An Event Data Set for the Arabian/Persian Gulf Region 1979-1997

Philip A. Schrodt and Deborah J. Gerner
Department of Political Science
University of Kansas
Lawrence, KS 66045 USA
phone: 785-864-3523 fax: 785-864-5700
p-schrodt@ukans.edu d-gerner@ukans.edu

Paper presented at the International Studies Association
Minneapolis, 18 - 22 March 1998

The Gulf data sets, KEDS program and other information
are available at the KEDS web site:
`http://raven.ukans.cc.edu/~keds`

This research was funded in part by the
University of Kansas General Research Fund
Grant 3500-X0-0038.

Abstract

This paper discusses a WEIS-coded event data set covering the Arabian/Persian Gulf region (Iran, Iraq, Kuwait, Oman, Saudi Arabia, Yemen, and the smaller Gulf states) for the period 15 April 1979 to 10 June 1997. The coded events cover international interactions among these states, as well as interactions with any other states or major international organizations. The data set is generated from Reuters news reports downloaded from the NEXIS data service and coded using the Kansas Event Data System (KEDS) machine-coding program.

The paper begins with a review of the process of generating a machine-coded data set, including a discussion of software we have developed to partially automate the development of dictionaries to code new geographical regions. The Gulf data are coded using a standard set of verb phrases (rather than phrases specifically adapted to the Gulf) and an actors dictionary that has been augmented only with the actors identified by a utility program that examines the source texts for actors not already found in the KEDS dictionary.

The Reuters reports generate 264,421 events when full stories are coded and 48,721 events when only lead sentences are coded. An examination of the time series that are generated when the events are aggregated by month using the Goldstein scale shows that they capture the major features of the behavior that we know to have occurred in the region. There is generally a high correlation ($r > 0.75$) between the series generated from lead-sentences and from full stories when the major actors of the region (Iran, Iraq, Saudi Arabia and USA) are studied. An exception to this pattern is found in interactions involving a relatively minor actor, the United Arab Emirates. Here the full-story coding provides far more events than the lead-sentence coding and shows greater variance even for interactions between major actors. We expect this will also be the case for other small Gulf states, suggesting that full-story coding may be necessary for a complete analysis of these actors.

This is a very boring paper. It reports on the development and basic validation of a WEIS-coded event data set covering the Arabian/Persian Gulf region (Iran, Iraq, Kuwait, Oman, Saudi Arabia, Yemen, and the smaller Gulf states) for the period 15 April 1979 to 10 June 1997. We generated the data set from Reuters news reports downloaded from the NEXIS data service and coded using the Kansas Event Data System (KEDS) machine-coding program.

We first discuss the process of generating a machine-coded data set, with some additional observations on parts of this process that we have partially automated. We then assess the face validity of these data by examining some time-series charts generated from these data using the Goldstein (1992) scale. We compare the data sets generated from coding all of the sentences in a story with the data generated from the lead sentences only, and conclude with some general comments about the future of machine coding.

The data sets—which are presumably the main reason you have any interest in this paper—can be downloaded from the KEDS web site `http://raven.ukans.cc.edu/~keds`. All of the utility programs discussed in this paper can also be downloaded from the web site.

Machine Coding of Event Data

The most visible change that has occurred in event data research during the 1990s has been the widespread shift to machine or machine-assisted coding. Event data collection projects that once would have required ten of thousands of dollars, a flock of student coders with a complex supervisory infrastructure and months of painstaking efforts can now be done by a single researcher in a few weeks when pre-existing dictionaries are used.[1] This development has, in turn, led to a series of less-obvious spin-offs such as the ability to experiment easily with new event coding systems such as PANDA (Bond, Bennett, & Vogele 1994, Bond et al. 1997) and with the coding of internal actors (Goldstein & Pevehouse 1997, Huxtable 1997;).

---

[1] This is not hypothetical: the Bosnia data set used in Goldstein & Pevehouse (1997) was coded under just these conditions.

The Kansas Event Data System (KEDS) project is one such machine-coding effort. Details on this work can be found in Gerner et al (1994), Schrodt & Gerner (1994), Schrodt, Davis & Weddle (1994), and Huxtable & Pevehouse (1996), as well as at the KEDS project web site. Almost of the initial work in the KEDS project involved studies of the Arab-Israeli conflict, although recently the program has been used by Huxtable to code West African events and Pevehouse to code the contemporary conflict in the Balkans.

Machine coding uses many of the techniques developed in other fields of computerized natural language processing (Salton 1989). The task of machine coding of event data is simplified by the fact that event data are defined by sets of transitive verbs and the coding needs only to focus on the basic subject-verb-object (SVO) structure of an English sentence. Consequently, the problem of event data coding is substantially easier than many other natural language processing projects. In event coding, the subject of the sentence is the *source* of the event, the verb determines the *event code*, and the object of the verb is the *target*.[2] Thus, consistent with similar work in linguistic pattern recognition for the purpose of coding political events (Lehnert and Sundheim 1991), KEDS can use *sparse parsing* of sentences, focusing primarily on the SVO sentence structure and word ordering rather than using full syntactical analysis. KEDS makes errors on oddly constructed or excessively complex sentences but is quite successful on the sentence structures most commonly used by journalists to describe events.

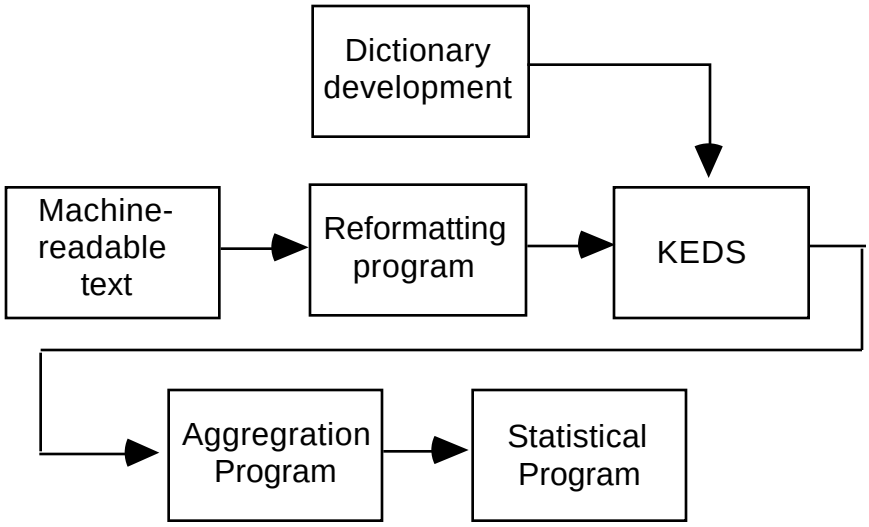
The ideal system for generating machine-coded event data would take a news source as the input and produce an appropriately formatted event data set as the output. While such a coding system is technically feasible with the integration of existing computer programs, in practice coding an event data set requires a few more steps than, for instance, pulling variables off the Euro-Barometer surveys. Figure 1 outlines the process involved in going from machine-readable

---

[2] Many discussions of event data use the word *actor* to refer only to the subject/source. In our work, "actor" refers to the set of entities that can be sources or targets.

text to data that can be analyzed with a statistical program. The remainder of this section will discuss the practical considerations at each step in that process.

Figure 1. Generating machine-coded data



STEP 1: Locate and reformat a set of machine-readable texts

The first step in machine coding is finding a source of machine-readable text. In our project, we have used Reuters reports from the NEXIS data service. Originally the text looks like

```
] ] ]
]                          LEVEL 1 - 5 OF 914 STORIES
] ]            Proprietary to the United Press International 1981
] ]                     <December> 29, 1981, Tuesday, PM cycle
] ]HEADLINE: World News Summary
] ] BODY:
]   Sen. Charles Percy, R-Ill., chairman of the Senate Foreign
]Relations Committee, asked his Israeli hosts to avoid the annexed
]
^
^
]Golan Heights in a helicopter inspection tour today of the tense
]Israeli-<Lebanese>frontier.
]
```

It needs to be converted to the KEDS input format:

```
811229UPI001
Sen. Charles Percy, R-Ill., chairman of the Senate Foreign
Relations Committee, asked his Israeli hosts to avoid the annexed
Golan Heights in a helicopter inspection tour today of the tense
Israeli-Lebanese frontier.
```

We have developed a fairly elaborate reformatting program—written in C—to remove all of the irrelevant information found in the NEXIS download and reformat the text. (The PANDA project used the macro language in Microsoft's *Word* program to perform the same task.) This program can filter the text to find only the lead sentence; it can also produce individual records for each sentence in the story. In full-story coding the program has the option of skipping sentences that appear to be direct quotes. The filter program also converts the English date given in Reuters to the YYMMDD numerical data format used by KEDS and assigns a unique identification number to the lead. Finally, the program checks each day for multiple sentences that have almost identical letter counts; this eliminates most of the stories rebroadcast by Reuters to correct spelling errors.

STEP 2: Develop the coding dictionaries

KEDS uses large dictionaries of proper nouns (*.actors*) and verb phrases (*.verbs*) to code the actors and events it finds in the source text. The dictionaries originally developed by the KEDS and PANDA projects are each the result of about two person-years of working through Reuters lead sentences to identify relevant verb phrases and assign these to an appropriate event code. These dictionaries contain most of the basic vocabulary used to describe political events. As we will discuss in more detail below, we have combined the KEDS and PANDA project dictionaries and an assortment of other *.verbs* dictionaries into a "standard" *.verbs* file that appears to capture most of the *international* events relevant to the WEIS coding scheme. We have used this standard dictionary to code the events analyzed in this paper.

Dictionaries for the coding of *internal* political events present additional challenges and it is not clear whether it will be possible to develop a general-purpose *.verbs* dictionary for this task.

Over the past year, we have developed a number of *.verbs* dictionaries to code internal events in a disparate set of states: Russia, China, Albania, Colombia, Mexico, Syria, Algeria, Pakistan and Nigeria.

Based on our experiences thus far, we suspect that the vocabulary referring to domestic events may vary too greatly to generalize across regions. First, internal events involve a much larger set of verb phrases than international events and many of these are idiosyncratic to specific states. For example, reports on Colombia and Mexico reflect a great deal of quasi-political criminal activity involving the trade in illegal drugs that is virtually nonexistent in the Middle East, Africa or Europe. Africa, in contrast, involves some quasi-political criminal activity involving the smuggling of diamonds that is not found elsewhere in the world. When we coded Albania for the 1996-1997 period, we encountered a series of events involving the collapse of pyramid investment schemes, followed by an almost complete breakdown of political order, followed by an international intervention that very quickly restored order: This sequence was quite distinct from the civil disorder we coded for Lebanon.

Second, despite the generally consistent style found in Reuters reports, there are certain distinct phrases that are employed by the reporters and editors in various regions. Each set of regional reports has a few idiosyncratic turns of phrase that we have not encountered earlier.[3] Because these phrases are common, they are discovered very quickly when spot-checking the dictionaries and actually simplify coding when events are reported using a small set of routine sentence structures.

Our conclusion from these projects is that anyone wanting to code internal events should spend time customizing the standard dictionaries. If a sequence of very unusual events has occurred—the collapse of the Albanian financial system, for instance—it may be advisable to develop dictionaries specifically to code that period. Routine internal behavior, on the other

---

[3]When developing a dictionary, there are times when one has a sense of looking over the collective shoulder of Reuters: For example an indicator that Reuters is using inexperienced reporters (or over-worked editors) in covering a local crisis are reports containing two consecutive apostrophes (' ') instead of a quotation mark (").

hand, can probably be coded reliably using standard dictionaries with some spot-checking for distinct phrasing and regionally-specific forms of political activity.

STEP 3: Supplement the *.actors* dictionaries

While the verb phrases used to describe political events differ little across geographical regions, there is substantial variation in the political actors, particularly once one moves to full-story coding. Consequently, any project focusing on a new geographical region needs to supplement the *.actors* file. When internal events are being coded, these modifications can be fairly extensive.

Over the past year, we have partially automated the process of identifying new actors by using a program called Actor_Filter. This program goes through a set of text records and, based on patterns of consecutive capitalized words, tabulates phrases that may refer to new political actors. The output of this program is a "keyword-in-context" (KWIC) file of the actors that cannot be found in an existing KEDS dictionary; these new actors are listed in order of frequency.

Figure 2 shows an example of the first records of this file from an Actor_Filter index on events dealing with Algeria. In the KWIC format, the actors are highlighted with <<...>>. The first two sets of records in Figure 2 show two common actors that were not already in the dictionary—the Islamic Salvation Front (209 occurrences) and President Liamine Zeroual (182 occurrences)—and also highlights an assortment of other proper nouns such as Air Algerie and Ali Belhadj.[4] The third set of records identifies an actor that is apparently very common in the texts—the *El Watan* newspaper, which Reuters frequently uses as a source—but which was not coded.

---

[4] These records also show some problems with the consecutive-capitalization rule, particularly when dealing with languages other than English. For example, Reuters does not capitalize the Arabic article "al-", so "Hafez al-Assad" is not seen as a single phrase. The French "L'Authentique" fails the test because of the contraction; "Liberte" because it is only one word. Obviously more sophisticated rules could be developed to deal with these cases, at the expense of a greater number of false positives. One still has to deal with idiosyncrasies such as the transliteration "El Watan" ("The Nation") rather than "al-Watan." (This is presumably due the transliteration employed on the masthead of the paper itself, and that in turn was probably chosen by some Spanish typesetter stranded in Algiers in 1923...)

Reuters will refer to a actor using a variety of different phrases. For example, Algerian President Liamine Zeroual might be referred to using any of the following formulation:

```
Algerian President Liamine Zeroual
Algerian President Zeroual
President Zeroual
Liamine Zeroual
Zeroual
```

As a consequence most major individual politicians such as chief executives require multiple entries. These are almost invariably subsets of the general structure

```
<nation name> <title> <first name> <last name>
```

and we anticipate that we could use this regularity to further automate the dictionary development process.

The Reuters editorial style appears to specify that whenever a relatively unknown actor is introduced in a story, he or she is identified by nationality. This characteristic makes the coding of international events relatively easy, because one knows the nationality of an actor even if the individual actors is not in the dictionary. It is less helpful in the coding of internal events, where the actor's national identity can generally be assumed from the context of the story. For example in the lead:

> Moslem guerrillas killed 14 people overnight in Tabainat village in Blida province, 50 km (30 miles) south of Algiers, Algerian security forces said on Monday.

both the guerrillas and the people killed are Algerian, but this is not stated explicitly.[5] The KEDS program contains some specialized routines (developed for the PANDA project) that identify geographical location, but we are not using them at the present time.

---

[5] In Gerner et al (1994) we describe a similar problem found when we tried to code chronologies from the *Journal of Palestine Studies*. *JPS* assumed certain nationalities were known, so that the statement "Israeli police beat Palestinian demonstrators" was always rendered as simply "Police beat demonstrators."

Figure 2.

Actor_Filter KWIC File Output

```
Islamic Salvation Front  (209)

960605  REUT-0004-04
The banned <<Islamic Salvation Front>> (FIS), which has been battling to topple
the Algerian government since 1992, has increased contacts with Libyan
authorities seeking information about the missing men, al-Hayat said

960611  REUT-0002-01
Lawyers defending <<Ali Belhadj >> a leader of the banned <<Islamic Salvation
Front>> (FIS), said on Tuesday that Algeria had arbitrarily detained their
client for the past year in a secret location

960612  REUT-0007-07
France has consistently refused to have any official contact with the <<Islamic
Salvation Front>> (FIS), deprived of almost certain victory in 1992 general
elections when the army intervened to cancel the second round of voting

==========================================

President Liamine Zeroual  (182)

960606  REUT-0003-22
<<President Liamine Zeroual>> has started a dialogue with political party
leaders and other prominent figures, excluding radical Islamists, in his quest
for a peaceful settlement after winning the presidential elections in 1995

960610  REUT-0001-02
It said <<President Liamine Zeroual>> accepted the resignation last week of
Mohamed Benchercheli appointed in 1994

960611  REUT-0002-03
They said they had appealed to the authorities several times, writing to
Algeria's justice minister and to <<President Liamine Zeroual>> among others
but to no avail

==========================================

El Watan  (137)

960609  REUT-0004-03
The newspapers, including the best known dailies <<El Watan >> al-Khabar,
L'Authentique and Liberte, did not appear on Sunday for the seventh day

960610  REUT-0001-01
The director general of <<Air Algerie>> resigned from the troubled state-run
airline, the Algerian newspaper <<El Watan>> reported on Monday

960610  REUT-0001-08
<<El Watan>> said there was also a sharp drop in revenue when <<Air Algerie>>
suspended direct flights to Paris last July after French aviation authorities
insisted it switch from Paris Orly airport to Roissy
```

STEP 4: Autocode the entire data set

Once the dictionaries have been suitably refined, the data should be recoded in fully automatic mode to ensure that the coding rules are consistently applied across the entire data set. If only part of the data set is machine-coded—with occasional records manually "corrected"—inconsistencies will be introduced into the time series that might show up as statistical artifacts later in the analysis. Autocoding also insures that the coding can be replicated by later researchers and can be updated.

STEP 5: Aggregate the data for statistical analysis

KEDS produces standard event data of the form

```
date <tab> source code <tab> target code <tab> event code
```

For example:

```
960101  IRN  IRQ  032
960101  IRQ  IRN  033
960101  ISR  LEB  220
960101  ISR  PAL  094
960102  JOR  ISR  081
960102  ISR  JOR  081
960102  ISR  JOR  032
960102  JOR  ISR  033
960102  ISR  LEB  220
```

For many applications, this nominal-level (categorical) event series must be aggregated before it can be used by standard statistical programs (such as SPSS and SAS) or graphical displays such as spreadsheets that expect an interval-level (numerical) series. This transformation is usually done by mapping each event code to an interval-level scale (for example, Goldstein 1992), and then aggregating the data by actor-pair and week, month, or year using averages or totals. It is *possible* to do this aggregation by scripting the data transformation facilities of a statistical program. However, this process tends to be very slow and awkward, particularly when dealing with a large number of actor pairs. We have usually employed a customized aggregation program, KEDS_Count, to automate this process.

Standard Coding

In our earlier work on the Arab-Israeli conflict, we refined the *.actors* and *.verbs* dictionaries by going through the source texts a number of times. This is quite labor-intensive and required many hours of human-coding over a number of years.

For this paper, we are experimenting with an alternative approach involving far less work: We use a "standard" *.verbs* dictionary and supplement the *.actors* dictionary using the Actor_Filter KWIC index. Using this method, a new set of dictionaries can be developed in a matter of hours. The objective of this exercise is to determine the overall quality of the resulting data. While a data set produced with "tweaked" dictionaries would be of higher quality, the effort spent on improving the dictionaries might be better spent refining the statistical models.

The standard verbs dictionary is a composite of several different dictionaries that were developed in conjunction with the KEDS project over the past eight years, including our original Levant dictionaries, the PANDA dictionary, Huxtable's West Africa dictionary and the Pevehouse's Balkans dictionaries.[6] We use these dictionaries to code into WEIS categories supplemented with 26 new codes that deal primarily with internal events (see Appendix). After merging these dictionaries, we eliminated assorted phrases that remained from times when KEDS contained bugs[7] (e.g. numerous "- $ * +" patterns, which are simply the KEDS default coding) and finally removed most verb phrases containing than a half-dozen words on the grounds that such phrases would be encountered only rarely. Our hope is that the resulting dictionary contains most of the verb phrases used by Reuters to describe international political events, as well as discard codes for a wide variety of athletic events, natural disasters, and fatal mishaps in the transportation sector.[8]

---

[6] These dictionaries were not developed independently so they contain very substantial amounts of overlap.

[7] More precisely, contained more bugs than it now contains.

[8] We are continuing to refine this list based on work we have been doing with internal events, and expect to have a "final" standard verb list by mid-summer 1998, coinciding with the final release of KEDS version 1.0.

We have also created a standard *.actors* dictionary that lists all significant states and international organizations in the international system, as well as major-power political leaders such as U.S. presidents, assorted European prime ministers, and heads of UN organizations. For purposes of coding the Gulf data set, this was supplemented with all actors identified in the Actor_Filter KWIC index as occurring eight or more times in the lead sentences. There are approximates 80,000 leads in the source texts, so we know that any single actor not included in the dictionary occurs in fewer than 0.01% of the texts (although obviously the total source texts containing uncoded actors is substantially larger than 0.01%.[9]).

Finally, the coding uses the KEDS complexity criterion

```
COMPLEX: VERBS[6] NOACTPRIOR
```

which means that a sentence is not coded if it contains more than six verbs, or if no actor is found prior to the verb.

The source texts were obtained using the NEXIS search command on the REUNA (Reuters North America) library:

```
(SAUDI! OR SAUDI ARABIA! OR IRAN! OR IRAQ! OR KUWAIT! OR GCC OR OMAN! OR
YEMEN! OR QATAR! OR BAHRAIN! OR UAE OR EMIRATE! OR DUBAI! OR ABU DHABI!)
AND DATE AFT ... AND DATE BEF ...
AND NOT (SOCCER! OR SPORT! OR OLYMPIC! OR TENNIS OR BASKETBALL OR NBA OR
T.STRM OR HEADLINE(HIGHLIGHTS OR (WORLD W/2 OUTLOOK) OR (KEY W/1 FACTS)
OR (EVENTS W/1 SCHEDULED) OR (HISTORICAL W/1 CALENDAR) ) )
```

This locates stories that mention any of the listed actors anywhere in the story, so the resulting data set contains events dealing with these actors and any other actor in the international system. The additional criteria are used to eliminate most chronologies, reviews, calenders and the like.

---

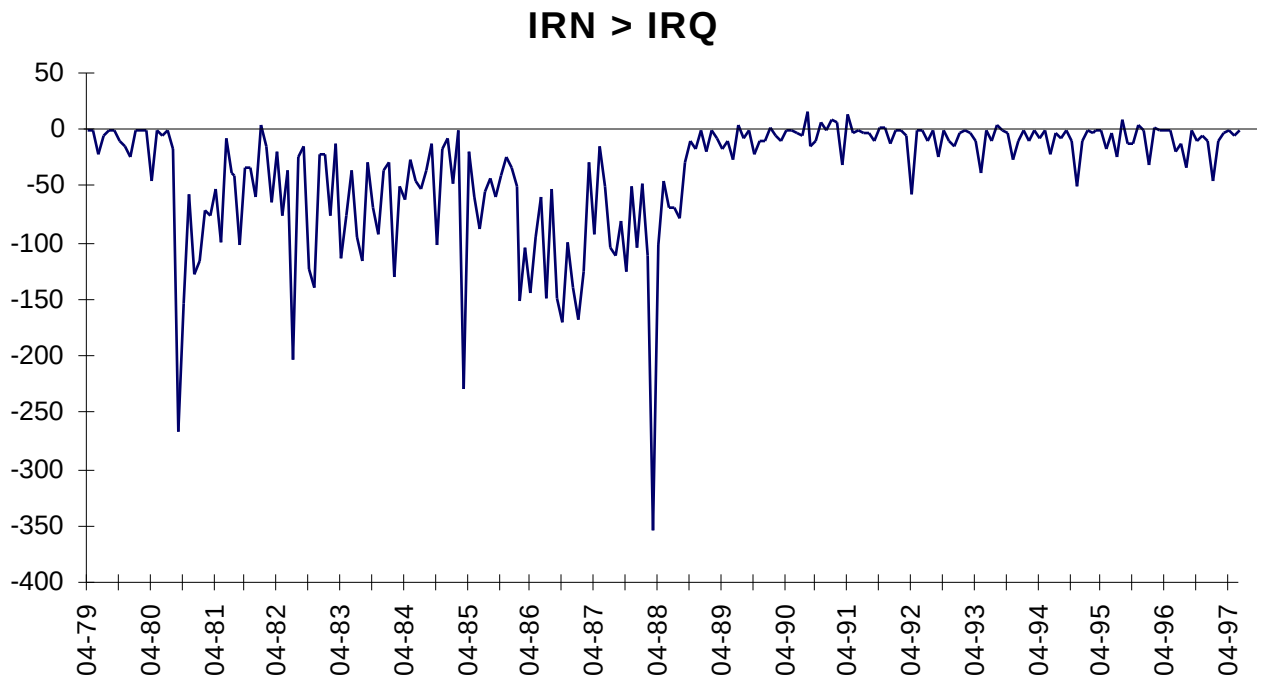
[9] Many of the proper nouns that occur fewer than 8 times involve individuals who achieve their "15-minutes-of-fame" as hostages or other victims of terrorism, leaders of transient political parties and the like.

Examples of time series

This section shows examples of the times series that were produced by aggregating the events by month using the Goldstein scale. Unless otherwise noted, the figures are based on the event data set generated by coding only the lead sentences.

Figure 3 shows the Iran → Iran sequence (the Iraq → Iran sequence is similar: the two correlate with r = 0.84). The major offensives of the Iran-Iraq war are clearly visible, as are the negotiations and overtures that Iraq made to Iran following Iraq's invasion of Kuwait. The 1990s, as expected, are characterized by sporadic disagreements but only limited activity compared to the 1980s.

Figure 3: Goldstein series for Iran → Iraq



The graphs of the Saudi → USA relationship in Figure 4 show a sequence of ebbs and flows in the relations between the two states (also note that the vertical scale is about one-tenth that of Figure 3). An examination of the individual events shows that the negative values correspond to

diplomatic disagreements between the two governments, for example on the controversy over the sale of AWACS aircraft in the early 1980s, and over United States' policy towards Iran and Israel later in that decade. The only large anomalous negative value occurs in January and February 1991, where there are a number of incorrectly-coded uses of force. These come from reports about the Second Gulf War where KEDS failed to accurately determine the appropriate actors. The other very large negative spike is due to the Khobar Towers bombing and the diplomatic disagreements between the two governments over the investigation of that incident. An examination of the individual events shows that the event record consists primarily of meetings, agreements and statements of policy support. While the lead-story sequence in Figure 4a shows a fair amount of detail, the full-story sequence in Figure 4b shows much more, with higher peaks during some periods of activity.

Figure 4a: Goldstein series for Saudi Arabia → USA, lead-sentence events

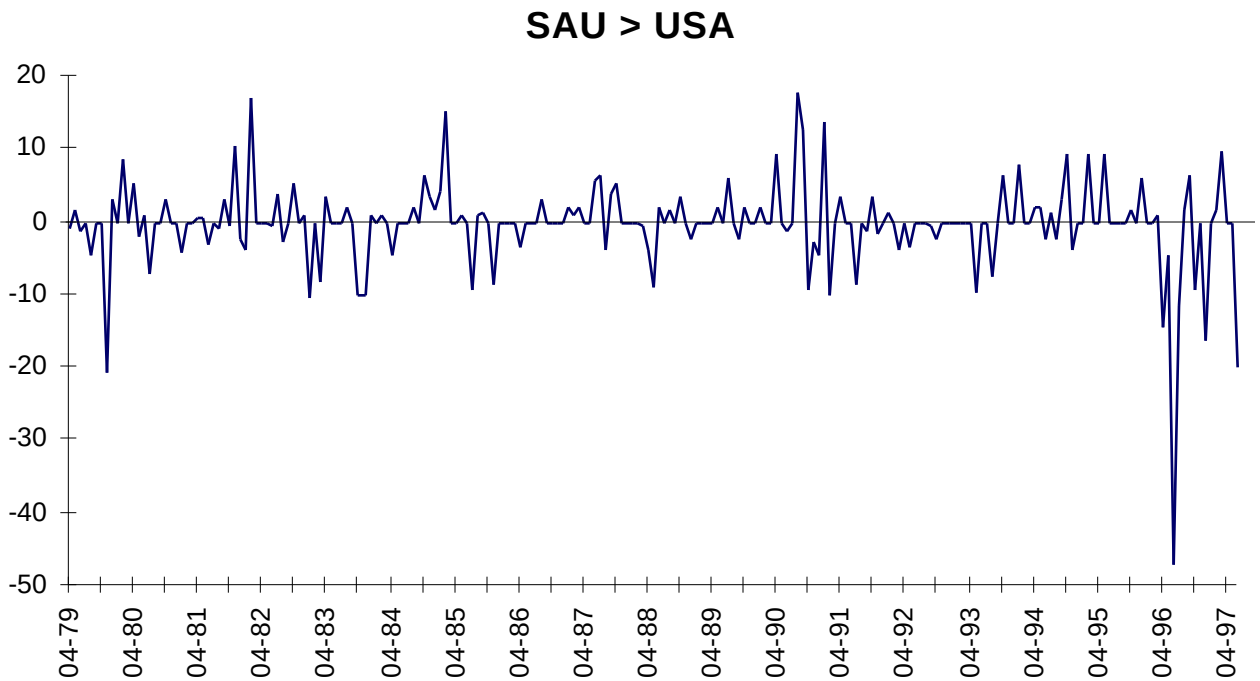


Figure 4b: Goldstein series for Saudi Arabia → USA, full-story events

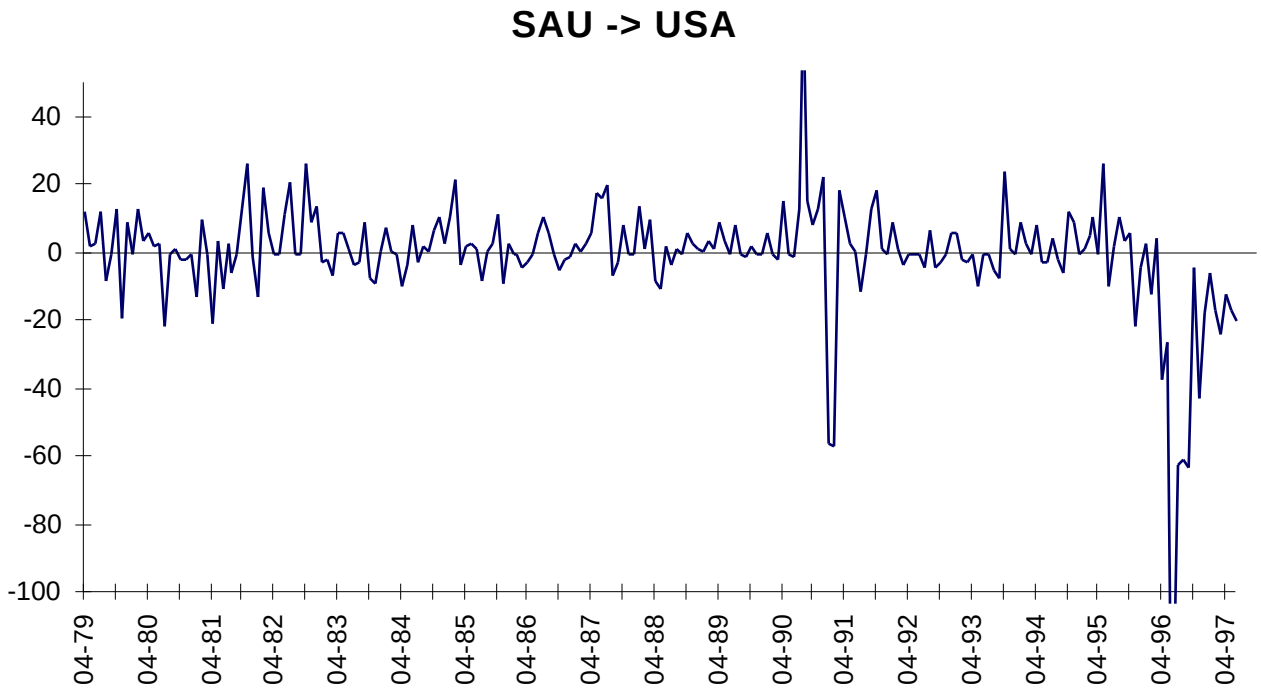


Figure 5: Goldstein series for USA → Iraq

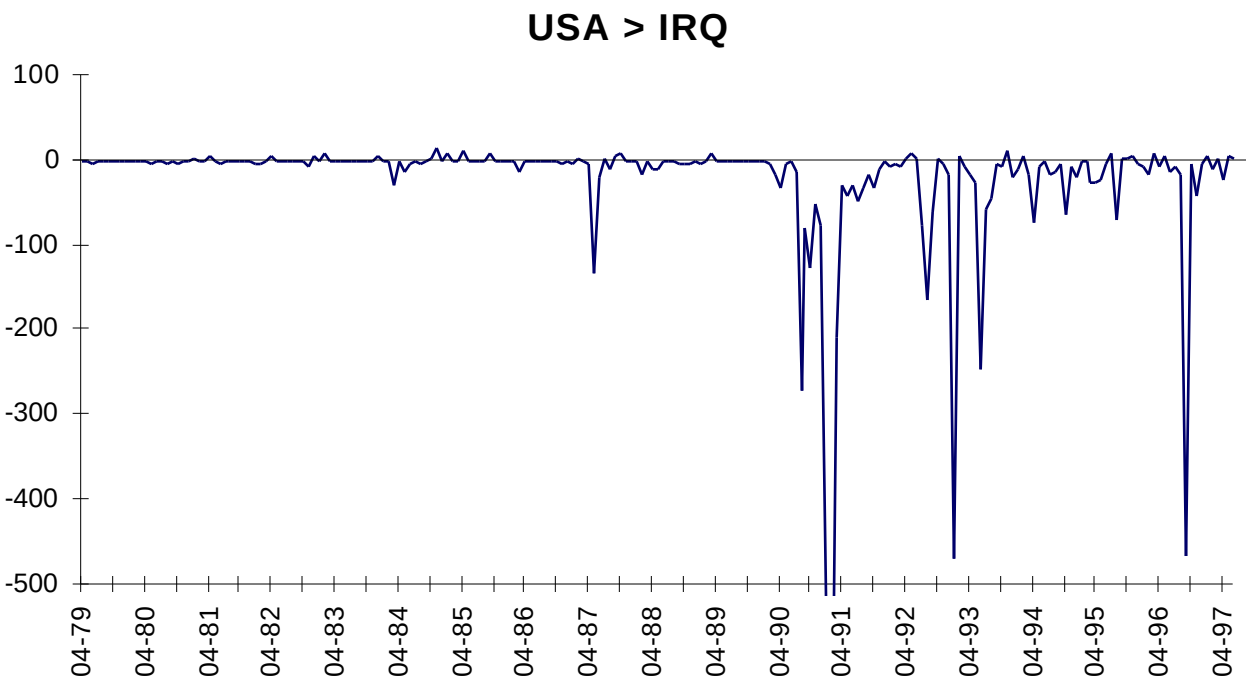


Figure 5 shows the USA → Iraq relationship. Unsurprisingly, the most conspicuous feature of the series is the Second Gulf War, where the Goldstein score actually goes down to -1000 in January and February 1991. The 1980s show a variety of consultations and agreements between the two countries (these are not very clear in Figure 5 due to the scaling); the negative spike in 1987 corresponds to the Iraqi missile attack on the U.S.S. Stark. The periodic US military attacks on Iraq in the period following the Second Gulf War are clearly visible, as are the diplomatic altercations and resolutions dealing with weapons inspections.

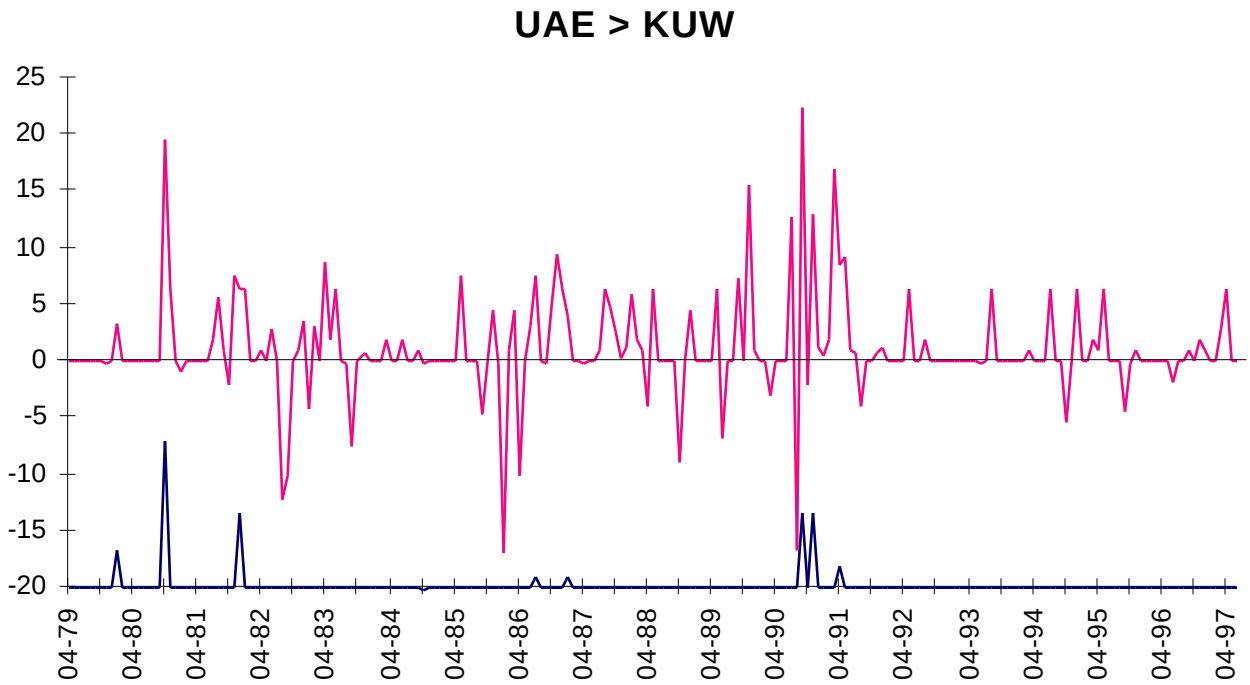
Figure 6 shows the UAE → Kuwait series for both the full-story coding and the lead-sentence coding; the latter has been displaced by -20 and rests on the bottom of the graph. As expected, most of the scaled events are positive, but there are a few instances of negative values despite the fact that Kuwait and the UAE are allies. These are due to three sources:

1. Disagreements over policy
2. Violent events due to criminal activity
3. Violent events that are due to coding errors.

The first category is legitimate—many of the cooperative events in the data set are the result of meetings, and sometimes meetings result in agreements, sometimes in disagreements. The criminal activity—usually a citizen of one state murdered or kidnapped in the other—is more problematic: this is not overtly political but it can have political implications. In the sequences we examined in detail, the violent events that are due to coding errors are relatively rare—the standard dictionaries and the complexity filter prevents most of these—but they still exist.

Erroneously coded violent events (WEIS cue categories 20 through 22) can be problematic because they have high Goldstein scores and therefore have a disproportionate effect on the aggregated measure. One solution to this problem would be to simply filter out such events in the computation of the Goldstein score for any dyad that was known, *a priori*, not to be engaged in military conflict. This would automatically remove the reports of deaths and seizures of persons and property that were due to criminal activity or incorrect coding; any remaining negative scores would be known to be caused by policy disagreements.

Figure 6: Goldstein series for UAE → Kuwait, full-story (top) and lead-sentence events



Comparison of Lead and Full Story Coding

The work we have done on the Levant has always coded only the lead sentences of stories. We used this approach for two reasons. First, the lead sentence is supposed to summarize the entire story and is therefore likely to follow a simple declarative form that can be coded correctly by KEDS. Second, our experience with Reuters reports in the Levant indicated that few events of any significance were reported only in the body of a story without also being covered in a separate story (and hence lead) of their own. In fact, the more common problem we encountered in reports from the Levant was the opposite: important events usually generated multiple stories.

We recognized, however, that this characteristic of Reuters might be specific to the Levant, which is reported very intensely. When our colleague Phillip Huxtable was working on a data set for West Africa—a region that does not receive intense coverage in the international press—he noted that Reuters would sometimes append a series of event reports to the end of a story that were unrelated, except in their regional focus, to the primary focus of the story. The underlying

editorial model appeared to be "Hey, if you're sufficiently interested in this out-of-the-way region to read this far, you'll probably enjoy this other stuff as well." Huxtable concluded that it might be possible to increase the density of events by full-story coding.

This observation, plus the fact that we'd never actually tested our assumptions about the adequacy of leads, led us to run some specific tests comparing lead-sentence and full-story coding on the Gulf data set. After downloading the full stories from NEXIS, we filtered the text to get all sentences that did not appear to be direct quotes.[10] Lengthy direct quotations tend to be very difficult to code and the main point that the speaker is making is usually summarized by the Reuters text outside of the quotations. The remaining sentences were coded using the same dictionaries and complexity filter used to code the lead sentences.

The full-story coding (and downloading) was substantially more time-consuming than lead-sentence coding. Coding the lead sentences required about 2-hours; coding the full stories required a full 24-hours. The full-story coding generated 264,421 events, as opposed to the 48,721 events generated from the leads.

Table 1 shows the correlation (r) between the monthly Goldstein series generated from the lead-sentence and full-story coding for 30 directed dyads. With the exception of dyads involving the United Arab Emirates (UAE), the correlations of the two series are quite high, usually above 0.75. This suggests that in most statistical studies involving linear models, similar results will usually be obtained with either approach. More generally, lead-sentence coding is probably quite adequate for exploratory work, given the much greater investment of time required to download complete stories.

---

[10] The filter skipped sentences containing a double-quote character (ASCII 34) that were either preceded by a period (.) or followed by a comma (,). This eliminated all correctly formatted sentences in Reuters that quote a speaker, while retaining sentences that contain short phrases placed inside quotes such as:

> Palestinian diplomacy has ended Lebanon's bloody "camps war" but analysts say it is likely to prompt a confrontation between Israel and Amal.

Table 1. Correlation of Goldstein-score time series generated with leads and full stories

	Target					
Source	IRN	IRQ	SAU	USA	KUW	UAE
IRN	---	.90	.73	.85	.76	.47
IRQ	.94	---	.98	.98	.93	.70
SAU	.77	.74	---	.64	.52*	.31
USA	.89	.96	.75	---	.88	.52
KUW	.68	.94	.43	.55	---	.35
UAE	.38	.38	.35	.65	.46	---

N=119

* Excluding Feb-91. With Feb-91 included, the correlation is 0.17

The lower correlations associated with the UAE are consistent with Huxtable's observations about West Africa: states perceived as peripheral by the international news media are more likely to be discussed only in the body of a story and not in the lead. For example, the SAU → UAE dyad contains 200 reported events in the full-story series, but only 6 (!) in the lead sentences. The IRN → UAE series is less dramatic, but still has 126 events in the full series versus 21 in the leads.

Table 2 shows the dyadic reciprocity correlation (X → Y x Y → X; see Dixon 1996 and Goldstein and Freeman 1990) for the lead and full-story sequences. These correlations show a very clear pattern, with the full story reciprocity being higher in all but two cases. Once again, the correlations for the minor actors are substantially lower than those for major actors; in some cases they are not even statistically significant. Some of the reciprocity in the full stories may be artificial in the sense that a long story is more likely to present "the other side" and thus generate a "reciprocal" event where none occurred. However, a long story is also more likely to present secondary events that did occur but which did not by themselves justify a separate story (and thus a lead). Furthermore, the fact that these series are aggregated by month should reduce the likelihood that the observed reciprocity is purely artifactual.

Table 2. Dyadic reciprocity in Goldstein-score time series generated with leads and full stories

Dyad	Full	Leads
IRN-IRQ	.95	.84
IRN-SAU	.85	.61
IRN-USA	.80	.71
IRN-KUW	.71	.32
IRN-UAE	.25	.06
IRQ-SAU	.96	.78
IRQ-USA	.97	.92
IRQ-KUW	.84	.77
IRQ-UAE	.00	.02
SAU-USA	.73	.53
SAU-KUW	.63	.35
SAU-UAE	.76	.93
USA-KUW	.27	-.02
USA-UAE	.22	.27
KUW-UAE	.86	.64

Finally, the similarities between the full-story and lead-sentence series can also be found in more complex measures of the series. Figures 7 and 8 give examples of the autocorrelation functions and cross-correlation functions of some of these series. The shapes of the curves are similar, although the strongest correlations are found in the full-story data.

Figure 7: Autocorrelation function of Goldstein series for Saudi → Iraq for full-story ("ALL") and lead-sentence events

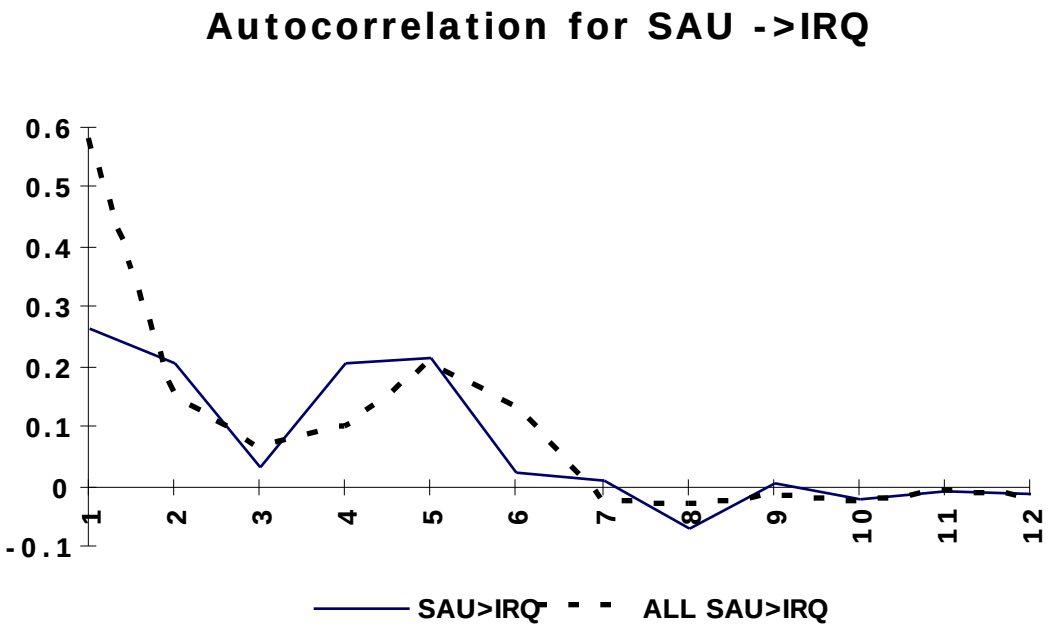
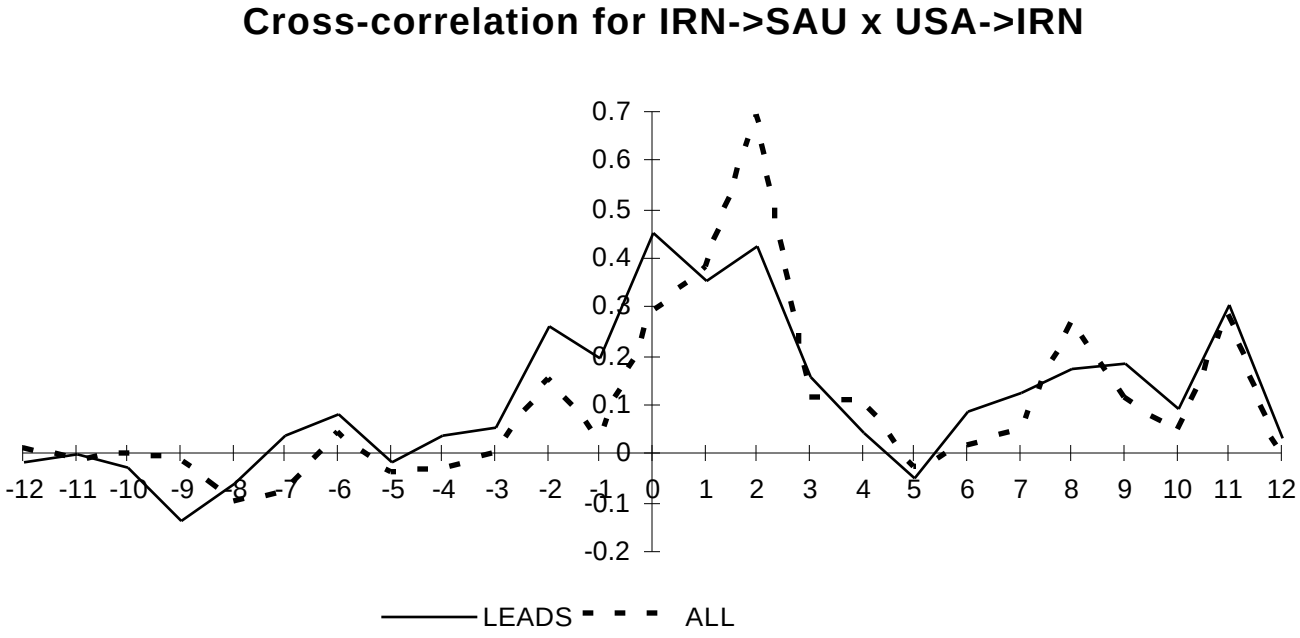


Figure 8: Cross-correlation function of Goldstein series for Iran → Saudi with USA → Iran for full-story ("ALL") and lead-sentence events



Conclusion

The results of this exercise demonstrate that by leveraging the existing investment in KEDS dictionary development, it is possible to generate new data sets with relatively little additional effort. At some point we hope to further refine the dictionaries used to code this data set, but the data produced from the standard verb dictionaries and actors identified by the Actor_Filter program are certainly credible.

In this concluding section, we will address three issues. First, what implications do these results have for the tradeoff between customized and general-purpose data sets? Second, how much effort should one spend improving dictionaries (a.k.a. "the psychopathology of dictionary tweaking")? Finally, what is gained from full story coding?

Custom versus General Event Data Sets

Imagine that academic researchers had access to a machine-readable copy of Reuters with the same ease that they can access human-readable copies of *The New York Times* (the traditional source of event data). Any decent research library has copies of *The Times* on microfilm. If we want students to examine events prior to the outbreak of the 1967 Middle East War, we just send them to the library and they read the relevant account at no cost.[11] Human-coded event data can take advantage of this accessibility, but human coding is extremely slow and expensive

If Reuters were similarly accessible, we would simply download the appropriate texts from the library via a high-speed communications link, reformat them, and code them. Under those circumstances, there would be very little reason for standardized event data sets: Because machine coding is completely reproducible, archiving dictionaries and search commands would be equivalent to archiving the original data, just as a research project that has constructed a scale from the American National Election Survey data reports only the questions used in the scale, not the transformed data.

---

[11] other than the cost of the aspirin required to deal with the consequences of using the typical microfilm reader.

This hypothetical situation would also lead to much greater experimentation with event data coding schemes, a development that we believe is long overdue. Most research designs that employ event data focus on only a small subset of political behaviors among a relatively small set of states. When dictionaries of verb phrases are available, it is relatively easy to develop a new coding scheme because it is much easier to assign a code to a known verb phrase than it is to anticipate that the verb phrase will occur. As we noted earlier, *.actors* dictionaries need to be supplemented with the names of regional actors, and internal political activity is frequently rather idiosyncratic.

That is our ideal scenario: customized, regionally-specific, machine-coded event data. Unfortunately, the current situation is far from this ideal. Reuters is not available cost-free from the library, but only through relatively scarce NEXIS subscriptions and through a very slow downloading process. For the past nine months, current reports from Reuters have not been available from NEXIS at all. From the perspective of machine coding, the books are still locked up in the monastery libraries, chained to the desks, and copied by hand onto expensive parchment.

This situation is presumably temporary. Archives of most news periodicals (although not Reuters) are already available electronically through services such as NEXIS and, increasingly, on the World Wide Web. Once in electronic form, the cost of moving those archives to a high-density archival medium such as DVD is minimal, and someone will presumably find a way to profitably sell these archives to libraries, just as microfilm is currently sold.[12]

---

[12] The economics of the NEXIS archives also make very little sense in the contemporary computing environment. The NEXIS system of storing files on a central server was economically rational when storage and processing power were sufficiently expensive to justify the time, bandwidth, and hardware dedicated to a NEXIS connection. But with the dramatic decrease in the cost of storage, it makes little sense to devote hours of low-speed connection time to transferring data that would fit on a $1 CD-ROM. It also makes little sense to process those archives centrally on a NEXIS mainframe when a desktop computer could accomplish the task as almost as easily. At some point, these technological changes will (presumably) cause a change in the attractiveness of NEXIS, although apparently NEXIS has anachronistic but exclusive contracts on many of its data sources.

Meanwhile, we *love* those educational subscription rates...

It would be very nice if this occurred with Reuters, but we are not holding our collective breath—Reuters has become notorious in the world of information services for resisting innovation. More likely we will find that by the year 2010, a more up-to-date company such as CNN, AFP, Bloomberg or even Microsoft will fill this niche, and provide full-text sources going back to 1990 or so. But until that point is reached, machine-coding will be constrained by the availability of source texts, so please make your coded data sets available on the Web!

The Psychopathology of Dictionary Tweaking

When are machine-coded data "good enough"? This is a very difficult issue, and some people appear to have a profound psychological block to using any data set that might, with additional tweaking of the dictionaries, be further improved. Yet if one backs off and looks at the overall research exercise, it is abundantly clear that a research project can very quickly reach a point of diminishing returns in dictionary development. Event data are produced from a nonrandomly censored sample of source texts, categorized with an imperfect set of codes, scaled arbitrarily into interval-level variables, and then applied in mis-specified models.

We suspect that the attractiveness of dictionary tweaking comes from the *illusion* of control over the analytical process—add a verb phrase, and you can see an event go from a use of force to an more innocuous verbal protest. The data are visibly better—progress! Let's grab another cup of coffee and keep on tweaking![13]

Don't kid yourself. The phrase you just changed may occur once or twice in the entire set of texts, and there's a non-zero chance that you've added something that will mess up other phrases that were previously coding texts correctly.[14] Even if you add hundreds of phrases, you are probably changing less than 5% of the coded events. By using the Actor_Filter program, you already know that no major actor references have gone unnoticed. Tens of hours of tweaking

---

[13] "Hi, my name is Jim and I tweak dictionaries"?

[14] We speak from experience....

may result in only negligible (and random) changes in the coefficient estimates of your final model.

Except for the addition of vocabulary to deal with regionally-specific activities such as drug wars and collapsed Ponzi schemes, the time and effort that one is tempted to spend on dictionary tweaking is usually far better spent on refinement of statistical models and estimation techniques. This can be a much less satisfying process: Instead of quietly punching the <Return> key and periodically solving little word puzzles, one must struggle to comprehend esoteric matrix notation, cope with partially-debugged algorithms borrowed from econometricians, deal with the unsatisifying tradeoffs between specification error and collinearity, and make unknowable decisions about error covariance structures. You have to try numerous model specifications that go nowhere, and when you finally do get a model specification that gives plausible results, you wonder whether you've merely out-foxed the significance test.

Coding is easy. Data analysis is hard.

Okay, we're over-stating the case here to make a point. We are certainly delighted to see additional work done with the KEDS dictionaries, and in some circumstances—particularly those where the lines between internal and international events are very murky, as in the Balkans, the Great Lakes region of central Africa, or southern Lebanon—our standard dictionaries may require substantial augmentation. The same is true if one is coding internal events or activities that go beyond the Westphalian framework of the WEIS system such as economic interactions, criminal activity, environmentally-induced stress or refugee movements. In these cases, it is probably a very good idea to develop a new coding system in addition to working on new dictionaries. WEIS wasn't intended to cover all circumstances, and it doesn't.

But even in these situations, recognize the point of diminishing returns. The event data projects of the 1970s were criticized at the time for "Beer-budget analysis on champagne-budget data." Thanks to the economics of machine coding, we can get champagne-quality data for beer-quality investments of time and effort, so invest that newly-available surplus on the models.

Full-Story Coding

The comparison of the full-story and lead-sentence data sets confirmed our expectations: First, for major actors there is generally a high correlation between the Goldstein series generated only with the lead sentences and those series generated with full stories. Given that full-story coding involves substantially more downloading time, lead-sentence coding is probably sufficient for most exploratory work. That said, there is substantially more variance in the full-story series even for major actors, and if one is developing a model that can exploit this level of detail—for example a regionally-specific early warning model—then the full-story series is preferable.

Second, full-story coding provides dramatically more information on minor actors, which is consistent with Huxtable's conclusions about West Africa. Information is "out there" in the wire service reports, even if one has to dig for it, and minor actors are under-represented—sometimes dramatically—unless the entire story is coded.[15]

Finally, it appears that the full-story series can be coded using the same dictionaries and techniques developed for lead sentence coding. As noted earlier, there is undoubtedly room for improvement in the coding through the development of regionally-specific dictionaries, but for a first pass through the data, the standard dictionaries are sufficient. This provides further evidence that these dictionaries have captured most of the English-language phrases used in Reuters to describe international political activity.

---

[15] This may also apply to many of the human coded event data sets. From early experiments in the KEDS project, we know that human coders are more likely to miss secondary events reported in a story and are less likely to code all of the possible combinations of multiple-actor meetings. Furthermore, if the human coder, rather than meticulously coding every single sentence, relies (explicitly or cognitively) on a summary of the story, the secondary events involving minor actors will be missed. The marginal frequencies on events involving minor actors in human-coded data sets such as WEIS and COPDAB generally look more like the frequencies of lead-sentence coding than the frequencies of full-story coding, even if the coders were supposed to be working with the entire story..

By way of comparison, COPDAB contains about 350,000 international events for the period 1948-1978. Our full-story data set on the Gulf therefore records, on an annual basis, about the same number of events as COPDAB records *for the entire world*, despite the fact that COPDAB was generated by full-story coding from multiple regional sources.

Bibliography

Bond, Doug, Brad Bennett and William Vogele. 1994. "Data development and interaction events analysis using KEDS/PANDA: an interim report." Paper presented at the annual meeting of the International Studies Association, Washington.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor and Kurt Schock. 1997. "Mapping Mass Political Conflict and Civil Society: The Automated Development of Event Data." *Journal of Conflict Resolution* 41,4:553-579.

Dixon, William J. 1986. "Reciprocity in United States-Soviet Relations: Multiple Symmetry or Issue Linkage." *American Journal of Political Science* 30:421-45.

Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. "The Machine Coding of Events from Regional and International Sources." *International Studies Quarterly* 38:91-119.

Goldstein, Joshua S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution* 36:369-385.

Goldstein, Joshua S., and John R. Freeman. 1990. *Three-Way Street: Strategic Reciprocity in World Politics*. Chicago: University of Chicago Press.

Goldstein, Joshua S., and Jon C. Pevehouse. 1997. "Reciprocity, Bullying and International Cooperation: A Time-Series Analysis of the Bosnia Conflict." *American Political Science Review* 91,3: 515-530.

Huxtable, Phillip A. 1997. "Uncertainty and Foreign Policy-Making: Conflict and Cooperation in West Africa." Ph.D. dissertation, University of Kansas.

Huxtable, Phillip A. and Jon C. Pevehouse. 1996. "Potential Validity Problems in Events Data Collection." *International Studies Notes* 21,2: 8-19.

Laurance, Edward J. 1990. "Events Data and Policy Analysis." *Policy Sciences* 23:111-132.

Lehnert, Wendy and Beth Sundheim. 1991. "A Performance Evaluation of Text Analysis." *AI Magazine* 12,3:81-94.

Salton, Gerald. 1989. *Automatic Text Processing*. Reading, Mass: Addison-Wesley.

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

Schrodt, Philip A., and Deborah J. Gerner. 1995. "Empirical Indicators of Crisis Phase in the Middle East, 1979-1995." Paper presented at the annual meeting of the Peace Science Society, Columbus, Ohio.

Schrodt, Philip A., Shannon G. Davis, and Judith L. Weddle. 1994. "Political Science: KEDS—A Program for the Machine Coding of Event Data." *Social Science Computer Review* 12,3: 561-588.

Appendix: Supplemental WEIS Code for Internal Events

014: Accommodate, cease-fire	015: Cede power	
026: Appoint or elect	027: Alter rules	
034: Vote		
055: Promise rights		
067: Grant position		
083: Ally	084: Merge or integrate	
096: Request policy change	097: Request rights	
113: Defy law		
123: Investigate		
133: Symbolic act		
151: Issue command	152: Claim rights	
161: War policies	162: Warn of problem	
196: Strike	197: Censor	198: Withdraw from
203: Ban organization		
213: Hijack or kidnap		
224: Riot	224: Assassinate or torture	226: Attempt coup