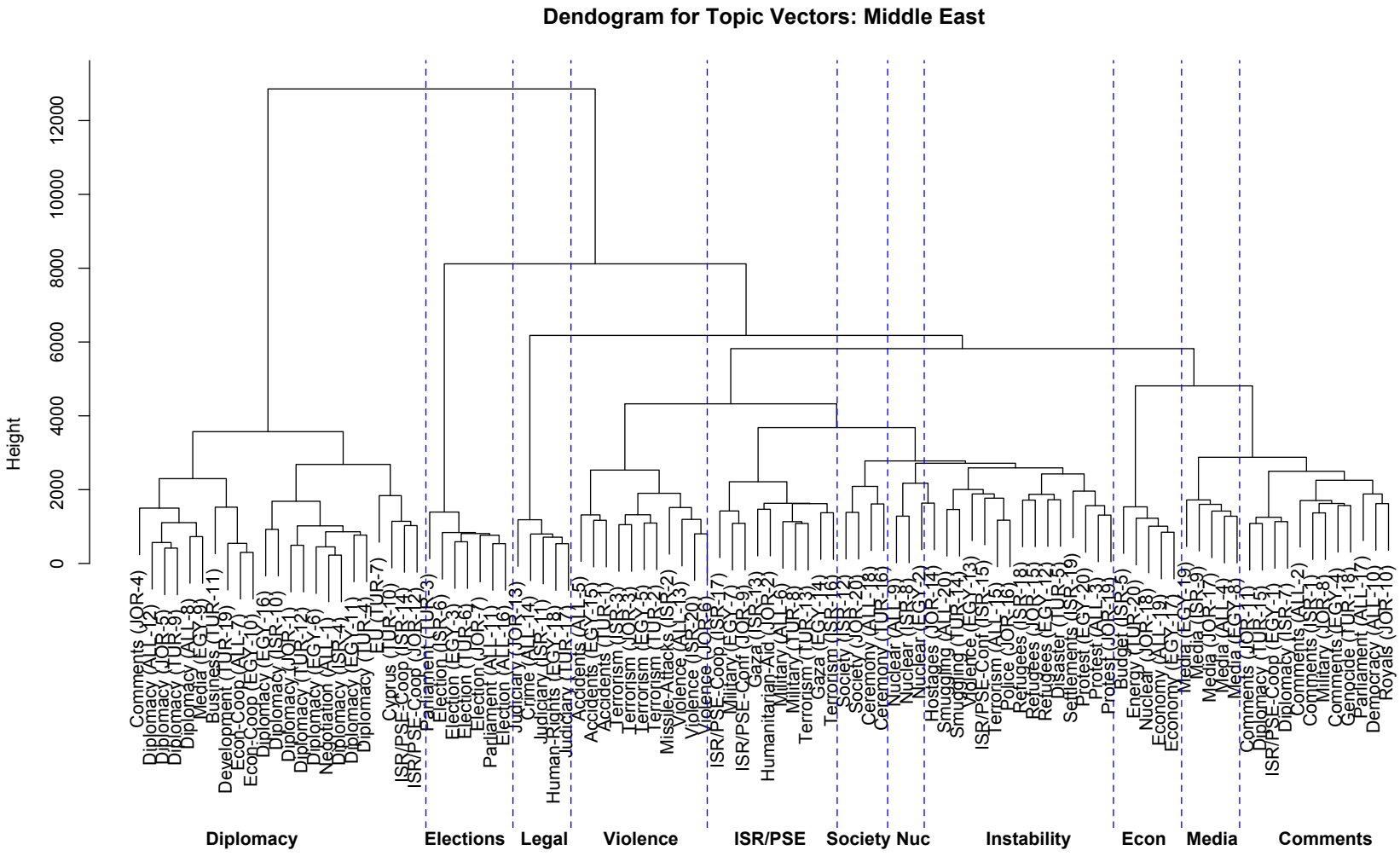

Figure 6: Clustering of topic vectors: Middle East

7 Conclusions

In general, we feel that this analysis has largely reinforced the legitimacy of the event data approach to summarizing political interactions by the categorization of news wire story texts, but at the same time indicated some major topics that existing systems are missing. We conclude this paper by summarizing our major findings, as well as suggesting a couple of possible applications for this approach.

Our primary result, of course, is that news wire stories do cluster into very plausible topics that are generally consistent with those assumed in the major event data ontologies. This result has been implicit in the event data approach from the beginning, but to our knowledge has never been tested using purely objective inductive methods. Furthermore, the fact that news stories appear to be clustered when read by humans—who comprehend the stories in the context of a deep body of linguistic, historical and cultural knowledge—does not mean they would cluster from the vantage point of a machine which is simply looking at the stemmed character strings. Yet they do, which is particularly important for fully-automated coding systems.

The majority of these topic clusters are general rather than country-specific, though almost all of the countries we examined had at least some country-specific topics. The results of the cluster analysis on the topic vectors also suggests that most of the topics found in individual countries are reasonably consistent with a topic from the analysis of the combined sample. Again, this has been implicit in the event data approach, whether in machine-coded or human-coded systems, but never explicitly tested. We also generally find support for the cooperation-conflict dimension implicit in the existing coding systems.

While the topic clusters are certainly consistent with event data coding, they depart from the existing approaches in two ways. First we did not find—nor did we expect to find—anything as specific as the sub-categories of any of the coding schemes, and some of the clusters that we did find probably cross multiple top-level 2-digit "cue categories" of WEIS, IDEA and CAMEO. Second, there are a substantial number of categories, notably those dealing with legislative and electoral behavior, crime, and some economic behavior, that do not correspond to categories in the existing schemes. This is not a failure of those schemes—they made no attempt to code such behaviors, as they were initially focused primarily on interstate behavior, and more recently intra-state conflict—but does suggest that considerably more could be coded from these reports than is currently being coded.

This latter finding has implications for event data coding more generally. While actor-

dictionaries and action-categories readily exist for coding political conflict and cooperation (e.g., bombings or verbal threats) between pairs of political actors (e.g., politicians and rebel groups), we lack comparable resources for the automated-coding of many other highly salient types international events. Business interactions between firms and governments, electoral externalities, criminal activities short of violence (e.g., illicit trades), and environmental disasters are but a few examples. Developing complete dictionaries and event-categories for each of these areas would involve considerable time and effort, a barrier to entry for many social scientists. Our approach could provide interested researchers with a technology leapfrog opportunity: LDA could be applied to corpora of relevant news stories—*when the full set of behaviors of interest are unavailable a priori*—to identify and code the most relevant of these, as well as their variation across time and space.

This is an important development with respect to the event/behavior categories. There is a large and well-developed body of techniques—"named-entity recognition"—for the identification of actors within a set of texts, as well as numerous sources for lists of actors such as multinational corporations, NGOs/IGOs, and political leaders. However, except for *WordNet*, which deals only with words, not phrases, no equivalent list of vocabulary exists for the identification of categories of behavior, and here LDA could be very useful. Because there is no syntactic component to LDA—it is simply looking at character strings—it could also be useful in the rapid characterization of news reports in languages other than English.

In the above analysis, we have not tried to classify the original texts into the topic categories, nor compared these with the results of automated event data coding. Classification is a straightforward procedure in LDA, as is automated coding using a system such as TABARI, and we intend to do so in the next iteration of our analysis. Generally, however, the latent topics seem to correspond to the high-frequency event data categories, with the possible exception of the generic "Comment" category, which seems to occur less frequently as a topic in the latent categories than it occurs in coded event data. However, when we do story classification, "Comment" may emerge as a residual category for stories which don't really fit any of the primary topics. Pretty much what it is in event data as well...

The coherence and plausibility of the latent topics opens the possibility of using these, rather than some *a priori* coding schemes, as the framework for event data coding, particularly if these latent topics could be shown to be stable over time, a dimension that we also intend to explore in a later extension of this work. The advantage of such an approach is that it would be based on the actual information available in the news reports, rather than assuming that reports would distinguish between types of activity arbitrarily specified in a codebook.

To an extent, we have already taken this approach in most of the event data analysis undertaken by the KEDS/TABARI-based projects at Kansas and Penn State. In almost all of these analyses, we use only the very broad categories of material cooperation, verbal cooperation, verbal conflict and material conflict, a typology which is even more general than the cue categories. While we have done some work with the 22 WEIS cue categories [Schrodt, 2006], we have not done any work at a detail below the cue categories, nor are we aware of any studies which worked at that level, except in aggregating the more detailed categories using the Goldstein [1992] scale or variations on it. If that is the case, the existing systems—human or machine—are trying to code *much* more information than is actually being used and, conversely, the automated coding systems may already be extracting almost all of the information needed for analysis, even if the event-by-event coding accuracy of those systems is still lower than human coding.[13]

However, while an inductive system such as LDA could substitute for the event *categories*, syntactically-based coding is still required to correctly identify the *source* (subject) and *target* (object) of the event. It may be possible to do this using general-purpose systems such as open-source parsers such as `http://opennlp.sourceforge.net/projects.html`, though whether there is sufficient information in the sentence-level units which yield subject and object information to also do classification using an LDA remains to be seen. LDA (or other latent topic) methods can certainly inform and refine the existing dictionary-based coding systems, but it is less clear that they can replace them.

That said, there would appear to be at least three clear applications of this method for improving existing coding.

First, despite the fact that the corpus of stories we were working with was supposed to only contain political interactions—and we would note with great relief that none of the latent topics concerned sports, a major problem in some downloads—there are still a sufficient number of non-political stories to generate some non-political latent topics, notably on accidents, business stories, and some amorphous "lifestyle/culture" categories. Classification using LDA might be a way of getting rid of these, though they might also be harmless: That could be easily ascertained by coding a set of stories so identified and seeing whether they generate any political events.

Second, this method would be very useful in a post-processing step where stories that did not generate events are checked to see whether the latent topics indicate that they *should*

[13]Or at least human coding that is done slowly and carefully while initially validating the coding protocol: This level of accuracy may not extend to data coded subsequently under less ideal circumstances [Mikhaylov et al., 2012].

have generated an event. Some subset of those could then be examined to determine whether they indicate the need for additional actor or event-phrase vocabulary.

Conversely, the latent topics could be used as a check on coded stories to see whether the machine-coding classification is consistent with the general topic assigned by an LDA classification. This may be less definitive: It is easy to imagine, for example, a single story that would generate both diplomacy and violence events because it reported on both an armed clash and an attempt to mediate the underlying dispute. If anything such stories are the norm rather than the exception in many conflicts. However, some subset of these stories might again lead to the identification of missing vocabulary.

This initial analysis also suggests several lines for future work. First, in addition to looking at the stability of topics over time, LDA could be used to compare how well topics correspond across various media outlets. For example, does AFP report proportionately more violence than Reuters? Does AFP report more country specific topics, while Reuters reports more general topics? How do the international sources compare to local and regional sources? What is the correspondence, or lack thereof, between news sources and social media data (e.g. Twitter, Facebook): Which topics of political interaction are similarly reported in both types of media and which are better reported in the news over social media?

Second, while in this discussion we have treated the country-specific topics as something of a nuisance, in fact they might be useful as extensions to the general event data categories, particularly if the relevant stories and vocabulary can be developed with relatively little effort, followed by event coding with automated methods. "One size fits all" is not how qualitative analysts look at politics, nor is it necessarily the best approach for event data. Our findings provide some support for both the generalists and "area studies" and LDA approaches could begin to allow scholars to have the best of both worlds.

Finally, a useful extension might be the application of "Nubbi LDA" [Chang et al., 2009], which can uncover relationships and networks between actors in a text; such as political actors in Wikipedia articles and individuals in the Bible. The actors are specified *a priori* but this is straightforward using existing actor dictionaries from various event data projects or NER software. Such an approach would allow us to begin to map the networks to countries in the world—and to identify the latent political interactions that underly these networks—via the news stories. More generally, this would allowing one to exploit the flexibility of LDA in identifying topics while using TABARI to make up for LDA's deficiency in identifying source and target, as well as providing a social network analysis component to the project.

8 Appendix

Table 7: European and Middle Eastern nation-states included in the sample

Nation States		
Afghanistan	Iran	Qatar
Albania	Iraq	Romania
Armenia	Ireland	Russia
Austria	Israel	Saudi Arabia
Azerbaijan	Italy	Serbia
Bahrain	Jordan	Slovakia
Belarus	Kazakhstan	Slovenia
Belgium	Kuwait	Spain
Bosnia & Herzegovina	Kyrgyzstan	Sweden
Bulgaria	Latvia	Switzerland
Croatia	Lithuania	Syria
Czech Republic	Lebanon	Tajikistan
Denmark	Macedonia	Turkey
Egypt	Moldova	Turkmenistan
Estonia	Netherlands	Ukraine
Finland	Norway	United Arab Emirates
France	Oman	United Kingdom
Georgia	Pakistan	Uzbekistan
Germany	Palestine	Yemen
Greece	Poland	
Hungary	Portugal	

Table 8: Word-pair similarity scores

Score	Word pairs
31	meet-talk
29	discuss-meet
26	meet-visit
22	discuss-visit meet-minist
21	discuss-minist
20	forc-militari
19	countri-region discuss-talk elect-parti minist-talk agenc-news
18	talk-visit countri-develop leader-talk agreement-meet countri-meet deleg-meet
17	meet-summit minist-visit attack-kill parti-polit deleg-discuss minist-presid countri-econom cooper-meet negoti-talk cooper-countri countri-relat summit-talk
16	elect-vote parti-vote report-sourc cooper-visit forc-troop meet-region agenc-report
15	leader-meet meet-relat forc-secur militari-troop parliament-vote opposit-parti attack-polic kill-peopl attack-terrorist relat-visit parliament-parti meet-peac news-report presid-talk cooper-discuss develop-meet
14	elect-parliament arm-forc agreement-sign agreement-talk meet-presid cooper-relat bilater-visit countri-visit issu-meet develop-econom agenc-sourc meet-negoti forc-soldier kill-polic cooper-develop develop-relat soldier-troop elect-polit govern-parti
13	news-sourc peac-talk alleg-investig deleg-visit polic-suspect attack-bomb meet-news investig-polic arrest-investig bilater-cooper democrat-parti attack-terror offici-talk agreement-countri discuss-issu leader-summit countri-discuss meet-report bilater-meet candid-elect cooper-econom elect-govern elect-opposit discuss-summit arrest-polic militari-soldier court-investig countri-polit
12	arrest-prison discuss-relat bilater-econom opposit-vote peopl-polic elect-leader develop-visit bomb-kill minist-offici leader-parti arrest-suspect arm-militari bilater-develop alleg-court deleg-minist countri-trade armi-troop leader-minist candid-parti investig-suspect report-websit countri-peopl discuss-offici charg-court govern-polit foreign-minist govern-leader meet-offici bilater-discuss govern-percent agenc-meet opposit-parliament discuss-leader peac-war democrat-elect agreement-discuss econom-economi

Table 9: Top Words for GRC (Topics 1-10)

Shipping	Corruption	Culture	Refugees	Economy	Econ-Coop	Immigration	Mil-Coop	Comments	Protest
ship	court	cultur	border	debt	cooper	immigr	defenc	issu	polic
island	investig	visit	aid	budget	relat	right	militari	name	attack
immigr	justic	event	cross	financ	visit	human	forc	solut	bomb
illeg	prosecutor	ceremoni	hospit	economi	bilater	war	defens	republ	protest
coast	charg	attend	road	deficit	econom	countri	energi	posit	fire
boat	former	anniversari	water	percent	countri	intern	arm	spokesman	demonstr
vessel	corrupt	citi	flood	econom	develop	secur	cooper	negoti	explos
earthquak	typo	celebr	evro	market	meet	illeg	agreement	stress	riot
port	alleg	ancient	region	fund	trade	asylum	staff	regard	car
migrant	resign	church	farmer	financi	invest	peac	gas	access	damag
quak	public	world	prefectur	crisi	tourism	refuge	exercis	foreign	embassi
crew	rule	centuri	doctor	rate	busi	polici	pipelin	question	offic
guard	minist	memori	traffic	govern	region	world	sign	former	youth
marin	file	honour	local	tax	tie	terror	oil	govern	outsid
merchant	scandal	exhibit	humanitarian	bank	agreement	stress	oper	koumoutsako	build
mile	judg	communiti	tourist	cut	discuss	peopl	visit	note	injur
author	compani	hellen	project	measur	minist	border	command	ad	anarchist
water	law	marbl	food	zone	sector	crisi	air	refer	caus
damag	judici	ambassador	health	auster	sign	global	project	relat	street
rescu	sunday	sport	home	growth	energi	nuclear	meimaraki	reiter	march

Table 10: Top Words for GRC (Topics 11-20)

Accidents	Election	Terrorism	Terrorism	Diplomacy	Cyprus	Drugs	Diplomacy	Diplomacy	Coalit-Govt
fire	elect	patriarch	educ	meet	island	polic	talk	news	parti
plane	parti	church	ministri	visit	talk	arrest	access	agenc	opposit
crash	percent	attack	social	minist	leader	charg	membership	meet	leader
firefight	vote	terrorist	minist	discuss	plan	prison	negoti	report	govern
forest	poll	condemn	bill	simiti	divid	suspect	summit	cooper	democraci
flight	socialist	terror	govern	talk	denkta	drug	join	sourc	polit
aircraft	opposit	right	fund	costa	solut	court	bloc	confer	elect
passeng	democraci	war	programm	issu	northern	illeg	leader	secur	costa
air	conserv	govern	reform	presid	negoti	sentenc	sign	discuss	coalit
helicopt	candid	ethnic	public	foreign	republ	traffick	start	region	main
jet	rule	violenc	sector	attend	peac	found	disput	websit	polici
accid	voter	properti	administr	prime	reunif	alleg	diplomat	ambassador	stress
kill	opinion	express	employ	met	settlement	trial	entri	countri	movement
pilot	seat	ecumen	economi	leader	north	prosecutor	agreement	presid	communist
blaze	parliament	militari	system	deleg	meet	murder	bid	mia	rule
train	costa	human	draft	counterpart	coup	convict	enlarg	process	left
near	govern	minor	environ	offici	communiti	extradit	protocol	organ	simiti
airlin	win	newspap	law	summit	referendum	author	port	repres	social
bus	leader	peopl	school	arriv	rauf	immigr	custom	anampa	spokesman
wildfir	coalit	journalist	labour	brief	reunit	investig	presid	assembl	address

Table 11: Top Words for NOR (Topics 1-10)

Diplomacy	Econ-Coop	Military	Immigration	Crime	Development	Nuclear	Oil	Refugees	Shipping
peac	cooper	militari	asylum	polic	develop	nuclear	oil	aid	ship
global	agreement	defenc	court	arrest	global	weapon	compani	refuge	rescu
via	oil	forc	law	suspect	fund	submarin	bank	humanitarian	vessel
intern	energi	defens	embassi	munch	countri	program	percent	peopl	crew
conflict	gas	troop	immigr	court	via	atom	mn	tsunami	water
resolut	sign	air	rule	paint	project	test	rate	food	boat
agreement	trade	plane	report	investig	donor	plant	invest	relief	coast
secur	barent	oper	author	attack	aid	missil	price	help	trawler
polit	develop	peacekeep	decis	charg	govern	secur	nok	countri	accid
support	countri	fighter	visa	scream	support	agenc	market	displac	fish
nation	talk	aircraft	refuge	crime	programm	wast	fund	disast	passeng
process	industri	command	countri	sentenc	educ	radiat	gas	nation	fire
negoti	border	exercis	alleg	stolen	corrupt	destruct	crown	intern	board
effort	region	soldier	govern	terror	poverti	inspector	energi	donor	polic
issu	discuss	secur	diplomat	theft	financi	document	tax	victim	mile
communiti	minist	mission	investig	prison	assist	inspect	export	camp	seeker
observ	bilater	train	applic	masterpiec	provid	programm	onlin	miss	captain
solut	econom	flight	claim	robberi	financ	energi	product	emerg	asylum
parti	visit	deploy	ministri	museum	money	mass	economi	govern	oil
war	resourc	send	broadcast	found	pledg	dispos	budget	region	helicopt

Table 12: Top Words for NOR (Topics 11-20)

Diplomacy	Election	Sri-Lanka	Myanmar	Peacekeeping	Diplomacy	Whaling	Royals	Cartoons	Nobel-Prize
cooper	elect	peac	democrat	rebel	meet	whale	children	protest	peac
visit	parti	talk	voic	kill	visit	climat	royal	cluster	award
relat	vote	rebel	genocid	attack	minist	emiss	school	embassi	prize
ambassador	poll	tiger	trial	militari	talk	ban	celebr	prophet	right
develop	coalit	govern	militari	soldier	discuss	global	coupl	cartoon	human
countri	opposit	negoti	websit	tiger	offici	forest	famili	bomb	committe
bilater	parliament	ceasefir	base	armi	prime	chang	visit	newspap	ceremoni
meet	govern	process	prison	war	deleg	fish	citi	ban	winner
discuss	percent	liber	ictr	troop	foreign	hunt	student	demonstr	world
deleg	labour	war	report	monitor	arriv	warm	hospit	attack	nomin
econom	seat	meet	suu	ceasefir	attend	trade	princess	publish	dissid
express	polit	island	daw	wound	met	environ	live	munit	laureat
republ	parliamentari	truce	record	bomb	confer	commerci	ceremoni	public	former
issu	socialist	envoy	wit	fight	leader	import	sonja	condemn	presid
agenc	conserv	broker	former	civilian	invit	research	mother	countri	won
polit	major	wickremesingh	command	truce	report	world	home	treati	campaign
news	voter	parti	arrest	east	jagland	carbon	wife	weapon	win
strengthen	support	agreement	democraci	liber	kjell	quota	church	convent	democraci
tie	rule	polit	crime	island	counterpart	environment	life	caricatur	jail
exchang	won	conflict	via	violenc	trip	meat	father	boycott	war

Table 13: Top Words for POL (Topics 1-10)

Election	Immigration	Election	Military	Business	EU	Economy	History	Crime	Diplomacy
percent	border	po	troop	compani	negoti	budget	camp	polic	visit
poll	visa	polit	countri	gas	summit	percent	anniversari	arrest	meet
elect	countri	pis	war	oil	talk	bank	communist	prison	presid
vote	citizen	parti	terror	contract	treati	rate	peopl	suspect	attend
parti	cross	politician	alli	suppli	enlarg	deficit	church	court	talk
po	travel	newspap	mission	energi	referendum	zloti	death	attack	discuss
platform	immigr	elect	secur	pipelin	bloc	econom	ceremoni	detain	deleg
civic	diplomat	platform	forc	tender	countri	financ	histori	charg	summit
pis	guard	civic	support	invest	join	growth	memori	kill	invit
support	worker	interview	attack	plant	constitut	economi	war	investig	offici
obop	eastern	articl	peac	firm	membership	central	former	prosecutor	arriv
survey	illeg	justic	withdraw	market	vote	polici	leader	sentenc	aleksand
district	custom	comment	allianc	industri	access	cut	victim	crime	minist
justic	region	opinion	told	board	candid	inflat	world	alleg	presidenti
law	join	law	militari	sale	propos	market	mark	extradit	day
candid	peopl	campaign	terrorist	economi	presid	spend	citi	releas	counterpart
sld	resid	candid	nation	purchas	leader	govern	massacr	murder	leader
sept	consul	daili	world	price	meet	reform	commemor	crimin	head
voter	foreign	polici	intern	project	compromis	currenc	celebr	spokesman	held
presidenti	refuge	presid	leader	power	agreement	forecast	histor	incid	ambassador

Table 14: Top Words for POL (Topics 11-20)

Accidents	Media	Constitution	Agriculture	Econ-Coop	Parliament	Budget	Military	Corruption	Military
crash	pap	law	ordin	cooper	parti	fund	militari	investig	missil
plane	agenc	court	amend	relat	elect	tax	soldier	report	defens
kill	news	amend	law	countri	sld	budget	defenc	servic	shield
peopl	report	legal	product	visit	coalit	bill	troop	secret	system
air	sourc	bill	import	meet	vote	pension	forc	radio	defenc
die	minist	regul	ban	bilater	democrat	financ	command	prosecutor	deploy
investig	stress	draft	meat	econom	pis	propos	armi	tv	plan
accid	independ	right	procedur	discuss	parliamentari	money	mission	offic	antimissil
land	prime	properti	method	develop	parliament	programm	arm	present	radar
near	ad	rule	educ	talk	law	percent	conting	committe	base
victim	told	justic	scope	energi	leader	project	oper	former	talk
flight	excerpt	compens	principl	agreement	justic	govern	attack	inform	interceptor
airport	meet	legisl	grant	trade	platform	employ	defens	alleg	negoti
death	presid	resolut	custom	joint	po	labour	multin	commiss	secur
injur	nation	claim	date	counterpart	govern	system	train	intellig	nuclear
pilot	accord	constitut	food	issu	left	cost	base	corrupt	threat
polic	talk	judg	agricultur	tie	civic	plan	kill	minist	instal
wife	confer	propos	trade	minist	opposit	zlbn	unit	public	agreement
flood	situat	appeal	establish	presid	minist	reform	staff	accus	element
tragedi	omit	decis	export	polici	selfdef	invest	equip	head	militari

Table 15: Top Words for EGY (Topics 1-10)

ISR/PSE-Coop	Nuclear	Election	Comments	Terrorism	Diplomacy	Military	Media	Media	Econ-Coop
peac	nuclear	elect	global	attack	visit	war	movement	agenc	cooper
intern	hostag	parti	via	bomb	meet	militari	agreement	news	trade
commentari	energi	opposit	world	kill	talk	weapon	dialogu	meet	countri
radio	compani	vote	democraci	polic	news	region	websit	sourc	econom
effort	oil	candid	women	suicid	discuss	forc	deleg	web	invest
region	gas	presidenti	countri	resort	arriv	resolut	meet	report	develop
situat	water	democrat	speech	explos	agenc	middl	sourc	site	relat
peopl	releas	parliament	polit	wound	deleg	troop	reconcili	musa	agreement
stabil	kidnap	constitut	histori	peopl	attend	countri	report	summit	meet
resolut	project	poll	cultur	tourist	region	alli	tv	discuss	bilater
stress	sept	rule	au	blast	offici	terror	dr	minist	sign
achiev	electr	parliamentari	contin	terrorist	bilater	attack	interview	held	visit
secur	unidentifi	polit	peopl	suspect	develop	rebel	govern	amr	industri
issu	atom	seat	event	church	summit	govern	umar	websit	field
support	tonn	reform	chang	dead	relat	secur	polit	develop	joint
aggress	suppli	amend	nation	terror	counterpart	destruct	discuss	committe	discuss
call	power	ndp	celebr	injur	middl	peac	newspap	situat	tie
communiti	aug	referendum	freedom	bomber	held	arm	issu	foreign	project
countri	natur	presid	middl	car	minist	threat	radio	deleg	ambassador
role	kidnapp	democraci	societi	condemn	tour	warn	sulayman	issu	sector

Table 16: Top Words for EGY (Topics 11-20)

Diplomacy	Refugees	Violence	Gaza	Accidents	Diplomacy	Economy	Human-Rights	Media	Protest
peac	plane	attack	border	kill	talk	percent	arrest	newspap	protest
talk	flight	televis	cross	injur	meet	bank	court	websit	demonstr
middl	refuge	kill	smuggl	accid	ceasefir	econom	prison	mb	polic
meet	airport	assassin	secur	polic	faction	fund	right	articl	squar
process	embassi	milit	tunnel	fire	offici	economi	charg	word	street
negoti	ship	statement	forc	crash	movement	market	sentenc	publish	peopl
settlement	travel	leader	polic	bus	leader	financi	trial	report	regim
visit	aid	terror	fire	road	truce	price	human	dr	thousand
effort	arriv	channel	milit	peopl	deleg	invest	polic	wafd	demand
summit	medic	broadcast	soldier	tourist	mediat	pound	detain	mosqu	upris
mideast	evacu	diplomat	offici	ship	agreement	rate	releas	assembl	clash
discuss	hospit	video	territori	hospit	propos	export	prosecutor	ahram	opposit
plan	return	newspap	troop	rescu	deal	govern	alleg	tv	forc
leader	author	publish	weapon	passeng	dialogu	increas	accus	entitl	power
region	airlin	war	rocket	train	negoti	countri	jail	law	antigovern
confer	humanitarian	claim	town	offici	summit	money	lawyer	page	riot
conflict	citizen	interview	armi	mile	prison	growth	tortur	sermon	rule
envoy	aircraft	cleric	blockad	die	reconcili	trade	investig	daili	call
violenc	home	ambassador	sourc	south	milit	aid	suspect	peopl	ralli
diplomat	oper	sadat	wall	driver	discuss	compani	author	governor	hundr

Table 17: Top Words for ISR (Topics 1-10)

Comments	Missile-Attacks	Terrorism	Diplomacy	Budget	Election	Diplomacy	Nuclear	Media	Diplomacy
world	rocket	attack	visit	govern	elect	peopl	nuclear	websit	meet
war	fire	kill	meet	fund	parti	intern	weapon	report	talk
middl	kill	bomb	middl	aid	vote	support	missil	news	offici
polit	milit	suicid	talk	budget	poll	right	program	sourc	leader
polici	attack	milit	discuss	minist	minist	nation	countri	agenc	ceasefir
countri	armi	leader	peac	econom	coalit	peac	sanction	wafa	agreement
antisemit	wound	terror	countri	money	govern	countri	atom	pna	faction
terror	southern	peopl	minist	cabinet	prime	regim	threat	occup	truce
conflict	shell	violenc	confer	financi	candid	region	develop	tv	discuss
speech	militari	terrorist	region	approv	parliament	world	militari	web	negoti
word	missil	bomber	diplomat	committe	leader	resist	energi	dr	agre
leader	town	strike	pere	bill	labor	aggress	programm	headlin	movement
newspap	northern	assassin	foreign	law	polit	occup	system	movement	summit
histori	citi	target	ambassador	bank	seat	stress	compani	site	senior
democraci	strike	civilian	relat	tax	percent	resolut	enrich	voic	govern
critic	air	condemn	deleg	economi	labour	confer	technolog	resist	uniti
power	tank	dead	leader	vote	parliamentari	polit	defens	statement	minist
articl	target	citi	trip	donor	resign	communiti	uranium	peopl	dialogu
view	mortar	war	arriv	financ	livni	human	arm	inform	propos
chang	raid	wound	cooper	cut	won	achiev	bomb	committe	cabinet

Table 18: Top Words for ISR (Topics 11-20)

Judiciary	Society	Gaza	ISR/PSE-Coop	ISR/PSE-Conf	Terrorism	ISR/PSE-Coop	Refugees	Settlements	Violence
court	hospit	border	peac	prison	report	resolut	cross	settlement	kill
investig	famili	ship	talk	releas	radio	call	church	settler	wound
polic	children	forc	negoti	soldier	defenc	ceasefir	border	protest	fire
alleg	school	troop	settlement	arrest	sourc	intern	water	polic	soldier
charg	student	aid	middl	milit	newspap	secur	food	demonstr	armi
right	doctor	flotilla	process	kidnap	defens	violenc	suppli	barrier	shot
human	life	plane	map	brigad	correspond	condemn	refuge	evacu	sourc
arrest	celebr	blockad	meet	jail	secur	nation	fuel	build	polic
crime	ceremoni	cross	plan	secur	minist	situat	humanitarian	citi	troop
suspect	live	raid	road	movement	terrorist	statement	worker	construct	attack
law	death	southern	quartet	captur	post	urg	allow	plan	citi
justic	women	deploy	effort	abduct	oper	action	aid	land	explos
accus	memori	armi	roadmap	arm	voic	immedi	relief	mosqu	car
sentenc	immigr	south	confer	martyr	plan	middl	peopl	wall	injur
indict	die	smuggl	solut	aqsa	attack	attack	unrwa	hous	bomb
trial	film	militari	agreement	leader	armi	civilian	blockad	fenc	near
former	father	weapon	freez	exchang	command	conflict	thousand	outpost	dead
legal	wife	peacekeep	envoy	citi	daili	humanitarian	travel	home	hospit
lawyer	educ	air	hope	forc	terror	forc	tourist	site	camp
judg	mother	activist	summit	gunmen	forc	escal	closur	thousand	town

Table 19: Top Words for JOR (Topics 1-10)

Diplomacy	Humanitarian-Aid	Terrorism	Comments	Diplomacy	Violence	Election	Military	ISR/PSE-Conf	Royals
meet	forc	attack	peac	visit	kill	elect	war	violenc	minist
visit	militari	bomb	agenc	agenc	polic	vote	region	peac	reform
talk	train	kill	effort	cooper	fire	law	middl	ceasefir	govern
summit	aid	suicid	news	news	attack	parliament	world	resolut	prime
discuss	troop	hotel	region	bilater	wound	parti	alli	withdraw	cabinet
offici	hospit	terror	websit	discuss	rocket	parliamentari	militari	call	appoint
minist	plane	terrorist	stress	relat	soldier	poll	weapon	middl	polit
leader	medic	peopl	process	meet	car	polit	countri	talk	resign
middl	humanitarian	alzarqawi	meet	talk	injur	candid	regim	stop	former
arriv	air	bomber	web	develop	citi	opposit	leader	escal	king
trip	flight	condemn	achiev	web	border	seat	polit	leader	reshuffl
tour	arm	target	support	websit	forc	govern	warn	forc	econom
expect	command	musab	stabil	countri	near	legisl	power	secur	replac
attend	suppli	claim	report	field	shot	front	analyst	meet	form
foreign	food	milit	import	sourc	hospit	women	polici	envoy	royal
hold	staff	dead	site	site	incid	committe	administr	militari	ministri
region	convoy	act	develop	deleg	town	deputi	conflict	region	chang
held	agenc	innoc	affirm	econom	port	voter	terror	pere	educ
peac	oper	blast	solut	tie	bomb	council	support	attack	offici
met	arriv	wound	majesti	report	sourc	democrat	fear	urg	corrupt

Table 20: Top Words for JOR (Topics 11-20)

Comments	ISR/PSE-Coop	Judiciary	Hostages	Refugees	Police	Media	Nuclear	Protest	Society
countri	peac	court	releas	refuge	arrest	newspap	agreement	protest	famili
peopl	talk	charg	hostag	border	intellig	sourc	trade	demonstr	children
confer	middl	sentenc	prison	camp	suspect	report	oil	right	father
nation	negoti	prison	kidnap	countri	offici	publish	project	human	daughter
polit	process	trial	unidentifi	return	secur	alyawm	econom	call	women
intern	settlement	lawyer	sept	home	attack	dustur	nuclear	mosqu	death
world	plan	defend	driver	cross	weapon	dr	invest	opposit	wife
support	meet	plot	aug	agenc	alleg	govern	sign	law	die
region	conflict	prosecutor	kill	travel	terrorist	issu	energi	activist	sister
summit	summit	convict	abduct	war	author	inform	compani	freedom	kill
terror	map	attack	compani	visa	investig	entitl	export	ralli	home
stabil	initi	death	freed	fled	smuggl	web	countri	govern	citi
uniti	road	militari	kidnapp	unhcr	inform	movement	water	violat	brother
issu	propos	jail	held	live	polic	tv	bank	peopl	celebr
dialogu	effort	accus	truck	land	plot	ra	industri	condemn	honour
speech	solut	suspect	milit	entri	oper	media	fund	demand	live
call	region	alleg	detaine	resid	accus	polit	economi	parti	princ
govern	palestinianisra	guilti	demand	embassi	agent	interview	sector	associ	student
secur	roadmap	judg	video	allow	detain	excerpt	market	march	ceremoni
role	withdraw	terrorist	jordanian	displac	explos	articl	price	street	raghad

Table 21: Top Words for ALB (Topics 1-10)

Diplomacy	Diplomacy	Development	Diplomacy	Violence	Election	Rebellion	Corruption	Meeting	Smuggling
agenc	membership	agreement	independ	polic	elect	ethnic	crime	meet	polic
news	countri	cooper	status	kill	vote	rebel	fight	minist	border
cooper	summit	sign	talk	explos	socialist	guerrilla	organ	foreign	traffick
relat	join	project	negoti	injur	opposit	peac	law	visa	arrest
meet	reform	educ	solut	peopl	democrat	forc	corrupt	repres	drug
region	invit	countri	resolut	hospit	parti	villag	agenc	agreement	illeg
countri	allianc	trade	intern	capit	poll	troop	news	sign	agenc
integr	stabil	develop	secur	blast	protest	weapon	justic	businessmen	cross
develop	sign	econom	envoy	villag	result	fight	institut	particip	news
visit	agreement	transport	issu	near	won	war	report	cooper	oper
stabil	integr	cultur	plan	car	seat	arm	govern	rotat	sourc
express	meet	energi	diplomat	fire	coalit	ceasefir	reform	procedur	smuggl
bilater	enlarg	news	communiti	mile	victori	border	terror	undersecretari	suspect
moisiu	access	agenc	futur	dead	ballot	govern	traffick	presid	immigr
report	region	region	meet	death	candid	armi	system	simplifi	seiz
support	support	economi	region	town	parliamentari	minor	human	driver	report
process	effort	infrastructur	propos	kilomet	parliament	secur	public	truck	peopl
sourc	hope	minist	provinc	victim	govern	violenc	legal	sent	vlore
polit	progress	school	support	die	leader	polic	measur	capit	crime
econom	step	programm	polit	accid	elector	insurg	inform	expect	custom

Table 22: Top Words for ALB (Topics 11-20)

Election	Meeting	Rebellion	Peacekeeping	Coalit-Govt	War-Crimes	Violence	Parliament	Media	Economy
elect	meet	provinc	defenc	minist	crime	polic	assembl	newspap	fund
parti	visit	ethnic	militari	prime	prosecutor	protest	news	polit	bank
democrat	minist	independ	forc	govern	investig	tv	agenc	democrat	econom
stori	prime	nation	arm	resign	war	attack	opposit	parti	invest
elector	talk	war	troop	deputi	court	incid	socialist	publish	percent
polit	luncheon	major	mission	post	prison	report	parti	report	project
opposit	counterpart	leader	peacekeep	ilir	former	websit	elect	koha	govern
vote	ceremoni	talk	command	cabinet	arrest	violenc	democrat	independ	power
local	host	bomb	armi	appoint	alleg	sourc	pd	privatelyown	budget
parliamentari	inaugur	forc	defens	former	charg	news	vote	leader	energi
main	confer	futur	oper	presid	accus	agenc	parliamentari	yesterday	busi
moisiu	hold	status	train	newspap	tribun	condemn	chairman	excerpt	economi
reform	particip	minor	staff	ambassador	offic	act	report	chairman	compani
parliament	recep	run	soldier	fato	suspect	northern	sourc	unattribut	financi
verifi	highway	percent	secur	foreign	sentenc	municip	deputi	peopl	growth
voter	chairman	former	conting	socialist	trial	kfor	polit	interview	countri
accuraci	travel	remain	news	diplomat	prosecut	radio	parliament	passag	tax
socialist	mayor	separatist	weapon	offic	file	terrorist	commiss	omit	electr
process	expect	breakaway	unit	decre	crimin	town	session	entitl	financ
vouch	joint	mission	exercis	nomin	justic	eulex	meet	ditor	money

Table 23: Top Words for ARM (Topics 1-10)

Protest	Crime	Genocide	ARM/AZB	Peacekeeping	Parliament	Energy	ARM/AZB	ARM/AZB	Diplomacy
polic	crime	genocid	conflict	militari	session	energi	conflict	border	visit
protest	polic	resolut	territori	forc	assembl	gas	settlement	region	meet
peopl	investig	kill	settlement	arm	parliament	nuclear	meet	protocol	discuss
ralli	crash	empir	region	villag	commiss	newspap	negoti	disput	minist
kill	murder	vote	peac	defens	deleg	transport	talk	diplomat	cooper
prison	record	bill	intern	region	sit	project	presid	relat	deleg
opposit	crimin	recogn	integr	fire	secur	power	cochairmen	tie	offici
arrest	court	recognit	statement	troop	discuss	pipelin	cochair	war	arriv
attack	offic	massacr	occupi	defenc	speaker	oil	peac	sign	presid
soldier	drug	call	resolut	posit	price	secur	process	enclav	defenc
citi	plane	parliament	militari	exercis	deputi	railway	minist	talk	issu
demonstr	traffick	committe	situat	armi	council	compani	discuss	ethnic	met
squar	dink	adopt	settl	monitor	particip	construct	mediat	conflict	agenc
clash	kill	victim	polici	ceasefir	parliamentari	militari	chair	normal	foreign
offic	terrorist	centuri	forc	district	committe	suppli	principl	territori	bilater
wound	peopl	relat	land	report	chairman	weapon	agreement	forc	servic
violenc	prosecutor	approv	war	command	issu	border	resolut	countri	counterpart
author	illeg	deni	independ	casualti	held	plant	settl	establish	press
capit	accid	condemn	posit	soldier	meet	region	posit	control	talk
bomb	flight	crime	occup	peacekeep	head	entitl	propos	kill	secur

Table 24: Top Words for ARM (Topics 11-20)

Media	Ceremony	Constitution	Parliament	Econ-Coop	Parliament	Election	Genocide	Development	Diplomacy
agenc	ambassador	polit	parti	cooper	law	elect	genocid	program	pace
news	relat	opposit	elect	relat	amend	presidenti	letter	educ	meet
report	church	constitut	opposit	develop	draft	vote	call	mln	discuss
sourc	express	reform	leader	econom	govern	poll	alleg	fund	deleg
tv	develop	author	candid	bilater	sign	observ	organ	amd	conflict
excerpt	visit	parti	polit	trade	legisl	candid	media	govern	repres
present	diplomat	power	presidenti	countri	legal	opposit	statement	scienc	assembl
defenc	peopl	referendum	parliamentari	meet	approv	parliamentari	histori	budget	issu
correspond	cultur	chang	republican	region	adopt	elector	confer	economi	organ
an	messag	state	parliament	agreement	servic	voter	cultur	implement	visit
privat	congratul	democrat	bloc	sphere	agreement	central	book	school	committe
station	anniversari	situat	coalit	import	tax	campaign	event	develop	region
omit	ceremoni	coalit	orinat	discuss	citizen	result	histor	system	cooper
passag	contribut	accord	democrat	mutual	decre	round	claim	project	settlement
presid	presid	societi	post	polit	right	violat	issu	social	special
programm	polit	forc	uniti	ambassador	read	station	sent	alloc	session
minist	award	countri	resign	level	regul	ter	monument	assist	polit
snark	friend	peopl	faction	note	custom	ballot	regard	financi	resolut
republ	countri	process	communist	economi	accord	mission	historian	sector	report
newspap	establish	parliamentari	appoint	invest	bill	percent	studi	financ	right

Table 25: Top Words for POR (Topics 1-10)

Judiciary	Violence	Military	Shipping	IOs	Budget	Mil-Coop	Diplomacy	Accidents	Culture
court	former	forc	ship	via	budget	summit	visit	fire	world
salem	coloni	militari	oil	global	deficit	meet	cooper	firefight	peopl
prison	independ	defenc	coast	communiti	percent	leader	meet	peopl	newspap
alleg	rebel	troop	water	speak	financ	talk	bilater	kill	articl
extradit	militari	mission	island	sport	cut	secur	relat	forest	histori
investig	war	noticia	fish	countri	debt	war	discuss	plane	life
charg	coup	secur	boat	santo	govern	diplomat	countri	injur	live
trial	nation	newspap	mile	angop	econom	missil	talk	mile	church
accus	tome	websit	tonn	ambassador	growth	discuss	minist	blaze	book
sentenc	govern	arm	vessel	cape	financi	attend	offici	passeng	death
arrest	armi	soldier	earthquak	particip	economi	presid	deleg	accid	time
abus	kill	report	tanker	deleg	bailout	issu	agreement	near	polit
former	forc	oper	port	meet	measur	minist	presid	die	war
suspect	leader	excerpt	fuel	tome	fund	resolut	trade	town	centuri
jail	peac	command	prestig	languag	auster	peac	counterpart	bus	noticia
judg	violenc	air	storm	organis	tax	plan	summit	land	campaign
crime	troop	polic	navi	attend	crisi	foreign	develop	hospit	site
deport	arm	guard	crew	diplomat	rate	defenc	econom	car	word
children	soldier	conting	km	play	spend	invit	sign	northern	publish
cbi	elect	deploy	quak	accord	zone	nuclear	socrat	road	cultur

Table 26: Top Words for POR (Topics 11-20)

Crime	Immigration	Election	Diplomacy	Drugs	Violence	Development	Coalit-Govt	Media	Constitution
polic	immigr	elect	countri	polic	attack	compani	minist	report	treati
disappear	law	socialist	presid	drug	terror	invest	presid	agenc	referendum
parent	propos	poll	polit	arrest	protest	energi	prime	lusa	vote
girl	illeg	vote	relat	seiz	polic	market	resign	news	constitut
miss	countri	parti	econom	suspect	terrorist	busi	elect	radio	ratif
investig	trade	percent	develop	cocain	fan	economi	lope	sourc	ratifi
suspect	legal	democrat	support	traffick	demonstr	industri	leader	web	parliament
daughter	worker	opposit	membership	investig	footbal	project	socialist	excerpt	reject
famili	measur	seat	express	oper	secur	govern	parti	site	approv
resort	visa	voter	futur	author	bomb	bank	govern	present	reform
holiday	bloc	socrat	stress	explos	strike	educ	abort	tv	voter
search	import	parliament	cooper	found	peopl	agreement	parliament	noticia	decis
coupl	market	win	posit	detain	fight	export	polit	passag	bloc
da	labour	won	enlarg	weapon	match	percent	appoint	omit	sign
mccann	border	candid	polici	uranium	kill	oil	former	newspap	opposit
apart	allow	major	peac	bank	organ	econom	guterr	rtp	presid
hotel	agreement	govern	process	pj	threat	electr	execut	websit	parti
evid	nation	leader	role	crime	march	stori	replac	minist	propos
vanish	employ	campaign	common	illeg	team	sector	post	quot	hold
child	servic	coalit	integr	car	violenc	tourism	democrat	broadcast	rule

References

Edward E. Azar. The conflict and peace data bank (COPDAB) project. Journal of Conflict Resolution, 24:143–152, 1980.

David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003. http://www.cs.princeton.edu/∼blei/papers/BleiNgJordan2003.pdf.

Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles L. Taylor. Integrated data for events analysis (IDEA): An event typology for automated events data development. Journal of Peace Research, 40(6):733–745, 2003.

J. Chang, J. Boyd-Graber, and D. Blei. Connections between the lines: Augmenting social networks with text. In Knowledge Discovery and Data Mining, 2009.

Jonathan. Chang. Package lda: Collapsed Gibbs sampling methods for topic models. http://cran.r-project.org/web/packages/lda/; Version date: 24-October-2010, 2010.

Jonathan. Chang. Package lda: Collapsed Gibbs sampling methods for topic models. http://cran.r-project.org/web/packages/lda/; Version date: 3-Nov-2011, 2011.

Ingo Feinerer. **openNLP**: **OpenNLP** interface. R package version 0.1. http://cran.r-project.org/web/packages/openNLP/index.html, 2007a.

Ingo Feinerer. **tm**: Text mining package. R package version 0.3. http://cran.r-project.org/web/packages/tm/index.html, 2007b.

Ingo Feinerer and Kurt Hornik. Introduction to the **openNLP** package. http://opennlp.sourceforge.net/README.html, 2010.

Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in R. Journal of Statistical Software, 25, 2008.

Deborah J. Gerner, Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. The machine coding of events from regional and international sources. International Studies Quarterly, 38:91–119, 1994.

Deborah J. Gerner, Philip A. Schrodt, and Ömür Yilmaz. *Conflict and Mediation Event Observations (CAMEO) Codebook*. http://eventdata.psu.edu/data.dir/cameo.html, 2009.

Joshua S. Goldstein. A conflict-cooperation scale for WEIS events data. Journal of Conflict Resolution, 36:369–385, 1992.

Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. International Organization, 57(3):617–642, 2004.

Russell J Leng. Behavioral Correlates of War, 1816-1975. (ICPSR 8606). Inter-University Consortium for Political and Social Research, Ann Arbor, 1987.

Charles A. McClelland. World Event/Interaction Survey Codebook (ICPSR 5211). Inter-University Consortium for Political and Social Research, Ann Arbor, 1976.

Charles A. McClelland. Let the user beware. International Studies Quarterly, 27(2):169–177, 1983.

Patrick McGowan, Harvey Starr, Gretchen Hower, Richard L. Merrit, and Dina A. Zinnes. International data as a national resource. International Interactions, 14:101–113, 1988.

Richard L. Merritt, Robert G. Muncaster, and Dina A. Zinnes, editors. International Event Data Developments: DDIR Phase II. University of Michigan Press, Ann Arbor, 1993.

Slava Mikhaylov, Michael Laver, and Kenneth Benoit. Coder reliability and misclassification in the human coding of party manifestos. Political Analysis, 20(1):78–91, 2012.

Philip A. Schrodt. Forecasting conflict in the Balkans using hidden Markov models. In Robert Trappl, editor, Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention, pages 161–184. Kluwer Academic Publishers, Dordrecht, Netherlands, 2006.

Philip A. Schrodt. TABARI: Textual Analysis By Augmented Replacement Instructions, 2011. `http://eventdata.psu.edu/tabari.html`.

Philip A. Schrodt and Deborah J. Gerner. Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. American Journal of Political Science, 38:825–854, 1994.

Philip A. Schrodt, Glenn Palmer, and Mehmet Emre Hatipoglu. Automated detection of reports of militarized interstate disputes using the SVM document classification algorithm. Paper presented at American Political Science Association, 2008.