Political Science: KEDS—A Program for the Machine Coding of Event Data

Philip A. Schrodt, Shannon G. Davis and Judith L. Weddle[*]
University of Kansas

Final revision: 28 June 1994

Abstract

This paper describes in technical detail the Kansas Event Data System (KEDS) and summarizes our experience in coding Reuters data for the Middle East. The components of KEDS are first described; this discussion is intended to provide sufficient detail about the program that one could develop a more sophisticated machine-coding system based on our research. We then discuss a number of problems we have encountered in machine coding, focusing on the Reuters data source and the KEDS program itself. The paper concludes with a discussion of future approaches to machine coding in event data research and other potential applications of the technology.

KEYWORDS: event data, natural language, full-text databases, international relations, social science.

Introduction

Event data are used in quantitative international relations research to reduce journalistic descriptions of international interactions to categorical data that can be analyzed statistically. An early paper by Charles McClelland (1967) provides the following general definition of international events:

> Event-interaction is meant to refer to something very discrete and simple—to the veritable building blocks of international politics. The content of diplomatic history is made up, in large measure, of event-interactions. They are the specific elements of streams of exchange between nations. Here are a few examples for hypothetical Nations A and B: Nation A proposes a trade negotiation, Nation B rejects the proposal, Nation A accuses B of hostile intentions, Nation B denies the accusation, Nation B deploys troops along a disputed boundary, Nation A requests that the troops be withdrawn, ... Each act undertaken by each actor in this illustration is regarded as an event-interaction. (pg. 8)

Two global event data sets have been developed—the Conflict and Peace Data Bank (COPDAB; Azar, 1982) and the World Events Interactions Survey (WEIS; McClelland, 1976)—and a variety of more specialized data sets covering specific regions or time periods are also available.

Event data are one of the most common types of information used in quantitative international relations research.[1] Historically, event data have been coded by legions of bored undergraduates and M.A. students flipping through copies of the *New York Times* and other printed sources. The Kansas Event Data System (KEDS) is an alternative system that generates machine coded event data directly from machine-readable sources such as wire service reports.

====================
TABLE 1 ABOUT HERE
====================

KEDS has been used in two major projects, both using Reuters news service leads (see Table 1).[2] At the University of Kansas, we have generated a 12-year event data series for the Middle

East (Egypt, Israel, Jordan, Lebanon, the Palestinians and Syria) using the WEIS coding system; this data set contains about 60,000 events. This data set is described and its reliability assessed with respect to human-coded data in Schrodt and Gerner (1994); no systematic differences were found between the machine-coded data and the human-coded data.

The Protocol for the Assessment of Nonviolent Direct Action (PANDA) at the Program on Nonviolent Sanctions in Conflict and Defense at the Center for International Affairs at Harvard (Bond, Bennett, and Vogele 1994) is using KEDS to code a superset of the WEIS categories (160 categories versus the 63 categories in WEIS) that provide far more detail on nonviolent events, substate actors and internal interactions such as strikes and protests. PANDA codes several contextual variables in addition to the standard date-source-event-target variables of event data.[3] Reuters reports dealing with the entire world have been coded for 1985-1994; the resulting data set contains about 500,000 events.

Machine coding provides at least two advantages over traditional human coding. First, coding can be done quickly and inexpensively; in fact the coding of a large machine-readable data set by a single researcher is feasible. This allows a researcher to experiment with a variety of different coding schemes, which is impractical with human coding. Second, machine coding rules are applied with complete consistency and are not subject to inter-coder disparities.

One presumed disadvantage of machine coding is the inability of a machine to take into account the political context of events. However, as we have argued elsewhere (Gerner et al., 1994), such contextual interpretation often introduces systematic biases into coding and probably creates at least as many problems as it solves, particularly when one is trying to maintain a data set over a long period of time. Machine coding, in contrast, is completely consistent and reproducible. Machines remains at a clear disadvantage in interpreting texts containing complex grammatical constructions.

Machine Coding

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 largely defined by sets of transitive verbs (verbs that have a direct object). In most cases, event data coding needs only to focus on the basic subject-verb-object (SVO) structure of an English sentence, so event data coding is substantially easier than many other natural language processing problems. In particular, it is easier than many content analysis problems where attitudes or affect must be inferred from a text.. 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.*[4] 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 it is successful on the sentence structures most commonly used to describe events.

The Program

KEDS is a Macintosh program with a standard Macintosh interface; it has a suggested memory application size of 1 Mb and has thus far worked on every Macintosh model we've tried and under both System 6 and System 7.[5] KEDS is written in Pascal and contains about 10,000 lines of code. Much of this code deals with the interface; the core routines handling parsing and coding involve only about 2000 lines of code.

KEDS will work with any input text containing records in the format

```
<date> <any other information>
<source text line 1>
<source text line 2> ...
<source text line n>
<blank line>
```

The source text is presently limited to 16 eighty-character lines containing a maximum of 1024 words. We download Reuters leads from the Nexis data service; a simple filter program is used to remove irrelevant information such as page headings and to format the text. The filter program also converts the English date given in Reuters to a YYMMDD numerical data format and assigns a unique identification number to the lead. Finally, the filter checks each day for multiple leads that have almost identical letter counts; this eliminates stories rebroadcast by Reuters to correct spelling errors.

KEDS is not a general purpose parser; instead it is optimized for event data coding. KEDS cannot cope with text such as political speeches or legal documents that use complex sentence structures. However, most Reuters leads are relatively simple declarative sentences that correspond closely to the SVO format assumed by event data. Consequently a simple parser can successfully convert a high percentage of Reuters reports into event data. KEDS processes a text in four stages:

1. Word typing

After the source text is converted to a standard form—all letters are changed to capitals and commas are delimited with spaces—the program checks each word in the text to see if it occurs in the actor and verb dictionaries. If the word is found, it is assigned the appropriate type (e.g. actor, verb, pronoun, conjunction); otherwise it is designated as untyped. Most of the subsequent parsing operations deal only the words that have been typed. The dictionaries are stored alphabetically in an array of pointers indexed on the first two letters of the entries so that only a limited number of words need to be checked to determine whether an unknown word is present.

2. Process local grammatical structures.

This processing includes assigning actor identities to common nouns (agents), assigning the references to pronouns, using titles to reduce two actor references to a single actor (e.g. "Israeli Prime Minister Rabin" is reduced to a single reference to Israel), recognizing compound noun phrases and eliminating subordinate phrases delimited by commas.

3. Event coding.

The program attempts to match the patterns associated with each verb in the sentence. Patterns typically distinguish between direct objects, as in the distinction between PROMISED MILITARY

AID and PROMISED TO VETO. If a verb phrase corresponding to an event is identified, the program then finds the source and target associated with the verb. The source is usually the first actor in the sentence; the target is usually the first actor after the verb that has a code distinct from the code of the source, or else the first actor before the verb that has a code distinct from the code of the source. If the source or target are compound phrases, these are expanded into multiple events. Only the first verb corresponding to an event is coded, unless the sentence is compound (i.e. contains a conjunction not associated with a compound actor), in which case each clause of the compound sentence is checked for an event.

4. Issue processing

In the PANDA coding, a final step identifies the location, issue, domain and context of the event. The location variable uses some syntactic information; for example prepositional phrases are recognized. The remaining variables are determined by using simple keyword matching on the original text after the subordinate clauses eliminated in step [2] have been reinserted.

Dictionaries

KEDS was designed as a general-purpose coding system rather than a WEIS-coding program, so most of its coding decisions depend on its dictionaries.[6] This section will describe the various types of coding information contained in these files.

A standard input format is used for the dictionaries. Words are entered in upper case; codes for actors and events are enclosed in square brackets. If two words must be consecutive, they are connected by an underscore; if the two words are separated by a space, other words can intervene. For example, the text "agreed to provide a loan" would match AGREED LOAN but not AGREED_TO_LOAN. All of KEDS's files are stored externally in ASCII format so that the files can be edited using a word processor and transmitted using electronic mail.

====================
TABLE 2 ABOUT HERE
====================

The current size of the dictionaries used in the two projects (as of December 1993) are shown in Table 2. Other work in automated text processing (Lehnert & Sundheim, 1991) indicates that dictionaries on the order of 5,000 phrases are necessary for relatively complete discrimination between political events described by English language news media sources, so these projects are probably beginning to close in on a relatively complete vocabulary. Despite the fact that PANDA is coding the entire world while the Kansas project codes only the Middle East, the actor dictionaries are about the same size because a fairly complete list of actors is required to identify potential targets.

Stemming

If a word ends in a space, it is used as a stem. Stemming refers to the process of reducing different forms of a word to a single root:

ACCEPT	ACCEPTS, ACCEPTED, ACCEPTING
SYRIA	SYRIA'S, SYRIAN, SYRIANS

KEDS handles stemming by matching patterns from the beginning of the word; a word is considered to match if every character in the root matches. In other words, SYRI will match all four forms of SYRIA but it will not match SYRACUSE. Long phrases are searched before shorter ones: for example SIGNALLED, SIGNED and SIGN are checked in that order. An underscore after the word means that the word will match only if it is followed by a space, so the root "OF_" will only match the single word OF whereas "OF" would match "OF, OFFER and OFFICIAL.

Stemming had two advantages in the early stages of our work when the dictionaries were relatively small. First, most regular verbs had to be entered only once.[7] Second, nouns related to a verb—for example MEETING/MEET or ACCEPTANCE/ACCEPT—would trigger a correct classification even when the actual verb found in the source text was not in the dictionary.

With more extensive dictionaries, however, stemming is the most frequent cause of wildly inaccurate coding errors. For example, in coding November 1993, the usually problem-free verb BEAT matched BEATY, the name of an American businessman released from prison by Iraq. A future version of KEDS may incorporate a facility for explicitly defining regular verb constructions and noun endings (e.g. plurals and adjectival forms) and drop stemming.

Proper Nouns: Actors

An *Actors* file contains proper nouns and associates each of these with a code.

```
ABU_SHARIF [PLO]
ACQUINO [PHL]
AL-WAZIR [PAL]
AMMAN [JOR]
AMNESTY_INTERNATIONAL [NGO]
ANKARA [TUR]
ANTIGUA [ATI]
```

Multiple nouns can map to the same code; for example ISRAEL, ISRAELI, ARENS, PERES, RABIN, TEL_AVIV and SHAMIR all have the code ISR.

In some circumstances, it is useful to have a single phrase generate multiple actor codes. In KEDS this is indicated by separating the actor codes with a slash ('/'):

```
EAST_AND_WEST_GERMANY [GME/GMW]
NORTH_AND_SOUTH_KOREA [KON/KOS]
G7 [USA/GMW/FRN/ITL/UK/JAP/CAN]
```

KEDS also recognizes compound nouns of the form

<actor$_1$> AND <actor$_2$>

<actor$_1$> , <actor$_2$> , ... <actor$_{n-1}$> AND <actor$_n$>

and does the appropriate duplication of events. For example, "The United States and Egypt approved of efforts by Israel and Jordan" would generate the four events

```
USA  <APPROVED>  ISR            USA  <APPROVED>  JOR
UAR  <APPROVED>  ISR            UAR  <APPROVED>  JOR
```

Common Nouns: Agents

The original event data sets were state-centered and made little or no distinction among substate actors. This convention is also found in many Reuters leads, where the names of the states are used to refer to actions of governments or foreign ministries (e.g. "Israel accused Syria..."). However, in many circumstances it is useful to differentiate the agent responsible for an event, for example distinguishing "Israeli soldiers", "Israeli police" and "Israeli settlers". This is particularly important in the PANDA coding scheme, which deals with many internal political activities such as strikes, elections and protests.

KEDS deals with substate actors using "agents". In some cases, an agent is implicit in a proper noun—for example GEORGE BUSH was president of the United States—and these codes are specified in the actors dictionary. In many other cases, the agent is identified by a common noun:

```
AGENT: DISSIDENT [OPP]
AGENT: ELECTORATE [CON]
AGENT: EMIGRANT [REF]
AGENT: EMIGRES [REF]
AGENT: ENVIRONMENTALIST [ENV]
```

KEDS attempts to assign an actor identity to all agents, using the following priority:

1. Implicit agents (e.g. GEORGE BUSH [USA:GOV])
2. <actor><agent> (e.g. FRENCH POLICE)
3. <agent><preposition><actor> (e.g. POLICE IN DAMASCUS)
4. an actor found within ±2 words of the agent

If none of these patterns occur, the agent is assumed to have no explicit actor and is then is treated as an actor when identifying sources and targets. The statement "Police fought demonstrators" will give generate an event of the form "*** POL *** DEM" where [***] is KEDS's code for an unknown actor.

Agents were one of the last features added to KEDS but their identification noticeably improves the coding. Agents, as nouns, are often the true subject or object in a sentence, and prior to the explicit coding of agents , KEDS often depended on treating adjectives (ISRAELI, SYRIAN) or the objects of prepositional phrases as if they were subjects or the objects of verbs. If no such adjective was found, the primary verb was incorrectly identified and instead a verb in a subordinate or adverbial clause was coded.

Pronouns

Pronouns occur frequently in Reuters:

> Turkey believes Iraq and Syria can cope with a decrease in vital water but they have lodged a protest with Ankara.

THEY refers to IRAQ AND SYRIA but the program must determine this in order to code the second clause of the compound sentence correctly. Ascertaining the references of pronouns is a very general problem in parsing;[8] the techniques used in KEDS are quite simple, but fairly effective when applied to Reuters leads.

HE SHE IT assign the first actor in the sentence

ITS assign the first actor prior to the pronoun

THEY assign either the first compound actor if one exists or else assign an actor followed by a word ending in 'S' or an agent (e.g. POLISH MILITIA, SYRIAN SOLDIERS)

Dereferencing is usually not critical in Reuters but it sometimes brings an actor reference to a point in the sentence where it will be correctly matched by a pattern. These rules are least effective on the pronoun IT because that word often refers to an activity rather than an actor.

Verbs

The verb file contains verb phrases and their associated event codes. This includes both simple verbs (e.g. VISITED) and verbs plus direct objects (e.g. PROMISED FUNDS). In the example

```
ACCEPT
- * PROPOSAL     [081]
- PROPOSAL WAS * [081]
- * CHARGES      [013]
- * FORMULATION  [042]
- * INVITATION   [082]
```

the root verb is ACCEPT; with stemming this will match ACCEPT, ACCEPTS, and ACCEPTED. The phrases that start with "-" are the patterns associated with ACCEPT and their codes; the "*" indicates where the verb itself should appear. For example ACCEPTED PROPOSAL will be coded 081 ("agree" in the WEIS coding system) while ACCEPTS FORMULATION will be coded 042 ("approve" in WEIS).

Patterns usually involve direct objects or modifiers such as NOT. The key to this scheme is ensuring that phrases are associated with a transitive verb rather than indicators of tense (e.g.

HAS, WILL) or forms of "to be" (IS, WAS, WERE). The important verb will often be an infinitive; for example in WILLING TO NEGOTIATE, the verb is NEGOTIATE. Pattern matching stops at any conjunction; this prevents a pattern from matching the direct object of the verb in a distinct clause of a compound sentence.

Patterns can also specify where the source and target are found in relation to the verb and associated words; these are indicated by "$" and "+" respectively. For example, the pattern

```
ADVISE
- + WAS * BY $
```

would make the correct source and target assignment on the passive construction "Egypt was advised by the United States." The symbol "%" specifies that a *compound actor* should be assigned to both the source and target. This is typically used when dealing with consultations—"Representatives of Syria and Jordan will meet in Cairo"—to indicate that the subject of the sentence contains both the source and target.

Paired and Subordinate Codes

The WEIS coding scheme often generates symmetric events of the form

<Actor$_1$> <Event$_1$> <Actor$_2$>
<Actor$_2$> <Event$_2$> <Actor$_1$>

For example, a meeting between Israel and Egypt would generate the pair

```
ISR 031 UAR         (meet with)
UAR 031 ISR         (meet with)
```

A visit by a Jordanian official to Syria would generate the pair

```
JOR 032 SYR         (visit; go to)
SYR 033 JOR         (receive visit; host)
```

In KEDS these are coded automatically by using a pair of codes separated by a slash ; for example

```
FLEW
- $ TO + [032/033]
```

would code the visit/receive pair.

KEDS can also set priorities when multiple verbs are found in a sentence. A *subordinate* code indicates that a verb is only to be coded if no other events are found. When a phrase with a subordinate code is encountered, KEDS continues to search for other verb patterns in the sentence rather than stopping. Its main use is coding comment events such as leads of the form

<actor$_1$> SAID <pronoun> <verb> <actor$_2$>

For example in coding "George Bush said he rejected Syria's assertion..." the relevant event is USA <REJECTED> SYRIA rather than USA <SAID> SYRIA. The combination of dereferencing the pronoun HE and using a subordinate code will handle this.

Other Parsing Features

Compound Sentences

When multiple patterns occur in a sentence, events are prioritized by:

- left to right order of verbs in the sentence
- length of patterns within a verb's pattern list

Events are only coded for the first verb unless the sentence is compound or the first verb is subordinate. KEDS recognizes compound sentences generated by the conjunctions AND and BUT after any AND in a compound noun phrase has been eliminated. In a compound sentence, the source is not changed unless it is explicitly reset by a $ operator in a pattern or if an actor occurs immediately after the conjunction. The target is reset according to the usual rules applying to verbs and targets.

Deletion of Subordinate Phrases Delimited by Commas

In Reuters leads, short phrases delimited by commas and phrases between a comma and the end of the sentence are usually irrelevant to the coding:

> President Mubarak, in a grim warning underlining Egypt's deepening economic crisis, will request emergency assistance from the IMF, the official UAE news agency said on Thursday

These phrases are eliminated if the number of words between the commas is greater than two and less than or equal to ten; the minimum allows the preservation of comma-delimited lists. The maximum and minimum length for an eliminated phrases can be set to alternative values by the user and commas inside numbers—e.g., "10,000"—do not trigger this feature.

Null Codes and Stop Words

The null code "---" is used to eliminate phrases that would otherwise be confused with actors or verbs. For example, the phrase "Israeli-occupied West Bank and Gaza" will generate both the ISR and PAL (Palestinian) codes as actors. By adding the null code

```
ISRAELI-OCCUPIED [---]
```

only PAL is generated as an actor. The phrase "The head of Lebanon's Catholics" would generate a verb identification for HEAD, a common verb[9], but adding the null code THE_HEAD_OF [---] eliminates this problem.

Null codes have proven surprisingly important in refining the coding of the system, particularly given the propensity of English to use words as both nouns and verbs. As discussed below, some of these verb/nouns present insurmountable problems, but null coding eliminates many troublesome phrases.

Issues

KEDS can code up to 4 sets of "issues": these are typically sets of words or phrases identifying the context or domain of an event. For example the PANDA "Issues" variable begins with the phrases:

```
ABORTION [T]
AIDS_ [E]
ANCESTRAL LAND [N?]
APARTHEID [H!]
ASYLUM [H]
BALLOT [G?]
BAN_THE_DEATH_PENALTY [P?]
BANKING [F!]
```

Issue phrases can be coded as dominant (!) or subordinate (?), or given a numerical priority between -254 and 255. The code for the issue with the highest priority will be assigned to the event. An issue can also default to the code of another variable; for example PANDA's "location" variable defaults to the source if no other location is found.

Discarding Events

Some news reports involve multiple international actors but no political event. Sports events are especially problematic given the propensity of Reuters to use national identities and martial metaphors in describing the athletic contests, particularly soccer ("Algeria blasts Spain in World Cup action").[10] Traffic accidents and natural disasters involving multinational fatalities are also a problem, as is transnational criminal activity. Such stories are discarded by specifying discard codes:

```
CODE 001 = ~~SPORTS
CODE 002 = ~~DRUGS

WORLD_CUP [001]
SOCCER [001]
MARIJUANA [002]
HEROIN [002]
```

If a discard phrase is found anywhere in the source text, no events are coded from the text.

User Interface

The main display for KEDS shows the source text, the coded events and some summary statistics. Unless a number of issues are being searched, coding occurs in less than a second after the source text is displayed. KEDS is designed for machine-assisted coding so, for example, it is possible to scroll forward and back among source texts, eliminated problematic phrases with comment delimiters, edit the source text and edit the coded events.

KEDS uses a "project file" to keep track of the actor and verb dictionary files, the source text file, and output files. The project file tracks up to 64 coding sessions, recording the coder, time and date, number of leads examined and the accuracy of the system during that session. These statistics can be displayed on the screen or written to a text file.

A *source window* initially displays the source text being coded along with its date and identification number. On machines with a color screen, KEDS has the option of showing how the text has been typed:[11] actors are shown in red, verbs in blue, agents in green, pronouns are replaced with their references and text eliminated by subordinate phrases or null codes is shown with a line through it. This typed display is particularly useful when one is trying to figure out why KEDS has made a set of coding decisions and in determining unintended word matches due to stemming.

The *event window* lists all of the events found in the source text; the display can be customized but typically includes the source and target, their agents (if any), the event code, the English interpretation of the event, and the specific text matched to generate the code. Incomplete events—those missing a source, target or event—are displayed in gray; complete events are in black. When using the program for machine assisted coding, clicking an event in this window allows the event to be edited.

In automatic coding mode, KEDS codes events without pausing to allow editing. The speed of autocoding is dependent on the size and organization of the dictionaries but using a Quadra 610, the Panda dictionaries and data from 1983, KEDS generated 8515 events from 7685 leads in 11 minutes and 34 seconds, or about 11 leads per second. In ideal circumstances, our 12-year Middle East series (62,000 events) would take about an hour and a half to recode. The output can be formatted in a variety of ways, including tab-delimited formats for use in database, spreadsheet or statistics programs.

Verbs and actors, as well as their associated codes, can be added, deleted or changed using a dialog invoked from the main program menu.[12] In the verb file, the dialog will list the patterns associated with the verb and these can be edited. The editing dialog also keeps track of the coder who added each phrase and the date the phrase was added.

Problems with Machine Coding

In the experience of PANDA and our project, KEDS assigns about 80%-90% of the same codes that a human coder would assign.[13] While these projects may be somewhat biased towards accepting KEDS's judgment and making coding interpretations that favor machine-codeable events, KEDS-coded WEIS data sets also compared quite well with independently produced, human-coded WEIS data (Schrodt & Gerner, 1994). The program is clearly in the ballpark.

Nonetheless, problems remain in the enterprise of machine-coding with both the Reuters source text and with KEDS itself.[14] This section will look at some of those issues, illustrating them with examples of problematic Reuters leads.

Reuters Problems

All factors considered, Reuters is a remarkably useful source. Using unedited Reuters leads, we were able to code 12 years of some of the most complex political interactions in the world (e.g. the Lebanese civil war; the Arab-Israeli peace negotiations) with a relatively limited vocabulary. This indicates that Reuters maintains fairly consistent levels of editorial control.[15] Reuters is an international news agency with a large number of readers who are not native speakers of English, so Reuters may use the English language more carefully than, say, the *Washington Post.* Nonetheless, four problematic characteristics of Reuters leads should be noted.

The first is the problem of attribution. The verb SAID, and related verbs such as REPORT, are used in two different ways in Reuters. In some cases, an entity that is not a political actor—typically a news service—is reporting an event. In other cases, a political actor is making an official statement. In some instances both uses occur in the same sentence, though no attribution in the lead is the most common form.

The KEDS coding system treats the isolated verb SAID as subordinate, which means it is coded (as a comment, WEIS 02) only if no other event is found in the lead. However, because SAID, combined a direct object, can in fact be the primary verb of the sentence, SAID has the longest pattern list of any of the verbs in our dictionary. This is partly a function of the WEIS coding scheme, which distinguishes between a variety of verbal behaviors, but is also a function of the fact that a great deal of international interaction is verbal behavior.

The subordinate SAID means that when one actor is making a statement concerning activities by two other actors, this will be coded as if the event actually occurred:

> *Lebanese* Prime Minister Rashid Karami said today the *Soviet Union* had agreed for the first time to help finance the *U. N.* peacekeeping force in Lebanon.

While Reuters tends to use verbs such as ACCUSED or ALLEGED in these circumstances, and reserves SAID for sources Reuters considers authoritative, KEDS codes a certain number of active events that are actually comments. The number of such cases is probably small but they add noise to the data.

A second problem in Reuters occurs with unidentified sources. This type of lead is relatively common, particularly in Lebanon, where violent interactions are reported involving "unnamed gunmen", "guerrilla groups" and other anonymous agents, and where unaffiliated violent events occur ("A bomb exploded near a checkpoint..."). In a protracted conflict such as Lebanon, these anonymous events probably have little effect on the overall event data series, since ample events occur where the identities of the actors are known. In addition, a group (or frequently, multiple groups) will usually claim responsibility for an action after the fact, and these claims generate events.

A third problem occurs with specialized vocabulary. While the Reuters vocabulary has generally been quite stable, one of the most difficult periods for us to code was the 1990-91 Iraq-Kuwait crisis. This period required substantial new vocabulary development because it dealt with behaviors we had not encountered in coding the low-intensity conflict in Lebanon and the Occupied Territories. During military operations, for example, ambiguous verbs such as ATTACK and FORCE, discussed below, usually refer to physical rather than rhetorical actions. One way to deal with this problem might have been to develop specialized dictionaries to deal only with this period, and then revert to our standard dictionaries once the crisis had ended.

Actors also change over time. The two most notable instances of this in our data are Boutros Boutros Ghali—who appears during 1982-1993 both as Egypt's foreign minister and as Secretary General of the United Nations—and the changes surrounding the collapse of the Soviet Union. We are in the process of adding a facility to KEDS that can restrict a code by time period.

Finally, while the majority of the reports in Reuters refer to political events, there are a number of feature stories providing human interest, political analysis or historical background. In general, such leads do not contain verbs corresponding to events and are not coded. Occasionally, however, they do, particularly when dealing with historical events. KEDS does not deal with conditional or hypothetical statements, so scattered through the machine-coded sequences are an assortment of non-existent acts of force between improbable dyads. These occur at random and would never be confused with the sustained conflict found in Lebanon, Desert Storm or the *intifada*, but any analytical techniques that were highly sensitive to unanticipated uses of force would need to filter them, much as U.S. nuclear warning systems learned to ignore the occasional flocks of geese and software glitches that computers mistook for incoming Soviet ICBMs.

KEDS Problems

The coding dictionaries used by KEDS must balance the need to find phrases that are sufficiently general to apply in multiple cases with the risk of coding errors when words are used in unanticipated contexts. The discussion below provides a number of examples where KEDS would incorrectly code events because of a particular grammatical construction used by Reuters, or because of ambiguities in the English language itself. This is by no means a comprehensive list of errors—Reuters and KEDS never fail to surprise us—but they are illustrative. In a number of these illustrations, we have made subsequent changes to the program and dictionaries to prevent the specific problems being demonstrated; in some the problems are irresolvable. In the examples, the *actors* KEDS identifies are in italics and verbs are underlined. The indented examples are actual Reuters reports; the illustrations used in the text have been created specifically to illustrate grammatical points.

Passive voice

> The *Soviet* embassy and trade mission in Beirut were hit by *Israeli* shellfire last night *and* suffered extensive damage, embassy staff said.

USR	ISR	223	(military engagement)
USR	ISR	222	(nonmilitary destruction)

The standard word ordering in an English sentence is subject-verb-object (SVO).[16] Because English, unlike many languages, does not provide for the declension of nouns, word ordering is the primary means by which subjects are distinguished from objects.

Once comma-delimited subordinate clauses have been removed, Reuters leads usually follow an SVO ordering when the subject and object are proper nouns. Because of this, the sparse-parsing approach of KEDS generally works quite well. The primary exception to SVO ordering is passive voice, which signals the reversal of subject and object (i.e. OVS). The sentences "Police fired on demonstrators" (active) and :Demonstrators were fired on by police" (passive) are identical in content but subject-object order is reversed.

Passive voice in English is usually signaled with the sequence <IS/WAS/WERE> <verb> <BY>. A number of patterns in the verbs list deal with this, but in retrospect, passive voice is sufficiently important that it should have been hard-coded into KEDS; a future version of the program probably will do this.[17]

Incorrect identification of passive voice usually results in the reversal of source and target but when statistically analyzing aggregate behavior in the Middle East, source-target reversal is less problematic than it might seem, since there is a great deal of tit-for-tat behavior in the region that averages out across dyads over time.

Ambiguous words

In a formal language, words are associated with a single meaning or a small set of related meanings. While this is often true when dealing with natural language—for example the English words ACCUSE and DENY are almost never incorrectly coded in Reuters leads—there are exceptions.

In our work, two words stand out as particularly problematic: FORCE and ATTACK. Both words can be used either as nouns ("A guerrilla force launched an attack") or as verbs ("Rebel radio said guerrillas would attack in order to force concessions") and occur frequently in reports about military conflict.[18] FORCE and ATTACK are further complicated because they can be used to refer both to verbal actions (persuasion and criticism) and to uses of force; both uses are common in Reuters. In our dictionaries, a large number of patterns are associated with each of these words to try to distinguish the noun usage from the verb usage. ARMS, BATTLE, BOYCOTT, FIRE, HELP, ORDER, PLAN, PLEDGE, STRIKE and SUPPORT are other examples of words that are used both as verbs and as nouns.

Another problematic situation arises from very short, common words. For example, BY is a useful marker for passive voice, but the *Random House College Dictionary* (1975) also lists 28 additional meanings for BY. There are 31 distinct meanings for IN and 25 meanings for TO. Consequently treating IN and TO as if they were *only* prepositions is a less than completely reliable strategy. A fluent speaker disambiguates these uses by context without even thinking—for example "The negotiators are going to the meeting to be held in the village in a week"—but a computer does not have this capability.

Proper nouns are rarely ambiguous, except where the role of an individual has changed over time. The notable exception is GEORGIA, which can refer either to a state in the southern United States or the country that was formerly part of the USSR.

Actor name occurring before the actual target

> *European Community* governments agreed in principle on Monday to a *German* proposal for *EC* financial aid to *Israel* to help it through the Gulf War
>
> EEC GER 071 (Extend economic aid)

This usually arises when there is an actor in the *direct* object of a verb and the correct target is in the *indirect* object. If the above sentence read "EC governments agreed to give financial aid to Israel" or "EC governments agreed to give Israel financial aid", KEDS would have coded it correctly. We've not determined any general way around this problem, though designated target codes (+) in patterns can deal with some of the common cases.

Event near the end of the lead

> A surprise setback today hit efforts to end the war between *Israeli* forces and Palestinian guerrillas when Syria rejected the idea of the entire *Palestine Liberation Organisation* moving to its territory
>
> ISR PAL 111 (Reject; turn down proposal)

Reuters will occasionally introduce an event by indicating why it is important, so that the relevant verb does not occur until well into the sentence. This provides ample opportunity for an extraneous actor to be associated with the verb.

A simple solution to this problem would be to avoid coding cases where an actor and verb are not found in the first few words of the sentence. Such sentences are unlikely to be in SVO format and therefore will probably be coded incorrectly. An analytical lead such as this example probably would have been preceded by an explicit SVO lead about the Syrian rejection (as well as indicating whose idea was being rejected) so the underlying event would have been coded from another story even if this lead was not.

Incorrect interpretation of conjunctions

> An artillery battle between *Israeli and Palestinian* forces in Beirut broke a 24-hour-old ceasefire today as President *Reagan* agreed in principle to send U.S. troops to help evacuate *Palestinians* from the city

ISR	USA	223	(military engagement)
PAL	USA	223	(military engagement)

In this example, the phrase ISRAELI AND PALESTINIAN was identified as a compound actor corresponding to the verb phrase BROKE CEASEFIRE and the target REAGAN. The resulting coding illustrates two problems arising from conjunctions.

First, the lead reports two events—a battle and a US agreement—separated by the conjunction AS. While AS is a perfectly legitimate English conjunction, it also has 26 other meanings (Random House, 1975), and therefore it is not in KEDS's list of conjunctions. If AND or BUT were used (or AS was recognized as a conjunction), KEDS would have correctly picked up REAGAN AGREED TO HELP PALESTINIANS, but this did not occur in the source text.

Second, while a number of verbs have patterns that are specifically looking for compound subjects—for example meetings and military engagements—BROKE CEASEFIRE does not, since this is typically done by only one actor. In this example, the ideal pattern would have focused on BATTLE BETWEEN and assigned a symmetric event with ISRAEL&PALESTINIAN as the compound actor. This would work despite the fact that BATTLE is actually a noun and the subject of the sentence, rather than a verb; this is a case KEDS would do the correct coding for the incorrect reason.

Excessive number of actors

> *Syria* said today the *U.S.* veto of a *U.N. Security Council* motion on *Israeli* settlements was "the most prominent phenomenon of *U.S.* hostility to the *Arabs* and *U.S.* support for *Israeli* plans to annex the *West Bank*"

This sentence contains nine actor references to six distinct actors (Syria, U.S., Security Council, Israel, Arabs, and Palestinians). The actors occur because a complex diplomatic process is being described—for example the object of the Syrian statement ("U.S. VETO ... ISRAELI SETTLEMENTS") involves three actors. Unless the multiple actors are neatly arranged in compound phrases, KEDS usually fails to correctly sort out the subject and object from the modifying phrases; the simplest way around this problem would be avoid coding sentences that contain an excessive number of actors.

Excessive number of verbs

> The PLO, raising the stakes before renewed Middle East peace talks, has accused the U.S. of cheating Palestinians by reneging on promises to grant Israel $10-billion in loan guarantees only if it halted all settlements in occupied territories.

This admittedly unusual—but authentic—sentence contains seven verb phrases: "raising the stakes", "renewed…talks", "accused", "cheating", "reneging on promises", "grant…loan", "halt…settlements". If the comma-delimited phrase "raising the stakes...talks" is removed, KEDS will actually code the sentence correctly because the initial part of the sentence has the SVO structure PLO ACCUSED U.S. but usually multiple verbs cause coding errors. Multiple verb phrases are very common in Reuters leads because a verb can be used in a variety of subordinate clauses, direct objects ("cheating", "grant"), and adverbial clauses ("only if it halted"). Except in the case of introductory clauses and reports (e.g. SAID), the verb determining an event is usually the first transitive verb in the sentence and is coded correctly.

Discussion and Future Research

KEDS was developed inductively. Because international event data are based on an SVO framework and most declarative sentences in English have that form, a few simple rules can be used to generate a large number of correctly coded events without human intervention.

In the early stages of developing the program, when our dictionaries were very limited, it was important to use shortcuts such as stemming to extract information in a syntactically incorrect fashion. However, as our coding dictionaries became more elaborate, the nature of the coding errors changed. Most verbs are now coded correctly and a higher proportion of the errors are due to incorrect source and target identification. These errors are caused by passive voice, subordinate and adjectival clauses, and ambiguous pronouns, and therefore can not be solved by simply adding vocabulary.

KEDS is thus facing something of a dilemma. A few of the remaining problems can be solved by adding specific grammatical features. We have done this with agents, we are in the process of working with prepositional phrases, and we expect to add a general facility for passive voice. Beyond this level, however, one needs to develop a much more sophisticated parser because most of the Reuters leads that create coding problems are quite complex.

There are, however, at least four arguments against developing a more sophisticated parser. First, sentences too complex to correctly code using the KEDS approach probably account for less than 10% of all Reuters leads. Second, we are not aware of any "off-the-shelf" parser-coders that perform substantially better than 90% on unedited source text.[19] Third, KEDS's sparse parsing is relatively robust in the sense that complex sentences are often assigned the correct code for the incorrect reason. Finally, we suspect that a more sophisticated approach will involve semantic information—information pertaining to word meanings rather than simply their type and relationship to each other—and this will involve a very substantial amount of work in terms of dictionary development. For problems more complex than event data coding, parsers with semantic information are often essential, but we suspect they are not needed for our problem.

This is not to say that an improved machine-coding system could not be created; in fact we would be absolutely delighted to see such an effort undertaken. It is simply to say that such a project will probably involve a substantial amount of work and from the standpoint of *our* research, it is more important to concentrate on refining the event coding schemes and the methods of analyzing event data rather than squeezing another 5% accuracy out of KEDS, particularly since that additional information may make very little difference in the statistical results obtained from the data.

Two facilities may be added, however. First, while KEDS currently attempts to code all source texts, cleaner data might be obtained if sentences that were clearly problematic—for example those showing excessive actors or verb phrases, those without a recognized verb early in the sentence, and those containing words known to be ambiguous—were skipped altogether or diverted to a separate file for human coding. Our impression is that Reuters contains a lot of redundant information and skipping a few complicated records is unlikely to significantly change the resulting data from the standpoint of statistical analysis.

Second, some specialized routines are clearly needed to deal with common grammatical constructs outside the SVO framework, for example passive voice and the use of prepositional phrases to identify agents ("police in Damascus"). At the present time these can be programmed into the core of the program but it may be possible to specify grammatical rules from external files, much as the actor and verb dictionaries are currently specified. The "augmented transition network" formalism common used to describe natural language structures (Salton, 1989, Chapter 11) may be one effective way of doing this.

Thus far we have been discussing the use of machine coding only for the purpose of generating international event data. However, the technology can be used more broadly.

First, the event data concept can be applied to fields other than international relations. For example, many of the events coded by the PANDA project are purely domestic interactions such as strikes and protests. Newswire sources such as Reuters, UPI and Agence France Presse cover a wide variety of social behaviors, and to the extent that actors are unambiguously identified and information can be coded from the SVO structure of a sentence, KEDS could be used to generate data from these reports.

Second, KEDS, in conjunction with the Boolean search capabilities of Nexis, can be used as a very sophisticated index or filter. Consider research on the topic of the dynamics of informal sectors in urban areas. A Nexis search can identify all information from defined sources on a given topic over a given time period. The search can be limited to headlines, which provide very limited information, or it can include whole news stories, but a whole story search will include feature stories with ill-defined lead paragraphs, obtuse language, and unspecified actors. Similarly, a subject search on "informal sectors" will yield an amorphous amalgam of stories, many of which will have no relevance to the specific interests of the researcher.

KEDS can be useful in this context in two ways. First, since the KEDS dictionaries are independent of the program, the researcher is not limited to current actor and verb lists, nor to using the WEIS coding scheme. Virtually any coding scheme that relies on an SVO structure can be applied to identifying and systematizing relevant information.

For example in the WEIS coding system, the lead "Sudan razes home and relocates thousands of squatters to camps" would classify Sudan as the source, "raze" as the verb, and no target; the WEIS code is "Force". Such a coding would be of little use. A new coding system, however, could include "squatter" in the actor file, (as well as other relevant nouns such as "informal sector," "slum," "urban poor," etc.). The new verb file would include "relocate" (as well as other verbs such as "evict") in a newly identified coding category of "Move". This system would then identify Sudan as the source, squatters as the target, and "Move" as the code.

Developing a new coding system obviously takes some time. Source material must initially be studied for relevant actors, verbs, and verb phrases, and the definition of what constitutes an actor may need to be broadened. For example, in the context of studying urban problems, one might want to include noun phrases such as "informal economy," "urban unemployment," "population," etc. in the actor file. The coding system might include codes for "increase," "decrease," "worsen,", "improve," "assist," and "hinder," as well as relevant WEIS categories such as "agree," "request," and "comment." However, even when the coding categories change dramatically, much of the vocabulary developed for WEIS would be still be useful. For example, while the vocabulary describing conflictual actions is quite different in the international and domestic arenas, the vocabulary describing cooperative actions much as meetings, agreements, promises and requests is fairly similar. The PANDA coding dictionaries, which deal with a much more detailed coding system than WEIS, already contain a large number of phrases dealing with domestic political behavior.

An even broader conceptualization of machine assisted coding can aid the research in a different way. A Nexis search on a broad topic area can generate hundreds of pages of information, much of it irrelevant. KEDS can assist in filtering this because unlike Boolean searching, it employs sparse parsing to focus only on subjects, objects and verb phrases.

This additional filtering can save hours of research time. For example, if one is specifically interested only in stories that relate to population growth in the urban informal sectors of the world's cities, all nouns that identify these sectors can be included in the actor list, and all verb phrases related to growth in the verb list. KEDS will identify all of the sources and verbs included in these files, when found together in the same source text, with or without targets. Whether or not the coding is accurate is irrelevant when the system is used as an identification mechanism. The researcher interested in "growth", for example, has simply to ask the system to identify those texts that contain the "growth" code, take the references, and refer back to the original articles for complete reports of the events. While this is a much less sophisticated use of technology than fully-automated machine coding of events data, it would be of substantial interest to anyone attempting to review the information contained in a plethora of news sources over an extended period of time.

In conclusion, it is our assessment that machine coding is a technology whose time has come. Machine-readable sources of natural language reports about political and social activities provide a tremendous resource for the systematic study of human behavior, but tools are needed to use those resources effectively. Based on our experience with KEDS, relatively simple programs running on inexpensive personal computers can be used to do a significant amount of processing of this type of data. We view KEDS as a first step, not the last word, and we would encourage others to work on these problems, as the potential rewards to social science research appear quite substantial.

While substantially faster than human coding, KEDS is still not quite fast enough to allow interactive experimentation with event coding schemes. Ultimately one would like to have a system where a researcher could change event definitions see these coding changes reflected, within a minute or two, in a statistical summary of the data, much as one can currently experiment with numerical transformations and subsets in SPSS or SAS. With source text files edited to eliminate uncodeable records, faster microprocessors, inexpensive parallel processing using networks and coprocessors, and greater efficiency within the KEDS program itself, this objective is probably attainable within the next five years.

Endnotes

[*] Authors' Note: An earlier version of this paper was presented at the 1994 meetings of the International Studies Association. Address correspondance to Philip A. Schrodt, Department of Political Science, University of Kansas, Blake Hall, Lawrence, KS 66045; 913-864-3523; fax 913-864-5208; Email SCHRODT@UKANVM.CC.UKANS.EDU. This research was funded in part by the National Science Foundation through grants SES89-10738 and SES90-25130 and the University of Kansas General Research Fund Grant 3500-X0-0038. Brad Bennett, Doug Bond, Ronald Francisco, Deborah Gerner, Tony Nownes, Jon Pevehouse, Julia Pitner, and Scott Savaiano have served as alpha- and beta-testers of KEDS and we are indebted to them for their feedback and patience.

[1] The advantages and disadvantages of events data have been extensively discussed elsewhere: see for example Andriole & Hopple 1984; Azar & Ben-Dak 1975; Burgess & Lawton 1972; Daly & Andriole, 1980; *International Studies Quarterly* 1983; Laurance 1990; McGowan et al, 1988; Merritt, Muncaster & Zinnes 1994; Munton, 1978; Schrodt 1994, Schrodt forthcoming.

[2] The lead is the first sentence of a news report and usually summarizes the contents of the report.

[3] PANDA codes an agent for the source and target, distinguishing for example between police and demonstrators within a country; the location where the event occurred, and the issue, domain and context of the event. KEDS is substantially less accurate on coding the latter four variables than in coding the standard variables, but it does a fairly good job with agents.

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

[5] We anticipate producing a Windows version of the program when the Macintosh version has stabilized. A copy of the program, its manual in Microsoft *Word* format, the actor and verb dictionaries, simulated data and coded event data for the Middle East 1981-1994 are available from Schrodt; we also anticipate depositing the coded data with the Inter-University Consortium for Political and Social Research.

[6] In an early version of KEDS, this generality extended to the other languages: in Gerner et al 1994 we report on experiments done in 1992 using KEDS on German language source material. Since that time, KEDS has been significantly enhanced to deal with additional English features and we've not kept up with the German development. Nonetheless, because KEDS is a sparse parser and needs to deal with only a small number syntactical structures (e.g. distinguishing the <adjective><noun> ordering of Japanese, English and German from the <noun><adjective> ordering of Arabic and the Romance languages) the program would probably be relatively easy to modify to handle other languages.

[7] The exception to this occurred when a verb root could be mistaken for a noun—for example FIRE—in which case multiple forms of the verb had to be entered explicitly.

[8] Pronoun references are very problematic because they often cannot be solved on a purely syntactic basis. In the sentence "Baker will meet with Mubarak when he goes to Geneva" the pronoun HE could refer to either BAKER or MUBARAK depending on who is going to Geneva. Sophisticated parsers (and humans) can often use semantic information to resolve references. In the sentence "John took the cake home and ate it", IT refers to CAKE because one does not eat HOME, whereas in "John took the baseball bat and broke it", IT refers to BAT rather than BASEBALL. KEDS does not use semantic information, but in most Reuters leads this is not required.

[9] For example, "Egyptian President Mubarak headed for a meeting with..."

[10] Given our focus on the Middle East, U.S. basketball star Michael Jordan was also problematic, particularly when we experimented with the U.S.-based United Press International news agency in addition to Reuters.

[11] For machines without a color display, a menu option displays the words and their types in tabular form in a dialog window.

[12] The editing dialog has to manage the entire list of verbs and actors, so modifications using these dialogs is somewhat slow; when an extensive set of changes must be made (e.g. creating a new

coding category using existing vocabulary) it is faster to edit the ASCII versions of the dictionaries with a word processor. We found it useful periodically to edit the actor and verb files to eliminate phrases that had been superseded by more general phrases and to check the organization of the verb patterns

[13] In an experiment where dictionaries were optimized for the coding of a single day, PANDA achieved a 91.7% machine coding accuracy; this probably represents the upper limit of accuracy for Reuters leads and a program using KEDS's sparse parsing approach (Bond, Bennett & Vogele 1994:9). These results are consistent with our sense that our current dictionaries for WEIS (a simpler coding scheme than PANDA) code Middle Eastern events at about 90% accuracy. 90% accuracy is lower than that of a single motivated coder working for a short period of time (for example a graduate student coding the 1990-91 Iraq-Kuwait crisis) but higher than the average reported intercoder reliabilities of about 85% in multiple-coder projects (Burgess & Lawton 1972).

[14] Additional problems occur because of ambiguities in the WEIS coding scheme; these are discussed in a longer version of this paper, Schrodt & Davis (1994), which is available from Schrodt.

[15] Reuters does not, to our knowledge, edit its reports to facilitate machine processing, but such editing is a possibility. International finance and trading firms figure prominently as customers of Reuters and these companies often use computerized filtering systems to automatically route news stories. The existence of such editing is conjecture on our part.

[16] This is also true of the Romance languages but is hardly universal: for example German uses a subject-object-verb order and Arabic uses verb-subject-object. We would anticipate that this basic ordering would be a crucial parameter in any future version of KEDS designed for coding multiple languages.

[17] We experimented with a general purpose pattern for passive voice in the German version, where it worked quite well.

[18] Articles—"a...force", "an attack"—often signal that the word is used as a noun, though at present KEDS ignores articles.

[19] As we note in Gerner et al 1994, there are numerous parsers and natural language understanding systems that are capably of doing much more sophisticated coding and classification than KEDS in *limited* behavioral domains and/or with *pre-edited* text, but that is not the problem we want to solve. KEDS is designed to take *unedited* Reuters leads covering a very *wide range* of political behavior.

Bibliography

Andriole, S.J., & Hopple, G.W. 1984. The rise and fall of events data: from basic research to applied use in the U.S. Department of Defense. *International Interactions 11,* 293-309.

Azar, E. E. 1982. *The Codebook of the Conflict and Peace Data Bank (COPDAB).* College Park, MD: Center for International Development, University of Maryland.

Azar, E.E., & Ben-Dak, J.,eds. 1975. *Theory and Practice of Events Research.* New York: Gordon and Breach.

Bond, D., Bennett, B. & Vogele, W. 1994. Data development and interaction events analysis using KEDS/PANDA: an interim report. Paper presented at the International Studies Association, Washington.

Burgess, P.M., & Lawton, R.W. 1972. *Indicators of International Behavior: An Assessment of Events Data Research.* Beverly Hills: Sage Publications.

Daly, J.A., & Andriole, S.J. 1980. The use of events/interaction research by the intelligence community. *Policy Sciences 12,*215-236.

Gerner, D.J., Schrodt, P.A., Francisco, R., & Weddle, J. L. 1994. The analysis of political events using machine coded data. *International Studies Quarterly 38,*91-119.

International Studies Quarterly. 1983. Symposium: events data collections. *International Studies Quarterly* 27.

Laurence, E. J. 1990. Events data and policy analysis. *Policy Sciences 23,* 111-132.

Lehnert, W., & Sundheim, B. 1991. A performance evaluation of text analysis. *AI Magazine 12*, 81-94.

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

McClelland, C.A. 1967. Event-interaction analysis in the setting of quantitative international relations research. Photocopy. Los Angeles: Department of Political Science, University of Southern California.

McGowan, P., Starr, H., Hower,G., Merritt, R.L., & Zinnes, D.A.. 1988. International data as a national resource. *International Interactions 14*,101-113.

Merritt, R.L., Muncaster, R.G., & Zinnes, D.A., eds. 1994. *Management of International Events: DDIR Phase II.* Ann Arbor: University of Michigan Press.

Munton. D. 1978. *Measuring International Behavior: Public Sources, Events and Validity.* Dalhousie University: Centre for Foreign Policy Studies

Random House. 1975. *The Random House College Dictionary*, rev ed. New York: Random House.

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

Schrodt, P.A. 1994. Statistical characteristics of event data. *International Interactions 20*,35-53.

Schrodt, P.A. 1994. "Event data in foreign policy analysis" in *Foreign Policy Analysis: Continuity and Change,* Haney, P. J. Neack,L., & Hey, J. A. K. eds. New York: Prentice-Hall.

Schrodt, P.A., & Gerner, D.J. 1994. Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. forthcoming, *American Journal of Political Science 38*

Schrodt, P.A., & Davis, S.G. 1994. "Techniques and troubles in the machine coding of international event data." Papers presented at the International Studies Association, Washington.

Table 1

Examples of Reuters Leads

July 17, 1990: Iraq President Saddam Hussein launched an attack on Kuwait and the United Arab Emirates (UAE) Tuesday, charging they had conspired with the United States to depress world oil prices through overproduction.

July 23, 1990: Iraqi newspapers denounced Kuwait's foreign minister as a U.S. agent Monday, pouring oil on the flames of a Persian Gulf crisis Arab leaders are struggling to stifle with a flurry of diplomacy.

July 25, 1990: Iraq has given Egypt assurances that it would not attack Kuwait in their current dispute over oil and territory, Arab diplomats said Wednesday.

July 31, 1990: Iraq has concentrated nearly 100,000 troops close to the Kuwaiti border, more than triple the number reported a week ago, the Washington Post said in its Tuesday editions.

August 2, 1990: Iraq invaded Kuwait, ousted its leaders and set up a pro-Baghdad government Thursday in a lightning pre-dawn strike that sent oil prices soaring and world leaders scrambling to douse the flames of war in the strategic Persian Gulf.

Source: Reuters

Table 2

Size of Phrase Dictionaries

	Kansas	PANDA
Actors	671	884
Verb phrases	3702	4292
Agents	not used	194

Authors' Addresses and Biographical Sketches

Philip A. Schrodt
Department of Political Science
University of Kansas
Blake Hall
Lawrence, KS 66045
phone: 913-864-3523 fax: 913-864-5208
Email: schrodt@ukanvm.cc.ukans.edu

Philip A. Schrodt is professor of political science at the University of Kansas; his research focuses on applications of formal methods to the study of international politics.

Shannon G. Davis
Department of Political Science
University of Kansas
Blake Hall
Lawrence, KS 66045
phone: 913-864-3523
Email: shandav@ukanvm.cc.ukans.edu

Shannon G. Davis is a doctoral candidate in political science at the University of Kansas; her dissertation focuses on small-group decision-making during the 1982-85 United States intervention in Lebanon.

Judith L. Weddle
202 N. Ridgeland Avenue
Oak Park, IL 60302
phone: 708-383-1433
Email: none

Judith L. Weddle is a doctoral candidate in political science at the University of Kansas; her dissertation focuses on the evolution of collective political institutions in marginalized urban housing.