Machine Coding of Events Data

Philip A. Schrodt and Christopher Donald
Department of Political Science
University of Kansas
Lawrence, KS 66045
913-864-3523

schrodt@ukanvm.bitnet
donald@ukanvm.bitnet

Version 1.1

April 1990

Paper presented at the International Studies Association meetings
Washington DC April 1990

This research is supported by the National Science Foundation Grant SES89-10738. Our thanks to Cory McGinnis for programming and Fritz Snyder of the University of Kansas Law School Library for assistance with NEXIS.

Abstract

This paper reports work on the development of a machine coding system for generating events data. The project is in three parts. First, the NEXIS database system is discussed as a source of events data. NEXIS provides a convenient source of essentially real-time events from a number of international sources. Programs are developed which reformat these data and assign dates, actor and target codes. Second, a system using a machine-learning and statistical techniques for natural language processing is described and tested. The system is automatically "trained" using examples of the natural language text which includes the codes, and then generates a statistical scheme for classifying unknown cases. Using the English language text from the FBIS Index, ICPSR WEIS and IPPRC WEIS sets, the system shows about 90%- 95% accuracy on its training set and around 40% - 60% accuracy on a validation set. Third, an alternative system using pattern-based coding is described and tested on NEXIS. In this system, a set of 500 rules is capable of coding NEXIS events with about 70% - 80% accuracy.

1.0. Introduction

In December 1989 and early January 1990, a group at the University of Kansas contributed a series of proposals to the DDIR Events Data project[1] which proposed developing software for coding events data directly from natural language (e.g. English) summaries of events. The original proposal was met with a combination of enthusiasm and skepticism. The enthusiasm stemmed from two sources. First, machine coding would allow the creation of a single English language[2] chronology of international events to which various event coding schemes such as WEIS, COPDAB and BCOW could be subsequently added. Second, the combination of machine coding and machine-readable news sources opened the possibility of dramatically reduced costs for generating events data. The existing events coding projects are highly labor intensive; this has meant that data are a number of years out of date by the time they became available to the research community.

The skepticism generally dealt with the issue of whether machine coding was possible, and if it was possible, whether it could be done with low-cost personal computers or whether it required more specialized equipment such as LISP workstations or supercomputers. There was an additional concern that event data coding was a sufficiently subtle exercise that machine couldn't—or shouldn't—do coding at all.

At the time of the DDIR proposal, we had been slowly working on the development of the system called WINR in conjunction with a National Science Foundation project entitled "Short Term Forecasting of International Events using Pattern Recognition". The focus of the NSF project is sequence recognition rather than machine coding, but since it required real-time events data on the Middle East, and existing events data sets ended over a decade ago, we needed to generate our own data.

In January, we obtained access to a machine-readable data source on Middle East events by downloading the "WIRES" file from Mead Data Central's NEXIS data service. With this source available, we felt that the most effective argument in favor of machine coding using conventional personal computers would be to create such a system. In February and March we accelerated the development of WINR, and an alternative, pattern-based system, KEDS, to the point where they could actually be tested on large data sets. This work implemented only the core algorithms without various filters and exception-handling routines, and is necessarily still tentative. We nonetheless feel it provides a useful lower bound on what can be done with machine coding.

This paper covers four topics. First, the general characteristics of the statistical analysis of natural language are discussed and the use of statistical, rather than linguistic, approaches is justified. Second, NEXIS is discussed as an events data source. Finally, two event coding systems, WINR and KEDS, are described and test.

---

[1] Data Development in International Relations, a collaborative data generating project funded by the National Science Foundation under the directorship of Dina A. Zinnes and Richard Merritt of the University of Illinois.

[2] Given the status of English as the dominant international language of commerce and science, we assume English would be the natural language of choice in any DDIR effort. Most of our techniques should work with other languages with minor modifications.

2.0. Statistical versus Linguistic Methods in Natural Language Processing

2.1. Statistical Processing of Natural Language

Within political science machine coding tends to be regarded as impossible; in fact it is a relatively straightforward computational problem for which a very substantial literature exists. The incorrect perception is due in large part to two factors. First, the initial efforts on the related problem of automated content analysis, using the General Inquirer program (Stone et al, 1966), were largely a failure, at least in political science.[3] General Inquirer was an effort of thirty years ago which ran on hardware possessing a fraction of the computational power of a Nintendo game: both hardware and theory have improved substantially. Second, machine coding is usually confused with the much more difficult problem of general natural language comprehension and machine translation. These are more complicated problems because they require mapping a domain with a complex structure (natural language text) into a range with a complex structure (another natural language or a knowledge base), whereas the machine coding problem maps into a very simple range (nominal codes).

The relevant literature for machine coding is that of automated information retrieval. Salton (1989) provides an excellent discussion of the general field of automated text processing with an extensive bibliography; Forsyth and Rada (1986) deal with the issue from a machine learning perspective. While this field is relatively new—taking off in the 1980s with the reduced cost of storage and subsequent increase in machine-readable text—it has been the focus of extensive work in computer science and library science. There are a variety of proven methods available; the work proposed here is merely adapts existing methodologies rather than inventing new ones.

Fundamentally, machine coding is a classification problem: the English description of each event must be associated with a specific category. Most approaches use variations on clustering and "nearest neighbor" algorithms: each individual event is described by a vector based on the words in the sentence describing the event; each category is associated with a point in that vector space[4], and classification is made by assigning the event to the category which it is closest to according to some metric (usually not Euclidean). Salton (1989: Chapter 10) describes a variety of methods for doing this; most are taken from the automated document retrieval literature (i.e. given a document, find all similar documents).

Converting this approach to a machine learning problem is fairly straightforward. One starts with a very large and representative set of cases (natural language events along with their associated event codes) and constructs—using either statistical or machine learning techniques—a vector space and a metric which maximizes the clustering of the events; each category is then assigned to the centroid of its cluster. Once the vector space, metric and centroids have been obtained, new cases are classified using the nearest neighbor method. The

---

[3] General Inquirer is still in use and has its share of advocates, but has never played a central role in international relations research.

[4] The dimensions of the vector spaces in question are usually associated with specific words or terms, and hence the dimensionality is extremely large: the use of one or two thousand dimensions is not uncommon. Conceptually, however, the system is similar to Euclidean nearest neighbor systems such as discriminant analysis.

method is conceptually identical to the technique by which discriminant analysis is used to predict unknown cases on the basis of a sample of known cases.

These methods are fundamentally statistical in nature and in that respect differ from linguistic approaches such as the Relatus system of Duffy and Mallery (1986). The advantage of the statistical approach is that it uses fairly straightforward machine learning methods, requires no pre-processing of information and will work in languages other than English; the disadvantage is that it is not sensitive to linguistic nuance and presupposes a great deal of regularity in the original text. Text describing political events—for example international events reported by the Foreign Broadcast Information Service (FBIS) or the international press—has a relatively limited vocabulary and should be fairly regular. The statistical approach is the more conservative technique: as Salton (1989: Chapter 11) points out, statistical and pattern recognition approaches are used in lieu of linguistic analysis in almost all existing document retrieval systems; Fan (1985; 1989) has also been using statistical methods for doing content analysis of wire service stories in studying public opinion.

2.2. A Typology of Machine Coding Schemes

Machine coding systems can be characterized on at least two dimensions. The first is whether they are primarily statistical or primarily linguistic, the distinction discussed above. The second is whether they are rule-based (deductive) or example-based (inductive).

There is no strict differentiation on either dimension: they are continua. Systems which are fundamentally statistical can still use some linguistic information. The likely order of increasing linguistic complexity would be:

- Pure statistical analysis
- Stoplists of common, irrelevant words. These can be ascertained statistically using sufficiently large samples, as discussed below, but are more efficiently provided externally since standard stoplists are readily available.
- Identification of parts of speech: the key distinction would be distinguishing verbs and proper nouns
- Simple syntactic parsing for compound subjects and phrases, direct objects and so forth
- Full English syntactic parsing
- "Deep structure" analysis to identify equivalent statement with very different syntactic structures: this is the apex of linguistic analysis and is still unsolved in its most general form.

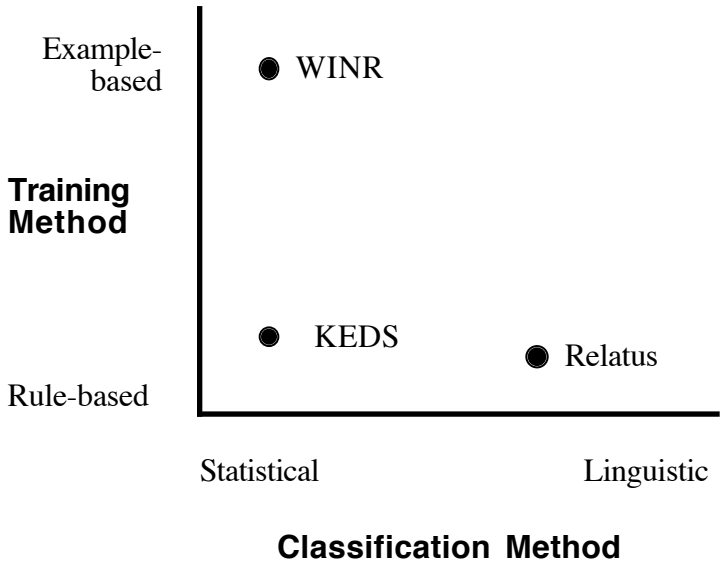
As noted earlier, Relatus tends toward the linguistic end of this continuum; WINR towards the statistical, though neither is a pure system.

The issue of rule-based versus example-based system is one of training. Example-based systems—which are based theoretically in the machine-learning literature of artificial intelligence (AI)—derive classification information directly from examples of the cases to be classified. Pure statistical systems are almost invariably example-based, but other machine-learning methods such as neural networks, genetic algorithms, and ID3 would fall into this category. The primary advantage of this approach is that no technical expertise is required of the

user: the machine extracts the relevant information directly from examples. This is also a disadvantage, there is no guarantee that the user actually has a logically consistent scheme[5]; the accuracy of example-based systems also tends to converge very slowly, following a classical learning curve, so a very large number of cases may be required for training. WINR is an example-based system.

In rule-based systems, which are based theoretically in the expert systems literature of AI, the rules for classification are provided externally by the user. The disadvantage of this approach is that the user must solve the classification problem first—which requires expertise and may be quite time-consuming in complex problems—and the computer simply implements that solution. The advantage is that the human usually has a good idea of the actual classification rules, particularly in the case of artificial classification schemes such as WEIS or COPDAB designed for rule-based implementation. The pattern-based approach, KEDS, is an example of a rule-based system.

As illustrated in the figure below, both of the systems discussed in this paper are towards the statistical end—though they incorporate some linguistic information in the form of stop lists and the identification of some proper nouns. As noted in the later discussion, further development of these systems would probably move them more in the linguistic direction; the issue of example versus rule-based training is still open.



2.2.2. Language

The basic principles of statistical and pattern-based information retrieval are largely independent of language. In principle, a system like WINR would work with Spanish, Russian, Arabic, Chinese or Japanese provided it had an appropriate tokenizing or stemming algorithm for

[5] If the user's examples are not logically consistent (more precisely, consistent with the underlying information structures of the machine learning system) this will be reflected by a failure of the learning curve to converge even when a large number of examples have been processed.

reducing verbs to their roots, a lexicon, and a suitable number of examples. Constructing these would not be trivial and would require some linguistic expertise[6] but might be worthwhile in situations where a language other than English would substantially augment regional coverage: the Arab Middle East, Spanish Latin America and Francophone Africa come to mind. Machine coding of additional languages requires standardized machine transliteration of non-Roman writing systems. For alphabetic systems such as Russian, Arabic and Japanese kana this is not difficult; ideographic systems such as Chinese, Korean and Japanese kanji are still problematic.

2.2.3. Quantity of Text

Up to a point, classification should become easier rather than more difficult as the length of the text increases, since a longer text will provide a greater number of features identifying the nature of the event. The point of diminishing returns is reached when the text incorporates multiple events: for example, if the leader of nation A protests an action by nation B and then launches into a long discourse on the prior friendly relations between A and B, that additional information will most probably lead to misclassification unless it is parsed to identify compound phrases within the text.

Because of the "pyramid style" of journalistic writing—the most important aspects of a story are given first, followed by the details—wire service stories are particularly useful input for machine classification systems. Fan (1989) uses wire service copy to study the effects of the media on public opinion; the document-retrieval system demonstrated by Thinking Machines Inc at the 1988 DDIR Events Data conference in Cambridge (which used a nearest neighbor system working in parallel) also analyzed Reuters wire service copy. Political speeches and editorials, in contrast, would be much more difficult, and become nearly impossible when they contain complex literary allusions, conditional phrases and so forth. Linguistic systems such as Relatus are designed to work with more complex text, but to our knowledge these have never been demonstrated as being able to operate in a machine-coding mode.

2.3. Validation

A machine coding system uses three levels of validation:

- *Within the training set.* The first test is whether the system can use its algorithm to correctly classify the cases on which it was trained. If it cannot, then the information structure of the algorithm itself is insufficient to classify the cases and a new approach needs to be used.
- *Split-sample tests.* The second test is against new cases which the system has not seen before. Failure at this point usually means that the training cases were not representative of the case universe; with a representative training set the split-sample accuracy should differ little from the training set accuracy.
- *Against human coders.* This test is actually redundant, since the cases in the split-sample

[6] It is quite likely that stemming algorithms and stop word lists have already been developed by the linguistic or library science communities for languages such as Spanish, Arabic and Chinese; English is actually one of the more difficult languages in this regard. It might be possible to use automated index construction techniques (see Salton, 1989: Chapter 9) to produce much of the lexicon, which would further reduce the labor involved.

test were also coded by humans. However, if the total case universe is not known when the system is first trained (e.g. if coding is being done in real time) this should be done periodically to insure that the system is still properly dealing with any new vocabulary or phrasing.

3.0. Events Data Analysis, Statistical Analysis, and Historical Analysis

The issue here, to borrow Hubert Dreyfus' phrase, is "what computers can't [and can] do". Events data collections have developed rather divorced from the creation of statistical techniques by which to analyze those data, a problem compounded by the tremendous time and expense required to collect those data.[7] While substantial effort has been devoted to the problem of measurement (for example, in the construction of the COPDAB scales), very little effort was spent figuring out what would be done with those measures.

The measurement problem is anything but trivial: there is no point in collecting data for statistical analysis which is more accurate than the statistical technique can use. To use a familiar analogy, there is no reason to create a drawing which is accurate to 1/300 inch if it is going to be printed on a dot-matrix printer at a resolution of 1/72 inch: the additional precision is lost.

Most existing general events data sets have internal intercoder reliability in the range of 80% to 90%,. Interproject reliability is less clear though Burgiss and Lawton (1972) report a 42% agreement at the discrete event level between the original WEIS and student coders at Macalester College; and Vincent (1983) notes that the correlation (r) between WEIS and COPDAB conflict scores (by year across countries) ranges from 0.92 in 1969 to 0.14 in 1972, with an average of 0.66. This is not a precise business: human coding of events data contains a lot of uncertainty.

The most likely measurement tradeoff is quantity versus quality: given finite resources, the greater the effort devoted to the reliability of the coding, the less the quantity of the data coded. Since, the most common statistical manipulation in events data is aggregation, the quantity-quality trade-off can be analyzed explicitly.

Assume—as is typical—that the events data are used to estimate a value, ô, (e.g. an hostility score) by taking the mean of a sample of observations:

$$\hat{o} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Assume that each xi can be decomposed into three parts

$$x_i = \mu + e_i + m_i$$

where μ = true value; e_i = true error (in other words, deviation due to the intrinsic variability of the underlying random variable) and m_i = measurement error. Assuming e_i and m_i have mean zero,[8] from the Central Limit Theorem, we know that ô will be distributed

[7] See Schrodt (1988a) and Laurence (1988) for further discussions

[8] A weak restriction since any known bias can be adjusted by subtracting a constant.

$$\text{Normal}\left(\mu\ ;\ \frac{\text{Var(e)} + \text{Var(m)}}{n}\right)$$

whatever the underlying distributions of e_i and m_i.

Suppose we have two measurement instruments A and B which have sample sizes N_a and N_b and measurement error variances v_a and v_b respectively. Assume that $N_a > N_b$ and $v_a > v_b$, in other words, A has a larger number of observations but also has a higher measurement error Let s_a and s_b be the error variances of the ô measured using A and B: under what circumstances will $s_a < s_b$?

Assuming without loss of generality that Var(e)=1, a bit of algebra will show that this is true provided

$$v_a < v_b\left(\frac{N_a}{N_b}\right) + \left(\frac{N_a}{N_b} - 1\right)$$

Since the second term is greater than zero, this means that $s_a<s_b$ so long as the variance of the less accurate measure increases proportionately to the increase in the sample size. For example, if method A provides twice the number of data points as method B, it can have at least twice the measurement error (Var(m)) and still produce more accurate (Var(ô)) measures of ô.

Reducing the variance of a sample mean is not the only item of interest[9] in events data analysis, though it probably is of interest in most existing applications. Our point is simply that the issue is open. In a world of finite resources—the world with which we are best acquainted—the quantity-quality tradeoff is real. For certain analyses, the reliability of human coded data will provide an advantage over the superior quantity of machine coded data. However, this cannot be taken for granted, particularly given the dramatic cost advantages of machine coding. We estimate that on a per-event basis, machine-coded data generated from machine-readable sources costs roughly 1/1000th as much as human coded data generated from hard-copy source. That data certainly does not have 1,000-times the error variance. When the alternative to machine-generated data is no data, the quantity-quality tradeoff must be carefully considered.

The related point on historical analysis concerns the level of detail required in events data. Events data are just that, data, something intended for statistical analysis. Events data will not replace books, chronologies or newspapers, nor are they intended to. The situation of CASCON versus BCOW is a case in point. CASCON, strictly speaking, is not a data set in the behavioral sense, because it is not intended to be analyzed using statistical methods. It is a form of hypertext designed to aid the memory of a human decision-maker. BCOW, in contrast, is designed for statistical analysis.

There comes a point where the background knowledge required to "understand" a set of data exceeds the capacities of either computer algorithms or statistical techniques. When that point is reached, one leaves the realm of the scientific and enters that of the humanistic. There is nothing

[9] For example, in some pattern recognition studies, the superior ability of a human to code discrete secondary events buried in a story might be a more important consideration.

wrong with this—after all, many behavioralist researchers teach humanistically at the undergraduate level. The humanistic approach, building on the massive experiential data base possessed by any socially functioning human being, will provide more complex generalization than a scientific model but at the price of incurring greater error and omitting a logical and explicit representation of those generalizations.

Much the same tradeoff exists with respect to statistical versus linguistic processing of natural language. Language comprehension is arguably the single most complex cognitive activity of human beings, and the fact that progress in computational linguistics has proceeded at a snail's pace should come as no surprise.[10] At the same time, natural language is a complex but highly regular human artifact, so statistical techniques which have proven useful in understanding other regular human artifacts might be expected to work with language. A computer is not human, and has neither human experience nor the specialized human "wetware" used in language analysis, and therefore it should not be at all surprising that a computer might best understand language through methods other than linguistics. Statistical methods provide one such alternative; pattern-recognition another.

Problems which can be solved by machines constitute only a fraction of all political problems, and an even smaller fraction of interesting political problems. This is hardly novel: machines have made dramatic inroads in replacing bank tellers; they have done little to replace short-order cooks.

4.0. The NEXIS Data Source

NEXIS is the on-line news data base of Mead Data Central.[11] While NEXIS contains over 100 different files, most of these are specific to narrow economic topics. The file we have been working with is the "WIRES" file, which contains wire service reports from Reuters (predominantly), Xinhua ("New China") General News Agency, United Press International, the London Daily Telegraph and an assortment of other sources.

While NEXIS provides the full text of articles, the "citation" form provides a date and a short "lead" which provides the gist of the article. A typical Reuters "citation" is

```
4. Copyright (c) 1989 Reuters The Reuter Library Report, March
31, 1989, Friday, AM cycle, 224 words, ISRAEL SUMMONS CANADIAN
AMBASSADOR TO DISCUSS PLO. JERUSALEM, March 31, ISRAEL-CANADA,
LEAD: Israel has summoned the Canadian ambassador to protest
Canada's decision to upgrade talks with the Palestine
Liberation Organisation (PLO), a Foreign Ministry spokesman
said on Friday.
```

Additional examples of NEXIS citations are provided in Figures 1 and 2.

---

[10] Theories of computational linguistics are available in abundance; practical applications are not, and to the extent that computational approaches to natural language have succeeded in practical applications, these have almost invariably been statistical rather than linguistic. This is not to say that a linguistic approach to this problem is impossible, but it is to say that until such an approach can be demonstrated, a great deal of skepticism is justified.

[11] At the University of Kansas, NEXIS is available during off-peak hours through the Law School subscription to the LEXIS legal service. The university pays a fixed subscription price for the service, rather than paying by citation; for faculty research purposes, it is effectively free.

NEXIS is searched using keywords, which can be arranged into Boolean statements of considerable complexity. For our project, a simple search command

```
HEADLINE(ISRAEL OR JORDAN OR EGYPT OR LEBANON OR SYRIA OR PLO
OR PALEST!)
```

is used to extract the records. This looks only at the headline of the article, and looks for any of the targets; the "PALEST!" construct contains a "wildcard" character which will match "palestinian", "palestinians" and "palestine".[12]

4.1. Density

One complaint about existing events data sets is the number of events per day. While the Middle East may not be completely typical for NEXIS, the density in this area seems quite high: it is consistently about 700 reports per month during the 1988-1989 period. In contrast, the ICPSR WEIS data set generated from the New York Times has about 700 events per month for the entire world during 1968-1978 (McClelland 1983:172); this is also roughly consistent with COPDAB's 400 events per month density for 1948-1978. The NEXIS density is even higher when one considers that many NEXIS reports would generate two or more individual WEIS or COPDAB events due to double coding, so a 1000 events per month density is probably in the ballpark. The WIRES is updated at least once a day, so the data are current to within about 24-hours. Existing events data sets, in contrast, are about twelve years out of date.

The approximate density of interactions within the Middle East is given by the following two tables. These are the result of machine coding NEXIS for actors and are probably not entirely accurate, though they will be close since the system had a fairly complete list of Middle Eastern actors.

1988

	ISR	PAL	LEB	JOR	UAR	SYR
ISR	--	1002	179	12	45	13
PAL	934	--	66	103	48	59
LEB	34	28	--	10	1	58
JOR	78	163	2	--	28	24
UAR	109	66	4	34	--	7
SYR	21	68	86	10	9	--

Total 3301

[12] This retrieves only a small number of irrelevant stories, primarily UPI stories on basketball player Michael Jordan and Reuters soccer, tennis and cricket results.

1989

	ISR	PAL	LEB	JOR	UAR	SYR
ISR	--	737	206	20	102	36
PAL	686	--	43	34	62	25
LEB	30	7	--	3	6	108
JOR	46	44	10	--	15	12
UAR	137	50	14	15	--	30
SYR	23	24	175	4	17	--

Total 2721

Since this period includes both the Palestinian intifada and the Lebanese civil war, this density is not necessarily typical for the entire world, but it is promising.[13] It should also be noted that these are counts of news stories—with most duplicates eliminated—and many news stories would generate multiple events: we guess that the event count directly comparable to WEIS or COPDAB would be about 50% higher.

The chart below shows the distribution of event counts for the Egypt-Israel dyad in COPDAB. The frequency of events varies substantially over time, rising to about 150 events per year during periods of military crisis (1955-1957; 1967-1973) and dropping off to fewer than 25 in other periods prior to 1974. After 1974 the level is about 75 events per year. If the 150% ratio of stories to events is accurate and 1988-89 has Israel-Egypt interaction levels comparable to 1974-1978 (which is at least plausible), then NEXIS is recording about twice as many events as COPDAB in the Middle East.

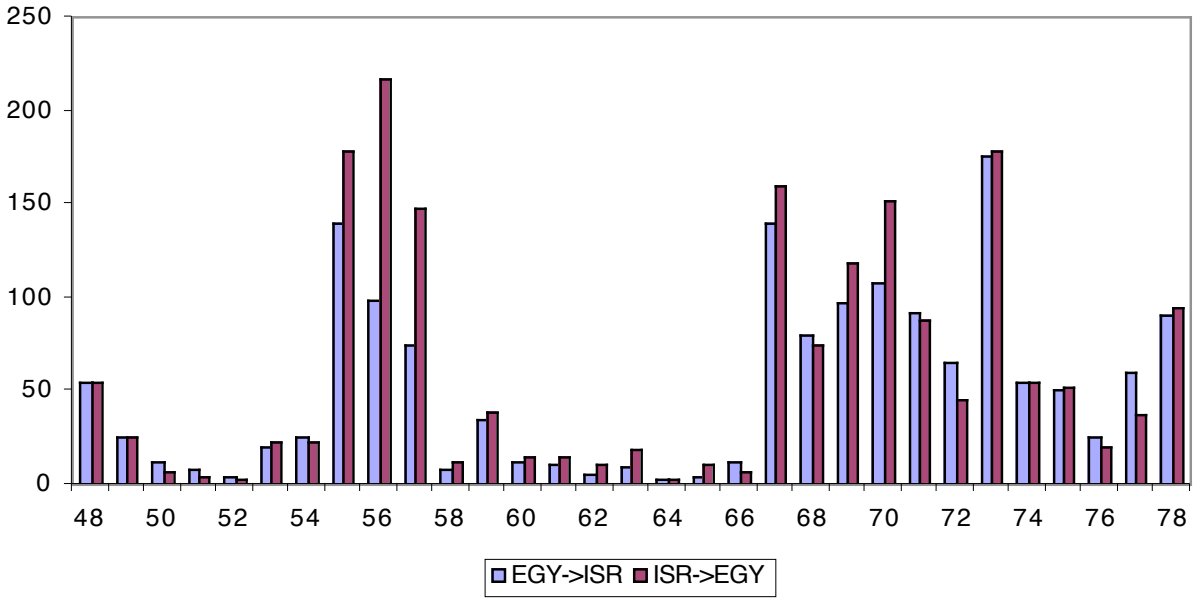
A wide variety of additional sources—for example the *New York Times, Wall Street Journal, Washington Post, Los Angeles Times, Times of London*, TASS, and a variety of Japanese news services—are available on NEXIS and if, as has been suggested, these newspapers provide complementary coverage of the world rather than overlapping coverage, these would potentially provide substantially enhanced coverage. However, this comes with three costs. First, the downloading time increases. Second, many of these "new" sources primarily use Reuters or UPI material; the apparently complementary coverage simply reflects different editorial selections.[14] Finally, searching the *NYT* and *LA Times* for Middle Eastern targets results in a very large number of false positives—as much as half the data—since Jordan is a very common surname in the United States. These can be weeded out, but increase the downloading time, particularly during basketball season.

Frequency of COPDAB Interactions by Year

---

[13] Note that coverage does not dramatically drop during 1989, in contrast to the New York Times international coverage, which focused almost exclusively on Eastern Europe and Panama during the autumn of 1989.

[14] This is true even for papers with reporters in the area: the herd mentality of the international press is well documented and the Reuters teletype is prominently available in the major international hotels in Middle Eastern capitals. Since most international events are either unpredictable acts of violence only occasionally witnessed firsthand by Western reporters, or are official statements released at press conferences, the presence on the ground of reporters will contribute little in the way of new events, however useful their presence may be for analysis.



A final conceptual problem in using wire service data is the large number of unofficial "conditional" reports—"a reliable source reports…", "the government of X is said to…" and so forth. Whether or not these constitute "events" is an issue which can only be resolved in the theoretical context of a specific problem: for some analyses this may simple constitute background noise which is better eliminated; in other applications (for example those dealing with perceptions) it may be very important. In the Middle East, particularly Lebanon, we may be getting an unrepresentatively high number of such reports, but they will occasionally occur in any politically active area.

5.0. Coding Data, Actor and Target from NEXIS Records

NEXIS is downloaded using a modem and an 800-number; the service operates at 1200 baud so this is slow[15] (a 1000-record file is about 500K, and takes about an hour and a half to download), but it is easily automated with the appropriate software.[16] Mead Data Central explicitly permits downloading and states that data can be held for 30 days and used for research purposes.

The NEXIS data comes off the wire looking like Figure 1. Reformatting this into something useful can be done with a relatively simple program, since the data is very regular in format despite all of the garbage in the file.

[15] It may also operate at 2400 but we haven't explored this.

[16] We used the Red Ryder (now called White Knight) software for the Macintosh, which contains a flexible programming language. Crosstalk is a comparable program for MS-DOS machines. The automated routines downloaded the data in the early morning hours without human supervision; only a couple of minutes of setup time were required.

5.1. Dates

Pulling the date out of the record is straightforward; in the examples in Figure 2, the date is the article date, but one could also get the story dateline, which is probably a more accurate.indication of when the event occurred. Because these are newswire stories, the story usually occurs on the same day as the event itself.

5.2. Actor and Target Identification

Actor and target coding initially appeared problematic but turns out to be relatively simple. The text is scanned using a list of potential actors. The first phrase encountered which is on that list is used to assign the actor code; the second phrase encountered which is different from the first is considered the target. This rule allows the system to cope with identifying both ISRAEL and SHAMIR as ISR actors while not coding "Israeli Prime Minister Shamir accused Jordan" as an ISR/ISR interaction. If a target distinct from the actor cannot be found, the story is assumed to be an internal event and is written to a reject file.

This system is obviously not foolproof but it seems to work in about 70% to 90% of the cases, both with respect distinguishing internal from external events, and in correctly assigning actor and target codes. The most common error, as would be expected, is transposition, and this is most likely to occur in UPI stories, which tend to use more complicated sentence structures than Reuters or Xinhua. The system we currently are using does not deal with compound subjects or objects (e.g. "Egyptian President Mubarak will visit Senegal and Mauritania next week"); we are currently working on some simple English parsing routines to correct this.

Automated actor assignment is particularly useful if one is dealing with internal actors, since these are substantially greater in number than international actors, and often change quickly over time. The same techniques used for assigning international codes work for internal actors, and we would suggest that any effort dealing with internal actors would be better served with natural-language data recoded for actor as the research dictates, rather than trying to derive a single internal actor coding scheme. In the Middle East, for example, some research would require coding all Palestinian groups as PLO, others would distinguish between parties within the PLO (e.g. Fatah, PFLP, DFLP), others only between parties participating in the Palestine National Council and those outside of it (e.g. Abu Nidal). These changes can be made very quickly using machine coding; similarly it is very easy to add new actors, change their names or constrain them in time.

5.3. Duplicates

Duplicate stories are very common on NEXIS, particularly for Reuters. Duplicates occur when Reuters issues an update to an earlier story, and can be identified by leads which are similar, but not identical to the original lead. On some days up to 40% of the stories will be duplicates.

We are currently detecting duplicates by computing a "signature" on each lead consisting of the count of the frequency of each letter in the text.[17] If the signatures of two leads on the same

[17] This undoubtedly seems like a strange approach to analyzing text but it is common in the statistical language processing literature.

day differ by less than a fixed threshold—currently set at 20—they are considered to be duplicates and only the most recent version of the story is coded.

This method captures the most obvious duplicates, for example those leads which are identical, differ only by a misspelling or word order, or which have added a couple of words. It also results in virtually no incorrect rejections; in fact the threshold may be too low. The method does not detect duplicates where substantial new information has been added to the lead (for example an additional sentence) and does not deal with the same event reported in two sources (e.g. Reuters and Xinhua).

Additional work is needed in this area. One approach would be to use the current system to do a first pass, code those events and then do additional duplicate filtering within sets of events which have the same actor, target and code. For example, a simple filter might allow only a single event with a given actor, target and code combination on any single day; this will result in excessive elimination in a few categories (notably 'Force') but would probably be relatively harmless in most. The nature of diplomatic discourse is such that the issuance of multiple, distinct warnings, accusations, denials, grants and so forth to a single target in the course of one day is unlikely, and if two such events are recorded (particularly in separate sources) they are probably duplicates.

5.4. Symmetrical events

In the existing WEIS and COPDAB data set, a large number of events are symmetrical or double-coded: for example a meeting between Egypt and Algeria is coded both as EGY 03 ALG and ALG 03 EGY; a firefight between Israel and Syria is coded as both ISR 22 SYR and SYR 22 ISR. In these situations, a single story leads to two events with transposed actors and targets.

To date we have made no attempt to deal with this situation, or to deal with multiple events in the same story. Symmetrical events are probably best dealt with at the level of coding, since symmetrical action seems possible in only four of the WEIS categories—'Consult', 'Agree', 'Reduce Relationship' and 'Force'. Multiple events in a single lead would require some English parsing. Fortunately, most of the NEXIS stories deal with only a single action—multiple events in a single day become separate stories—so this is a relatively minor problem.

6.0. The WINR Machine Coding System

We have experimented with two machine coding systems. WINR[18]— an almost pure machine-learning, statistical system—was our initial attempt and the one with which we've done the most experimentation. It is a useful illustration of how a machine learning system might work, and might even be useful in some specialized circumstances, such as coding-by-example in non-English source material. Both systems generate WEIS codes but should in principle be able to code COPDAB, and possibly BCOW.[19] For reasons discussed below, most of our current development focuses on a pattern-based alternative to WINR, but that system incorporates some

[18] A recursive acronym for "WINR Is Not Relatus".

[19] BCOW is more difficult because it codes the continuation of cessation of events as well as their initiation. This characteristic, rather than the greater number of BCOW categories, might cause problems.

of WINR's characteristics.[20]

6.1. Programming

Natural language processing has acquired a quite inappropriate mystique of being suitable only for specialized workstations such as $30,000 LISP machines and Kurzweil document processors. In fact, it is well within the domain of common, upper-end personal computers[21] such as the Macintosh II series or IBM AT and PS/2 series provided suitable programs are used. This section will briefly indicate how this can be done; an earlier version of this paper or the source code itself (both are available, "as is", from the authors) provides greater technical detail.

Our project works in both the MS-DOS and Macintoshes environments; we avoid mainframes. Most of the work reported here was done on a circa-1988 Mac II. The programs are Pascal: Turbo Pascal 5.0 on the MS-DOS machines and TML Pascal II on the Macs. Both systems have optimizing compilers and include extensions to allow bit-level manipulation, direct addressing and other register-level programming, so these Pascals provide most of the advantages of C while being more readable.

In both WINR and KEDS, the bulk of the computational time—outside of I/O (input-output)—is spent searching for individual words. The appropriate structure for such a search is obviously a binary tree, which reduces (on average) the number of terms which must be compared to log2(N) where N is number of words in the system.[22] Thus a system needing to search a base of 16000 words would require on average only about 14 comparisons. By further optimization, for example by packing tokens into 32-words and using 32-bit comparison instructions rather than character strings, even this computation can be significantly speeded up.

While the English language contains several hundred thousand words, the vocabulary used in the international news media is substantially more restricted, and the vocabulary required for

---

[20] In other words, WINR was something of a deadend, but it was an instructive deadend.

[21] At the risk of beating a dead horse—but a horse which seems to have many lives, as it keeps cropping up—a point should be made concerning the issue of hardware. It is well known that the "personal" computer of today has computing resources comparable to a university mainframe in the early 1970s, and has vastly superior software (e.g. operating system, compilers and debuggers). The "neccessity" of LISP machines, RISC workstations and supercomputers in political science research is comparable to the "necessity" of BMWs among yuppies. While it is doubtlessly more fun to drive to the grocery in a BMW than in a Chevy Nova, the Chevy is quite sufficient for that task. Using a LISP machine on most political science problems is comparable to renting a backhoe to dig a flower bed. It is quicker, cleaner and cheaper to use a shovel.

Compounding this problem, interpreted languages such as LISP and Prolog, the dynamic scoping found in many object-oriented languages, and complex multi-tasking, networked environments such as UNIX are tremendously expensive in terms of machine cycles. They are wonderful general-purpose development tools, but in a specialized production environment, they provide performance akin to downhill skiing in snowshoes. There ain't no such thing as a free lunch, and at the core of those fancy operating systems and languages, the CPU is still running machine code. The higher the level of implementation language is from that code, the less efficient the program. UNIX and LISP: Just say No.

[22] This holds for a balanced tree where all words are equally probable. Neither case holds in this problem: the tree is built by taking words as they are encountered, which unbalances it but means that more frequent words are likely to be higher in the tree. As is well known, word frequencies generally follow a rank-size law distribution (Zipf's Law); the imbalance of the tree and skewed distribution of the words probably about cancel out so the log2(N) approximation is still valid.

classification even more so. WINR's analysis of 8000 IPPRC WEIS descriptions required a vocabulary of fewer than 5,000 words, well within the constraints of a modest personal computer.

WINR is quite fast: on a Macintosh II it processed about 100 cases per second. Fully half of this time is disk I/O, rather than computation. The computationally intensive work is identifying tokens and either updating (training) or calculating (classification) the category vectors. The system could be speeded up considerably through parallel processing; farming out the search problem to individual processors on either a token-by-token or case-by-case basis, with a central processor handling the I/O and aggregating the vectors.

We are currently experimenting using parallel processing with T800 transputers, and also intend to work with a group of networked Macintoshes using the "Mac-Cube" software developed at Cal Tech. As a rough guess, we should pick up about a 50% increase per additional processor using Mac-Cube (communication delays limit additional gains)—that is, 1 additional processor would give 1.5 times the processing speed, 2 would give twice, etc —so with a network of, say, 11 Macintoshes we could process about 600 cases per second.[23] The entire 30-year ICPSR COPDAB has about 150,000 cases, and at these speeds could be recoded in 500 seconds, or about 8 minutes.[24] Even an unassisted Mac II could recode all of COPDAB in less than an hour.

This is not to suggest recoding COPDAB in eight minutes is a useful exercise in political analysis: it is merely to suggest that existing hardware is more than up to this task.provided it is carefully programmed. Our WINR software required about 90 seconds to analyze a set of 8000 IPPRC WEIS cases.[25] If a more elaborate and accurate system is slower by a factor of 10—which frankly is difficult to imagine unless one is being completely oblivious to conserving machine cycles[26]— then one could still recode the whole of COPDAB in an overnight run, or experiment with a 1000-case subset in runs of about two minutes using off-the-shelf equipment. The constraint is software, not hardware.

Any machine capable of doing serious statistical analysis or desktop publishing is sufficient for the task of machine coding. If classifying ten events per second is considered an acceptable speed for a machine coding system—which would allow 430,000 events to be coded in a 12-hour (i.e. overnight) run—then one has about 200,000 machine instructions available to code each event. If one cannot classify an event using 200,000 instructions, then one should be looking for

---

[23] In other words, a modest-sized departmental computer lab is a workable parallel computer. The Cal Tech system uses standard AppleTalk wiring, so if you've got a networked Macintosh data lab, you've got a small parallel computer. The Mac-Cube code is available from Cal Tech.

The T800 transputers are even faster, running at 10 MIPS with virtually no operating system overhead. Based on early experiments they should classify at about five times the speed of the unassisted Mac II. A small farm of 4 T800s should be able to classify about 2000 cases per second, though the system may become I/O bound prior to that point. The T800s are not cheap—about $1500 a processor, including 1 Mb RAM—though on a per MIPS basis they are about 100 times cheaper than a supercomputer…

[24] WINR requires two passes through the data, hence the apparently doubled time.

[25] Each run involved training on a 2000-case set, reclassifying the training set, and two classification runs on 3000-case sets. These did not involve retokenizing.

[26] For example using LISP…

a better algorithm, not faster hardware.

6.2. The Algorithm

The classification metric used in WINR is similar to that used in fuzzy sets; this extends work done earlier in Schrodt and Leibsohn (1985). With each word w we associate a vector v_w indexed on the event codes (c) where

$$v_{wc} = \frac{\text{\# of occurences of w in events of type c}}{\text{total occurrences of w}}$$

In other words, v_{wc} is simply the conditional probability that w is found in an event of type c given that w occurs in a sentence. If, for example, we were using a set of 22 codes (e.g. the 2-digit WEIS), the maximum value of v_{wc} would be 1.0 (which would occur if a word were only found in events having a single code); the minimum value would be 0.045 (1/22, which occurs if the occurrences of the word are spread equally across all of the categories). These vectors are determined empirically from the frequency of tokenized words in the English text in the training set.

The initial classification criterion is the code c which maximizes

$$\underset{c \in C}{\text{Max}} \sum_{w \in S} v_{wc}$$

where C is the set of event codes and S is the set of tokenized words found in the text describing the event. In other words, the system simply sums the conditional probabilities and takes the maximum value.

WINR EXAMPLE

```
Source text:    "Mubarak Receives Message From Mitterand"
Tokenized words: RECEIV  MESSAG
WEIS Codes           01    02    03    04    05    06    07   ...  22
Vectors:  RECEIV    .00   .25   .30   .00   .02   .10   .33   ... .00
          MESSAG    .00   .30   .30   .20   .10   .10   .00   ... .00
Total               .00   .55   .60   .20   .12   .20   .33   ... .00
Classification: WEIS code 03
```

In this example, there are two key words in the source sentence, "receives" and "message". Each word is found in sentences in various WEIS codes—code 02 is "comment", 03 is "consult", 04 is "approve", 05 is "promise" and 06 is "reward". These probabilities were determined empirically from the actual data. When the probabilities are summed, the maximum category is 03, "consult".

WINR implements this basic algorithm with a few minor modifications:

1. Elimination of stopwords on an entropy criterion

The information theoretic concept of entropy (see Pierce 1980) is ideally suited to a

classification problem such as this. This measure—identical to the H_{rel} used in some of McClelland's early studies—is effectively a measure of the ability of a feature (in this case, a word) is discriminate between categories. A word with zero entropy is associated with only a single category (in other words, it classifies perfectly); one with high entropy is scattered more or less equally across the various categories. Entropy is defined as

$$E = \sum_i p_i \log_2(p_i)$$

where p_i = proportion of times the word is observed in category i. In the IPPRC WEIS, "low" entropy tends to be anything below 0.5; "high" entropy above 1.5.

High entropy words were eliminated in a two-stage process. First, the data were tokenized using a "universal" stoplist of about forty common words (e.g. numbers, days of the week) and proper nouns (actors). The entropy of the remaining words was then computed and all those words above a threshold (set somewhat arbitrarily at 1.5) were added to the stoplist. The final stoplist contained about 700 words for the IPPRC WEIS set.

The elimination of high entropy words decidedly speeds up the computation; interestingly it has only a small, though positive, effect on the classification accuracy. The sample of the high-entropy words from the IPPRC WEIS and their associated entropies are given below:

JOHN	2.335	BOTH	2.025	CALL	1.534	BANK	2.406
ARMS	2.644	BEEN	2.013	BACK	2.032	ARMY	1.887
AREA	1.899	ANTI	1.561	BOMB	1.609	BUSH	1.561
DAYS	2.026	DOES	1.712	CUBA	2.395	COSTA	1.609
CUTS	1.560	CAMPAI	1.609	DROP	1.609	FUEL	1.561
FIRE	2.372	FULL	2.398	GERMAN	2.534	AFGHAN	2.339
AFRICA	2.623	CHANCE	2.271	AGENCY	2.045	NATO	2.518
MEET	1.839	ATTACK	2.170	AIMED	1.609	COMMAN	1.792
DECLAR	1.979	EFFECT	2.146	ESTABL	1.793	GENSCH	1.561
ABANDO	1.609	EXERCI	1.561	DIRECT	1.885	COULD	1.588
ARMED	1.970	GOOD	1.609	HAIG	2.190	HIGH	2.164
HOLD	1.834	GULF	1.847	HELD	2.397	HELP	1.923
HEAD	1.561	INDIA	2.344	INDIAN	1.550	GUARAN	1.677

For the most part these are common verbs (e.g. CALL, DOES, COULD, BEEN), proper nouns (AFRICA, JOHN, BUSH, CUBA, COSTA, NATO, INDIA) or common improper nouns (ARMY, DAYS, FUEL). The presence of verbs such as MEET, BOMB, DECLAR and ABANDO might be unexpected as one would anticipate those to be strongly associated with specific categories. For example, MEET should be strongly associated with the "Consult" category (03) or BOMB with "Force" (22). However, they aren't: there are quite a few common

verbs which are frequent but spread quite evenly across the categories.[27]

The creation of a stoplist by entropy calculations has the obvious weakness that if a particular actor (i.e. noun) is strongly associated with a type of action—which might be true simply because the actor appears infrequently—that noun will not be picked up through an entropy test. However, this does not seem to be a problem in the actual tests—for example the entropy test correctly picked up the names of major individuals active in the Falklands/Malvinas conflict (e.g. Galtieri, Thatcher, Haig, Reagan, Weinberger, Pope John Paul) even though the Falklands data was a relatively small percentage of the data.

2. Keywords

A few of the WEIS categories are strongly associated with single categories. These were detected, using machine learning methods, by looking for high-frequency words with exceptionally low entropy. The IPPRC WEIS produced the following set of words with entropy less than 0.5 and a frequency greater than 10; the two-digit numbers are the WEIS categories they are most strongly associated with:

```
03   'MEETS '     05   'ASSURE'     08   'AGREE '
09   'CALLS '     10   'URGES '     11   'REJECT'
13   'PROTES'     14   'DENIES'     15   'DEMAND'
16   'WARNS '     20   'EXPELS'
```

This set corresponds quite closely to the keywords in the various WEIS categories; these were the only words with high frequency and low entropy.

3. Tokenizing

Most natural language processing systems have a means of reducing verbs to their roots, a process known as "stemming."[28] WINR (and KEDS) uses a cheap and dirty stemming technique: words were truncated to six characters,[29] which we refer to as "tokens". This works relatively well for English: for example REJECTS, REJECTED, REJECTING all go to REJECT. In cases where it does not work—for example SHOT, SHOOTS, SHOOTING—separate counts are maintained. This approach is simpler and computationally more efficient than a system with greater information on rules of English verb formation (e.g. Lovins, 1968) but results in relatively few incorrect roots because of the limited vocabulary used in describing international events. A formal stemming algorithm would probably improve the performance of the system by a couple percentage points, but not dramatically.

The tokenizing process also automatically eliminated any word of three or fewer characters.

---

[27] Most problematic: SAYS. One would expect SAYS to be strongly associated with the "Comment" category but it has one of the highest entropies of any verb and is therefore largely useless.

[28] Tokenizing to root words is done for the purposes of computational efficiency; the actual classification algorithm will work just as well with complete words. Lovins (1968) provides a much more complete system for the derivation of English "stems" which we may eventually incorporate into KEDS; van Rijsberger (1979) also discusses stop word and stemming systems in detail.

[29] Six is not a magic number; due to a programming error we inadvertently did a couple of runs with 5-character tokens and this did not noticeably degrade the performance.

This eliminates many common prepositions (e.g. TO, BY, AT), some forms of "to be" and "to have" (e.g. IS, ARE, HAS, HAD), and most of the abbreviations used in the ICPSR WEIS (e.g. PM, FM, PRS and all of the country codes). This deletion probably had no negative effects on the performance of the system.

The IPPRC and ICPSR WEIS descriptions reduce from anywhere between one and more than a dozen tokens; there is no obvious pattern to this. The 8000-case IPPRC WEIS contains about 4000 unique tokens.

4. Elimination of singletons

A potential problem occurs with words which are very infrequent in the training set, in particular words which occur only once and therefore are strongly associated with a single category; we refer to these as "singletons". These enhance the internal consistency of the test but result in "overtraining" because if the same unusual word occurs in a different context it will probably result in misclassification.

The solution is just to eliminate the use of any low-frequency word occurring in only a single category. In the tests reported here, "low frequency" was <2. The elimination of singletons reduces the internal consistency of the test from 95% to 90% in the IPPRC WEIS, but has little effect in the FBIS or ICPSR WEIS sets; it raises the external accuracy by about 3%. This may be too low a threshold, and one might eliminate any low-frequency cases where there is an equal distribution among categories.[30]

7.0. Experimental Coding of WEIS Summaries using WINR

The system described in the DDIR proposal was tested on four different data sources: a FBIS index, ICPSR WEIS, IPPRC WEIS and NEXIS. This system was done in entirely a machine-learning mode except for the introduction of a few words such as common four+-letter words (THAT, AMONG) and a list of actors borrowed from the actor-target assignment program.

7.1. Data Sets

Four data sets, of increasing linguistic complexity, were processed. These were subjected to an assortment of presumably harmless transformations, such as converting to capital letters, eliminating all punctuation, and truncating the description to 255 characters if necessary. Duplicate descriptions were discarded if, when tokenized, they were identical, so most symmetric interactions are represented only once in the data.

In the two WEIS sets, an attempt was made to get equal representation of all of the categories. For the ICPSR WEIS, fifty samples of each type were taken (if available); for the IPPRC WEIS 100 samples. The remaining cases were used as the validation set; this is somewhat problematic because the validation set is nonrandomly depleted in the less-frequent categories.

Since the objective of this work is developmental and the ultimate objective is coding NEXIS,

---

[30] For example a word occuring twice, once in 02 and once in 12. These are not captured by the high entropy measure since, strictly speaking, they have dramatically reduced the uncertainty, from 22 categories to only 2. Still, this configuration provides less confidence than a word which occurs 10 times in 02 and ten in 12, even though the entropy is the same.

not WEIS, these data sets are a convenience sample rather than a random sample. The benchmark data set, on which most of the actual development was done, is the IPPRC WEIS.

FBIS Index:

About 450 entries from the Newsbank Inc. Index to the FBIS Daily Reports[31] from the Middle East in April, 1989 were coded. Because the FBIS categories are very abbreviated, the four-letter tokenizing threshold was not used; the standard stoplist was used. The only high entropy word in the set was 'TO'. The training set was the first half of the data; the validation set the second half.

ICPSR WEIS:

1200 cases taken from the standard ICPSR WEIS data. The cases have both the actor and target in the Middle East and are taken from an assortment of years. The ICPSR WEIS descriptions are quite short and generally are in an abbreviated English.

IPPRC WEIS:

This is the [in]famous "fugitive" WEIS was collected by the International Public Policy Research Center for Richard Beal's National Security Council event data project of during the early years of the Reagan administration. The source is the New York Times, and project was directed by former McClelland students, so it is probably fairly close to "true" WEIS.[32] The descriptions are substantially longer than the IPPRC WEIS and appear to be similar to NYT leads. The sample is about 8000 events from all actor (no Middle East selection) for 1982, with duplicates eliminated. Roughly 2000 events are used in the training set; and two validation sets of 3000 events each were tested. The second validation set is relatively undepeleted in events, except for the very rare ones, and this is reported in the tables.

7.2. Results

The overall results are reported in the following tables. Each table presents a matrix giving the "true" two-digit WEIS code (row) against the category into which the description was classified (columns). "Acc" is the percentage of correct assignments[33]; N is the number of cases in each category. The "Total" accuracy is the total for the table; "Unknown" is the number of

---

[31] This publication is available in hardcopy or microfiche and can be found in most reference libraries. Recent indices are also available on a CD-ROM. Unfortunately, Readex Inc has been less than enthusiastic about the prospect of using this information for machine coding purposes—in distinct contrast to Mead Data Central—and it is not possible to download the information directly. We entered this information by optically scanning the hard copy, a relatively time-consuming and error-prone process in comparision to NEXIS.

[32] Hey, we got the tape from Lew Howell and this is all we know about it. That's why the data set is called fugitive, right?

[33] This is a simple percentage and is not directly comparable to Scott's pi (see Krippendorf 1980, chapter 12), the measure of intercoder agreement commonly used in measuring intercoder reliability which adjusts for classifications correct by chance. With the 22 coding categories of WEIS, the simple accuracy and Scott's pi measures are similar: accuracy is greater than Scott's pi by about 5% to 10%, depending on the exact distribution of categories, for the ICPSR and IPPRC WEIS data.

cases which could not be classified because of unknown words or singletons.

The overall accuracy of the WINR system is summarized below

	Internal	External	Learning
FBIS Index	96%	51%	64%
ICPSR WEIS	94%	33%	64%
IPPRC WEIS	90%	43%	62%

The internal accuracy of data sets is comparable; the external accuracy differs considerably on the simple external test, and then converges again when learning is incorporated. Unsurprisingly, the FBIS index entries have the highest accuracy since these are both very abbreviated and almost entirely comments and consultations.

The "Internal Consistency Test" is the result of recoding the training set after the classification matrix had been determined: it is basically a measure of the extent to which that matrix has incorporated the information in the training set itself. The results are reported below for the IPPRC WEIS; from purposes of brevity only ten WEIS categories are reported, though these ten constitute 83% of all events in the ICPSR WEIS (McClelland, 1983).

IPPRC WEIS INTERNAL CONSISTENCY TEST

Code	Acc	N	02	03	04	08	09	10	11	12	14	22
02	0.800	100	80	7	..	2	1	..	..	1	1	1
03	0.959	100	..	96	1	..	..	..	..	..	..	..
04	0.980	100	..	..	98	..	..	..	..	..	..	..
08	0.910	100	1	1	..	91	..	..	1	1	..	..
09	0.730	100	1	..	1	4	73	3	1	1	1	..
10	0.930	100	..	..	..	1	..	93	..	..	..	1
11	0.880	100	..	..	2	1	1	2	88	..	..	..
12	0.910	100	..	1	..	1	..	..	..	91	2	1
14	0.990	100	..	..	..	..	..	..	..	..	99	..
22	0.860	100	2	..	1	4	..	..	..	1	1	86

Total 0.900 1984 Unknown 0.000 0

. The "External Validation Test" is the test against the remaining cases that were not in the training set. The tables below show this for each of the three data sets.

FBIS EXTERNAL VALIDATION TEST

Code	Acc	N	02	03	04	08	09	10	11	12	14	22
02	0.703	54	38	5	2	..	..	2	..	1	..	5
03	0.602	73	15	44	13	1	..	..	..	..	..	..
04	0.200	5	3	1	1	..	..	..	..	..	..	..
08	0.333	9	3	..	..	3	3	..	..	..	..	..
09	1.000	1	..	..	..	..	1	..	..	..	..	..
10	0.666	3	1	..	..	..	..	2	..	..	..	..
11	0.250	4	2	..	..	..	1	..	1	..	..	..
12	0.200	10	6	1	..	..	..	1	..	2	..	..
14	0.428	7	4	..	..	..	..	..	..	..	3	..
22	0.571	14	3	1	..	..	..	1	..	..	..	8

Total 0.507 213 Unknown 0.093 20

ICPSR WEIS EXTERNAL VALIDATION TEST

Code	Acc	N	02	03	04	08	09	10	11	12	14	22
02	0.172	110	19	9	5	7	1	6	2	3	6	3
03	0.777	36	..	28	1	3	..	..	..	..	1	..
04	0.000	0	..	..	..	..	..	..	..	..	..	..
08	0.000	0	..	..	..	..	..	..	..	..	..	..
09	0.000	0	..	..	..	..	..	..	..	..	..	..
10	0.000	0	..	..	..	..	..	..	..	..	..	..
11	0.513	37	4	3	1	1	2	2	19	1	..	..
12	0.285	438	32	12	22	21	6	20	23	125	62	13
14	0.882	17	..	..	..	..	..	..	1	..	15	..
22	0.327	394	35	4	3	4	1	12	1	7	97	129

Total 0.334 1048 Unknown 0.023 25

IPPRC WEIS EXTERNAL VALIDATION TEST

Code	Acc	N	02	03	04	08	09	10	11	12	14	22
02	0.096	458	44	8	45	34	23	24	11	25	18	5
03	0.679	458	4	311	2	5	8	3	1	5	3	1
04	0.445	74	3	1	33	2	..	..	..	3	3	1
08	0.696	56	2	..	5	39	1	1	..	..	2	..
09	0.158	107	8	3	9	4	17	10	1	9	2	3
10	0.514	68	3	2	2	..	5	35	..	1	2	..
11	0.584	113	3	2	1	4	1	1	66	2	7	3
12	0.365	400	19	6	21	8	13	10	17	146	20	5
14	0.939	33	..	..	..	..	..	..	..	1	31	..
22	0.189	58	1	..	3	12	1	..	..	4	4	11

Total 0.433 2052 Unknown 0.036 74

The low external accuracy of the ICPSR data may be due to the fact that the training set depleted the validation set for many of the categories, and almost half of the validation set was a single category, 'Accuse'. The obvious downfall in the IPPRC set is in the 'Comment' category — less than 10% of these are categorized correctly, and comments are about 25% of the total data. Generally there is only a very weak pattern in the incorrect categorizations of comments.

7.3. Variations

In the process of developing the system, several variations on the basic protocol were studied; those which worked are briefly reported here. Unless otherwise specified, all of these experiments used the IPPRC data.

7.3.1. Iterative Learning

The natural extension to a simple training/validation protocol is allowing additional feedback when a case is misclassified. This approach is typical of many machine learning systems, and has the advantage that additional information is only added to the data base when it is needed, providing a corrective to “overtraining” on the training set.

In the “iterative learning” results, whenever a case is misclassified, the information on that case is added to the distribution matrix. Unsurprisingly, this helps the overall accuracy considerably, bringing it to around 63% for all of the cases. The 63% includes the initial misclassification as an “error”; when the validation sets are retested after iterative learning phase (which will then presumably correctly classify most of the previously misclassified cases, though singletons still cause some to be unclassified), the accuracy goes to about 80%, roughly the inter-coder reliability for human coders. Iterative learning dramatically improves the accuracy in the ‘Comment’ category on the IPPRC set, which would suggest that comments were being misclassified in part because of different word combinations.

7.3.2. Near Misses

While WEIS is technically categorical, it is frequently used to generate either binary or interval data, so it is useful whether the errors made are random or are clustered around the true value. In general, as is obvious from the table, they are not strongly clustered, but widening the error limits provides some additional accuracy

Validation Set Accuracy

+ 1	49.3%
+ 2	54.4%
+ 3	60.2%

7.3.3. Word Location Tests

One would expect that most of the information required to classify a case is in the first few words, and more generally the key word is frequently the first verb. Since the stoplist removes a large number of the nouns, that verb is usually the first or second token. Therefore some experiments were done which distinguished between a token as first word, second word and anywhere else[34], and also looking at only the first four tokens.

Neither experiment made any systematic difference. Words with location information improve the internal accuracy somewhat when singletons were not excluded, but then lost accuracy on the external validation test. Considering only 4 tokens had about +1% of the accuracy of considering all tokens, which eliminates the possibility that words later in a statement (which tend to be tangential to the main aspect of the event) are introducing errors which are not present in the initial word. If processing time were a critical limitation (which it isn't) it could be reduced by only looking at a limited number of tokens.

7.3.4. Alternative Classification Metrics

We did some limited experimentation with some alternatives to the proportionality metric. Of particular interest was seeing whether the entropy measure itself could be used as a method of weighting. Two metrics were studied:

$$p_{wi} = \frac{N_{wi}}{N_w(e_w+1)} \qquad (1)$$

and

$$pwi = \frac{N_{wi}}{e_w+1} \qquad (2)$$

where N_{wi} = frequency of token W category i, N_w = total frequency of W and e_w = entropy of the distribution of p_{wi}. Both of these have the effect of weighting the effects of low-entropy words more strongly than high-entropy words.

---

[34] This was done by making the sixth letter lower-case for the first word; the fifth for the second word, and otherwise doing nothing.

Metric (1) has no positive effect—in fact it seems to reduce the accuracy (with or without iterative training) about about 2%. Metric (2) looks rather promising—it raises the accuracy of the IPPRC validation test to 47% without learning and 64.5% with learning. The gain is probably due to the fact that measure (2) effectively weights words by their frequency and uses entropy to adjust for cases which are scattered across several categories, whereas the original measure used only proportions and consequently over-emphasized the impact of low frequency words.

7.3.5. NEXIS

As an experiment in the spirit of "and now for something completely different…", we tried the unreasonable exercise of coding NEXIS using the frequency matrix trained using the complete IPPRC WEIS. If IPPRC WEIS descriptions are in fact similar to New York Times leads, and if NYT leads are similar to Reuters, Xinhua and UPI leads, then there should be at least some correspondence between them. The NEXIS target set was Middle East stories from November 1989 through February 1990; codes were assigned to NEXIS events while training KEDS and are probably about 90% accurate.

There is no particular reason this should work, and it didn't: the agreement with NEXIS is only about 26%. The NEXIS set contains a number of "internal" events for which there would be nothing comparable in the IPPRC WEIS, but even eliminating this problem the agreement is well below 35%. The NEXIS set turns out to contain a substantially larger (or different) vocabulary than the IPPRC WEIS, so the two seem to be less comparable than would appear at first. The highest accuracies were on the categories which used keywords and these, rather than the complete matrix, probably account for most of the correct classifications.

7.4. Discussion

WINR is in many ways an extreme test. It is almost a purely statistical, example-based scheme—the program had no linguistic information except for the initial stoplist. The natural language input was completely uncontrolled—neither the vocabulary nor syntax were restricted.[35] The FBIS and ICPSR sources are obviously very abbreviated forms of English, and even the IPPRC categories are probably influenced by the WEIS categories, but it is still relatively unprocessed data. Finally, this is a very simple classification system, and nowhere close to the state of the art. In particular, there was no attempt to estimate optimal weighting vectors for the various words[36], nor was there a sophisticated attempt to generate an optimal training set. For all of these reasons, this should be interpreted as a lower bound on the accuracy of machine coding, not as either an upper limit or typical performance.

Several general conclusions can be reached from this exercise. The basic representational scheme is clearly sufficient, since the internal consistency is greater than 90%. The performance drops to around 40% on a simple external validation test; this can be increased to about 62% with iterative learning (80% accuracy on reevaluation of the entire set). The external validations are still about 20% below the within-project coder reliabilities in the existing projects, though

---

[35] In the case of the Xinhua General data, the sentences were not necessarily even grammatically correct, and more generally misspellings and garbled words were fairly common.

[36] For example using regression analysis or some comparable linear technique.

they may be approaching the between-project reliabilities using human coders.

It is not entirely clear how many events are required for the system to learn a category; that answer will be dependent on the category itself. For example, the 100 example per category protocol used on the IPPRC WEIS was sufficient to obtain accuracies greater than 85% on three categories ('Deny', 'Demand' and 'Warn') and around 70% on four others ('Consult', 'Promise', 'Agree', and 'Reduce Relationship'), but produced less than 20% accuracy on 'Comment' and 'Force'.

The improvements in these experiments have tended to be incremental and linear. Three changes to the basic scheme—keywords, elimination of singletons, and an entropy-weighted metric—each added between 3% and 4% to the external accuracy. There were no magic bullets to cause dramatic increases in accuracy; at the same time, the method does not seem to be hitting limits, and it is certainly closing in on the reliability of the large human-coded data sets. As noted earlier, we have ample machine resources—both speed and memory—for additional improvements before the system begins to slow. The two remaining dramatic improvements are probably to be found in optimal weightings and a system for parsing compound phrases.

That being said, WINR is probably not the way to go in the long run. Pure machine learning is a useful academic exercise since it is not language dependent which would provide an inexpensive means of coding non-English text. But for our purposes, where the NEXIS event source is in English, it makes sense to incorporate some linguistic knowledge.

8.0. A Pattern Based Approach to Coding NEXIS

An alternative approach to coding is to use human-designated patterns rather than statistical inference. The disadvantages of this approach is that a human must interact with the data and the patterns will be language-specific. The advantage is that the human coder will bring to bear linguistic knowledge and will probably reach a convergence on the common categories much more quickly.

The pattern-based program we have developed—KEDS[37]— is very simple and requires little storage space and only modest machine speed. In the training phase, the system is primarily constrained by the time required to read the candidate event, so even a very simple computer (e.g. IBM PC) would be sufficient. Since patterns require only a few bytes each for storage—probably around 16 on average—even a modest 640K machine could probably handle about 20,000 patterns. Search time, as before, is log2(N) so a 20,000 pattern set would require only about 15 or so comparisons on average, modest requirements for machines which can do about a half-million comparisons a second.

KEDS simply looks for target strings of tokenized words (the four-letter minimum no longer applying) with only a single generic operator, *n where n is an integer, which allows a pattern to skip words. So, for example, the pattern

LEFT *5 TODAY

would match "left today", "left Cairo today" and "left for Saudi Arabia today" but not "before he

[37] Kansas Events Data System. We are referring to the current system as KEDS-X (experimental) since we anticipate substantial improvements over the next few months.

left, the ambassador said that the meeting held today".

Based on the WINR experience, 'Comment' was left as a "default" category—if the system cannot match any of its phrases to the sentence, the case is assumed to be a comment. This has two advantages. First, it makes statistical sense to leave some category as the default, and the variety and complexity of comments makes them a good candidate.[38] Second, the comment category will contain almost all of the cases misclassified because they contained unknown phrases.

The '00' code is used to designate "Not an Event". In the training set, this disposes of sports reports and arguments within Israel concerning foreign policy (which contain references to other actors and are picked up as external events). This is a only a simple filter and is not intended to be a longterm solution to this problem.

8.2. Results

The results of the NEXIS test will be available at the ISA or on disk from the authors; a sample is attached to the paper.[39] We trained the system on NEXIS data from November 1989 through February 1990, a set of about 2000 events. The system was then "tested" on March, 1990 data. The original data set was coded to greater than 95% accuracy; both of us did the coding which took about six person-hours.

8.2.1. Training

The following observations are impressionistic but indicate some of the strengths and weaknesses of KEDS versus WINR.

As with WINR, a small number of keywords is sufficient to capture correctly most of the events in categories such as 'Deny', 'Accuse', 'Consult' and so forth. Unlike WEIS, the obvious tokens KILLED, WOUNDI and WOUNDE pick up a large percentage of the 'Force' events, which are rather prevalent in this region. There are a few few multiple-word phrases which have almost the same specificity as keywords, though there are not a lot of them.

If one ignores compound statements—which a simple parser could take care of—and internal events which could be eliminated with a better filter, there were very few false positive classifications in the training set. If the system actually finds a pattern, it will probably do the classification correctly in most instances; a formal test of this using the IPPRC WEIS is reported below. Human scanning of the default category—'Comment'—alone should be sufficient to get the error rate down to existing intra-project reliability levels. While we only coded for 2-digit WEIS categories, coding for 3-digit codes would be only slightly more difficult.

In the training phase, coding proceeds quite quickly, limited (on a Macintosh SE/20) only by the coder's reading speed and tolerance for eyestrain. The system quickly stabilizes at about an 80% or higher accuracy rate, so most of the "coding" is simply approving the system's selection,

---

[38] Ironically, McClelland (1983:172) reports "…the Comment and Consult categories were added as virtual afterthoughts during the designing of the category system." They also account for more than one-third of all events in the ICPSR WEIS dataset.

[39] The entire data file is about fifty pages long so we have not included it; we would be happy to send it in ASCII format to anyone interested; specify Macintosh or MS-DOS format.

which takes only a keystroke. We were coding at better than 200 events per hour, though it is not clear whether this could be done at a sustained rate or by less motivated coders.[40]

The 2000 event training set resulted in about 500 phrases; these were sufficient to code the entire data set to better than 95% accuracy; a sample of these phrases is provided in this paper.[41] We guess that the phrases follow a rank-size law in terms of their utilization; 20% of the phrases, plus the default category, probably account for about 80% of the classifications. We have done some limited optimization of the list—noting for example that the phrases LEFT FOR and LEFT *3 FOR will match the same phrase and eliminating the more specific—but have not done so systematically; this might reduce the size of the phrase list by about 10%.

KEDS is somewhat slower than WINR, coding about 20 events per second on a Macintosh SE in batch mode; this speed can probably be at least doubled through software optimization.

8.2.2 Testing KEDS

For the proof of the pudding, we downloaded NEXIS data for March, 1990 and coded it using a fully automated system. Downloading took about an hour; the reformating and actor code assignment about seven minutes; the event assignment about a minute. All of these processes ran unattended except for loading disks and changing a few file names.[42]

An additional test was done by using the KEDS system on 5000 cases of the IPPRC WEIS. This is a strong test in the sense that KEDS was trained solely on NEXIS and any correct classifications would have to be due to the natural language content shared between NEXIS and the IPPRC data plus any statistical commonality due to the journalistic sources.

The results of the classification are presented in the final table; the overall accuracy is about 38%, which is a considerable improvement on WINR's 26% accuracy on the NEXIS data, and overlaps the 35%-50% accuracy levels of WINR without iterative training. Unsurprisingly, the 'Comment' category is quite accurate (73%) in terms of the number of IPPRC 'Comments' correctly classified into the 02 category; interestingly 'Protest'(13) is even more accurate (76%) and 'Promise' (05) is fairly good (60%).

Since KEDS classifies as a comment anything it has not already seen, a measure of considerable interest is the number of false positives (the percentage errors measured by the columns of the table rather than the rows); this is given below.

---

[40] The ambiguity of this figure is due to our doing this coding amid assorted interruptions and also the fact that by the time the program was up and running, we'd already accumulated a dictionary of about 100 phrases, including many of the most important.

[41] We also have not checked the inter-coder reliability of Schrodt and Donald, which is unlikely to be in excess of 90% on the cases which the machine system is not already correctly classifying.

[42] Actually the entire process could be automated with a simple script at the operating system level. Our production system for "real time" data will automatically dial NEXIS, download data from the previous day, reformat and code it without human supervison.

Percentage of False Positives in IPPRC WEIS Classification

Code	%wrong	N	Code	%wrong	N	Code	%wrong	N
01	0.714	14	02	0.808	2854	03	0.257	700
04	0.308	107	05	0.351	174	06	0.522	23
07	0.000	1	08	0.588	34	09	0.717	60
10	0.756	213	11	0.120	83	12	0.057	300
13	0.288	73	14	1.000	2	15	0.213	47
16	1.000	1	17	0.656	32	18	0.500	2
19	1.000	4	20	---	0	21	0.788	113
22	0.464	153						

As would be expected, the 'Comments' category has a huge number of false positives: 81%, and over half of the data set. However, some of the categories are quite accurate, for example 'Accuse' (6% error ; N=300), 'Reject' (11% ; N=83), 'Demand' (21%, N=47) and 'Consult' (26%, N=700). If one were doing an analysis which looked only at these categories, the coding is already within a useable range of accuracy. KEDS may significantly undercount events of these categories—erroneously classifying them as comments—but those events which are actually reported to be in those categories will have been accurately counted.

8.3. Discussion and Enhancements

KEDS-X is only a prototype and one which has received very little training. Some known problems which could be solved in a more sophisticated system:

1. The system does not deal with either compound phrases, subjects or objects. Most such cases could be handled with a relatively simple English parser.
2. KEDS stops when with the first valid pattern is found. An alternative system would find all valid patterns and use some precedence rules to decide between them.
3. The only precedence rule is that longer phrases are given priority over shorter patterns. Some more sophisticated precedence rules might include
   - Negation
   - Conditional statements
4. The only filtering for "events" was the necessity of identifying distinct actors and targets, so there a fairly large number of non-events in the set.
5. There seems to be a sporadic bug in the search procedure itself, which we've yet to squash.

Our unsystematic training is evident from some obvious misses in the March 1990 test. For example, PROMIS is not a token for the 'Promise' category, so the system misses some easy classifications; the absence of the infinitives TO MEET and TO VISIT cause misclassification of several 'Consult' events. Accusations, demands, protests, agreements and uses of force are generally picked up accurately; some of the less frequent categories such as 'Threat' and 'Reward' are less accurate. True to Murphy's Law, early March features a political dispute

between Algeria and Egypt over a soccer match, which generates an assortment of incorrectly labeled non-events.

The reject list is almost perfect—we spotted only two stories which were incorrectly rejected and both of these should have been caught since the targets were on the actor list; a program bug is at work. Actor and target coding in the selected stories is generally workable but suffers from both compound subjects and the inability to pick up indirect objects. There are quite a few incorrectly selected stories which actually deal with internal Israeli politics, though March has been a particularly bad month for that.

Would we use this data for research? Perhaps. In lieu of hand-coded data, probably not. In lieu of no data, probably yes. The marginal cost of generating this data—after the sunk costs of developing the program and coding the training set—was about five minutes (at most) of human labor. Could we get more information on Middle East politics during March 1990 using five minutes of hand coding?—probably not.

Gerner (1990) uses this data in a study of Israel's reaction to adverse international opinion as reflected in the change in the number of deaths of Palestinians inflicted by Israeli forces in dealing with the intifada. The only WEIS categories in the study are the negative interactions 'Accuse', 'Protest', 'Demand' and 'Warn', which based on the IPPRC WEIS tests have low false positive classification rates. Gerner's test showed not only a statistically significant correlation between deaths and international protest, but also found that non-Arab accusations and demands had a significant negative effect on the change in deaths but that accusations and demands from Arab states had no significant effect. The ability of KEDS-generated coding to reflect such a relatively subtle political distinctions provides at least some indication of its validity.[43]

A critical difference between the KEDS test on NEXIS and the WINR analysis of the FBIS Index and WEIS sets is that NEXIS is totally free-form data: it is straight off the news wires with no editing prior to coding. The Xinhua General frequently is not grammatically correct English; the Reuters reports occasionally employ excessively florid language, such as "clashed" for "disagreed" and "blasted" for "criticized". While IPPRC and possibly even Readex may have been influenced by the WEIS coding category, Reuters, Xinhua General and UPI are not. This is natural language.

An additional advantage to KEDS is that the training is cumulative, transferable and explicit. Any additional training will pick up where the old training left off—subsequent effort adds to the "knowledge" already incorporated. The training undoubtedly follows a classical learning curve—in other words, shows diminishing returns—but it does not start from over from the beginning. At the end of our comparatively short training phase, we were occasionally running better than 90% accuracy on new events, so even this simple system is working fairly well.

The knowledge is transferable—the results of the Kansas training could be transferred to another institution by mailing a disk. Some of the phrases used to code the Middle East will undoubtedly prove inaccurate when used in other areas of the world, but that can be resolved through some retraining—most of the information will be relevant.

---

[43] To a human these distinctions are not subtle; by the standards of events data, where for example WEIS and COPDAB disagree on the direction of change of US-Soviet relations about 30% of the time (Howell, 1983), this is doing rather well…

Finally, the coding rules are explicit, which cannot be said of human coding no matter how elaborate the coding manuals and coder training. Some of our coding is questionable due to ambiguities in the WEIS scheme itself; our inter-coder ambiguities are also embedded in the phrase set. But all of these decisions are explicit and one could code data in 1999 using precisely the same coding rules and precisely the same interpretation of those rules as we used in 1990.[44] This allows a greater degree of reliability in the coding, which in many applications (notably those studying change) is more critical than validity.

9.0. Conclusion

The purpose of this project has not been to develop a state-of-the-art machine-coding system: it has been to demonstrate the possibility of a system which has been asserted to be impossible. It is a Wright Flyer; not a Boeing 747: we can't carry 400 people at 1000 km/h, but we can certainly get our machine off the ground and down again in one piece under its own power. To reiterate our major points:

- NEXIS data provides much greater event density than the *New York Times* and any of the WEIS or COPDAB sets we have looked at. The basic downloading and reformatting of NEXIS can be entirely automated; the major constraint is downloading time.
- The coding of actors and targets appears more straightforward than we anticipated; very simple rules seem to handle actor and target assignment with about 80%-90% accuracy and simple parsing would correct most of the remaining cases.
- The pure machine learning system proposed in WINR will handle IPPRC WEIS with 90+% internal consistency; and about 45% - 65% external consistency on 2-digit WEIS codes. We do not consider this to be acceptable as a final performance but it is certainly a good start for a prototypical system.
- A pattern-based system, KEDS, seems better suited for the NEXIS data; it runs with an accuracy of about 80% on NEXIS data. In a cross-check against the IPPRC WEIS, it performed at about the same level as external validation checks on WINR. KEDS also has very low false positive classification rates on some of the categories. The pattern-based system is particularly well-suited for machine-assisted coding.
- All of these programs run quite efficiently on upper-end personal computers such as the Macintosh II and IBM AT systems with a megabyte or so of memory; if speed is not a factor they will run on a basic 640K IMB PC-style machines. The Macintosh WINR implementation codes over 100 events per second; KEDS is substantially slower but has not been optimized. While some parallel processing hardware might be useful for experimental and development work, the basic software can be run on existing, relatively inexpensive equipment.

A basic software development scheme has been outlined throughout this paper through the identification of some key open issues in the existing system. Simple parsing and better filtering are the two key issues, and we are currently working on adding this feature to KEDS. In addition

[44] It also goes without saying that unlike work study students, the computer does not have midterms or term papers, leave early for vacation, become emotionally upset with its significant other, graduate, or require retraining every fall.

to this, a better "knowledge base" is needed in terms of actors and event patterns, and these two processes need to be integrated. The software as a whole could use additional optimization for both serial and parallel processing. We hope that within a few more months we will be able to get most of the WEIS categories into the general range of human inter-coder reliability (70% to 90%) and possibly shift to 3-digit WEIS codes, though our project currently requires only 2-digit codes. Additional results will probably be presented at the APSA.

Bibliography

Burgess, Philip M. and Raymond W. Lawton. 1972. *Indicators of International Behavior: An Assessment of Events Data Research*. Beverly Hills: Sage Publications.

Duffy, Gavan and John C. Mallery. 1986. "RELATUS: An Artificial Intelligence Tool for Natural Language Modeling." Paper presented at the International Studies Association, Anaheim.

Gerner, Deborah J. 1990. "Evolution of a Revolution: The Palestinian Uprising, 1987-1989." Paper presented at the International Studies Association, Washington.

Fan, David. 1985. "Lebanon, 1983-1984: Influence of the Media on Public Opinion." University of Minnesota. Mimeo.

Fan, David. 1989. *Predictions of Public Opinion*. Westport, CT: Greenwood Press.

Forsyth, Richard and Roy Rada. 1986. *Machine Learning: Applications in Expert Systems and Information Retrieval.* New York: Wiley/Halstead.

Howell, Llewellyn D, Sheree Groves, Erin Morita and Joyce Mullen. 1986. "Changing Priorities: Putting the Data back into Events Data Analysis." Paper presented at the International Studies Association, Anaheim.

International Studies Quarterly. 1983. "Symposium: Events Data Collections." *International Studies Quarterly* 27.

Krippendorff, Klaus. 1980. *Content Analysis.* Beverly Hills: Sage.

Laurence, Edward J. 1988. "Events Data and Policy Analysis" Paper presented at the International Studies Association, St. Louis.

Lovins, J. B. 1968. "Development of a Stemming Algorithm." *Mechanical Translation and Computational Linguistics*. 11:1-2, 11-31.

McClelland, Charles A. 1983. "Let the User Beware." *International Studies Quarterly* 27,2:169-177

Munton. D. 1981. Measuring International Behavior: Public Sources, Events and Validity. Dalhousie University: Centre for Foreign Policy Studies

Pierce, John R. 1980. *An Introduction to Information Theory.* New York: Dover.

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

Schrodt, Philip A. and David Leibsohn. 1985. "An Algorithm for the Classification of WEIS Event Code from WEIS Textual Descriptions" Paper presented at the International Studies Association, Washington

Schrodt, Philip A. 1988a. "Statistical Characteristics of Events Data". Paper presented at the International Studies Association, St. Louis.

Schrodt, Philip A. 1988b. "Experimental Results on Event Coding *The New York Times* and FBIS". DDIR-Update 3,2 (October): 5

Stone, P.J., D.C. Dunphy, M.S. Smith and D.M. Ogilvie. 1966. *The General Inquirer: A*

Computer Approach to Content Analysis. Cambridge: MIT Press.

van Rijsberger, C.J. 1979. *Information Retrieval* (2nd edition). London: Butterworths.

Figure 1
UNFORMATTED NEXIS “WIRES” DATA

```
F53. Copyright (c) 1989 Reuters The Reuter Library Report, March 31, 1989,
Friday,
FAM cycle, 571 words, ISRAEL, U.S. TRADE SNUBS OVER PLO, By Paul Taylor,
FJERUSALEM, March 31, ISRAEL, LEAD: Israel and the United States have snubbed
Feach other in an apparent diplomatic tit-for-tat over Washington’s contacts
with
Fthe PLO, officials said on Friday.
F54. Copyright (c) 1989 Reuters The Reuter Library Report, March 31, 1989,
Friday,
FAM cycle, 224 words, ISRAEL SUMMONS CANADIAN AMBASSADOR TO DISCUSS PLO
F>>>.np
Fj5
F                              LEVEL 1 - 903 STORIES
F5JERUSALEM, March 31, ISRAEL-CANADA, LEAD: Israel has summoned the Canadian
Fambassador to protest Canada’s decision to upgrade talks with the Palestine
FLiberation Organisation (PLO), a Foreign Ministry spokesman said on Friday.
F55. Copyright (c) 1989 Reuters The Reuter Library Report, March 31, 1989,
Friday,
FAM cycle, 206 words, ISRAELIS BAN SOUTHBOUND CARS FROM SOUTH LEBANON BUFFER
ZONE
FJERUSALEM, March 31, ISRAEL-LEBANON, LEAD: Israel and its South Lebanon Army
F(SLA) allies have banned refugees fleeing fighting in Beirut from bringing
cars
Finto Israels self-declared security zone for fear of bombs or smuggled arms,
Fsecurity sources said on Friday.
F56. Copyright (c) 1989 Reuters The Reuter Library Report, March 31, 1989,
Friday,
FAM cycle, 104 words, ISRAEL AND U.S. AGREE TO DEVELOP "STAR WARS" RESEARCH
FCENTRE, JERUSALEM, March 31, ISRAEL-SDI, LEAD: Israel and the United States
have
Fagreed to develop a 35-million-dollar computerised research centre for the
U.S.
F"Star Wars" programme, an Israeli defence source said on Friday.
```

Figure 2
REFORMATTED NEXIS DATA WITH DATE, ACTOR AND TARGET CODES

890331 SAU UAR Reuters
King Fahd of Saudi Arabia left Cairo for home on Friday after a four-day trip that strengthened Egypt's position as peace-broker in the Arab-Israeli conflict.

890331 ISR USA Reuters
Israel and the United States have snubbed each other in an apparent diplomatic tit-for-tat over Washington's contacts with the PLO, officials said on Friday.

890331 ISR CAN Reuters
Israel has summoned the Canadian ambassador to protest Canada's decision to upgrade talks with the Palestine Liberation Organisation (PLO), a Foreign Ministry spokesman said on Friday.

890331 ISR LEB Reuters
Israel and its South Lebanon Army (SLA) allies have banned refugees fleeing fighting in Beirut from bringing cars into Israel's self-declared security zone for fear of bombs or smuggled arms, security sources said on Friday.

890331 ISR USA Reuters
Israel and the United States have agreed to develop a 35-million-dollar computerised research centre for the U.S. "Star Wars" programme, an Israeli defence source said on Friday.

890331 SYR ISR Reuters
Syria accused Israel on Friday of stirring up trouble in Lebanon and vowed to confront such schemes regardless of both the sacrifices and consequences.

890331 LBY UAR Reuters
Libya voted against Egypt's return to the Arab League's satellite communications organisation, Arabsat, conference sources in Oman said on Friday.

890331 MOR SYR Reuters
Moroccan Foreign Minister Abdellatif Filali will pay an official visit to Syria starting April 2, the Foreign Ministry said on Friday.

890331 NIG UAR Xinhua General
the nigerian air force (naf) is seeking cooperation with the egyptian air force in maintaining and servicing its equipment and faczilities [sic], chief of air staff air marshal ibrahim alfa said here today.

890331 ISR PAL Xinhua General
israeli troops shot and wounded eight palestinians during clashes friday in the occupied west bank and gaza strip, reports coming from jerusalem said.

890331 PAL ISR Xinhua General
three palestinian guerrillas were killed before dawn today in clashes with an israeli patrol in south lebanon, according to radio israel monitored here.

SAMPLE FBIS INDEX DESCRIPTIONS

025:Abd al-Majid, Algerian Ambassador On Cooperation

033:Arab League Official Arrives In Cairo 2 April

065: Membership In Arab Mining Company Restored

081:Electricity- Grid To Link With Asian Countries

032:Egyptian Assistant Foreign Minister Arrives 5 April

033:PRC Vice Minister Concludes Visit, Departs

033:Ghall Meets Ethiopian Envoy 10 April

031:Mubarak Receives Call From Mitteran On Lebanon

072:Air Force To Receive U.S. F-16, EC2 Planes

071:Soviet Union To Finance New Projects

031:Mubarak Receives Message From Mitterand

082:Ghana, Congo Seek Improved Relations

032:Hungarian Deputy Prime Minister Arrives

081:Sports Agreement Signed With Hungary

173:Israel Threatens To Destroy Iraqi Reactor

160:Spokesman Warns Israel Against Aggression

032:Italys Prime Minister De Mita Pays Visit

142:Agricultural Cooperation With Israel Denied

211:Israeli Convoy Advances Into Liberated Areas

211:Israelis Reinforce Positions; Overflights Noted

212:Terrorists Aboard Lebanese Vessels Detained

032:French Interior Minister Arrives On 3-Day

033:Mubarak Receives French Interior Minister

033:West German Parliamentary Delegation Arrives

031:Sidqi Receives Message From Iraq Ramadan

033:Ramadan Receives Egypts Sidqi Arrival

032:State War Production Minister Leaves For Baghdad

223:Egyptian Military Plane Reportedly Downed

024: Egyptians Say Plane Mistakenly Shot Down

141:Official Denies Nuclear Cooperation With Egypt

031:Merhav Briefs Egypt, PRC, USSR On Plan

SAMPLE IPPRC WEIS DESCRIPTIONS

94:Pope John Paul II calls for the survival of Solidarity, saying the union has become an integral part of the heritage of the workers of Poland and of other nations.

101:Saudi FM Saud al Faisal says that in return for Israeli recognition of Palestinian rights and the return of occupied lands, the Saudi gvt is prepared to accept Israel.

94:Saudi FM Saud al Faisal wants Americans to debate and rethink their policy in the Middle East.

121:Saudi FM Saud al Faisal charges that Israeli policy is to try to precipitate war.

121:Saudi FM Saud al Faisal charges that if USA policy in the Middle East is not changed, conflict in the region will occur.

95:West German Chancellor Schmidt appeals for greater understanding by nations on both sides of the Atlantic of the different economic, geographical, and political problems all allied leaders face, despite their common objectives.

94:Polish PM Jaruzelski calls for a meeting with ten European Community ambassadors.

182:Israeli jet fighters violate Israeli airspace.

121:Iraq charges that Israeli jet fighters violate Iraqi airspace.

21:Israel refuses to confirm the Iraqi charge of Israeli violation of Iraqi airspace.

121:Soviet Communist Party paper Pravda charges the Central Intelligence Agency with taking part in a conspiracy to overthrow the Polish gvt.

121:Soviet Communist Party paper Pravda charges the special services of certain other NATO countries with taking part in a conspiracy to overthrow the Polish gvt.

31:Polish leader Jaruzelski meets with Western European diplomats.

31:Western European diplomats meet with Polish leader Jaruzelski.

121:Polish leader Jaruzelski attacks USA PRS Reagan's policy of economic embargo as interference in Poland's internal affairs.

SAMPLE ICPSR WEIS DESCRIPTIONS

223:syr and isr exchanged fire

121:syr accused isr of intrusion

223:syr said its forces inflicted heavy damage on isr

223:isr said two were wounded

121:isr accuses syr of violating cease-fire by continuing to fire

223:isr tanks fired on syr

121:syr in note to uno accused isr of preparing a raid into syr

223:syr and isr exchange fire

223:isr and syr tanks clash on border in two hour battle

121:isr prm calls use of syr tanks in dmz a serious violation of the armistice agreements

160:isr warns if harrassment continued isr would determine her response

121:syr says that isr tractor had entered the dmz

121:syr said isr opened fire first

121:isr charges syr had fired on isr fishermen both yesterday and today

223:syr shore positions and isr patrol boat exchanged fire on the sea of galilee

160:isr fm warns syr about continuing border incidents

82:isr and syr accepted proposal for a meeting of the isr-syr armistice commission to avert large scale violence

54:syr reaffirms armistice pledge to isr

31:syr meet isr

150:syr demands eviction of isr from dmz

211:syr seized funds from jor for overthrowing syr

191:syr postponed isr-syr uno meeting

121:jor accused uar of helping smuggle arms into jor

192:jor recalls envoy to uar after nasser attacks king

112:uar accused jor of serving interests of imperialists

112:uar declared that jor and sau were lackeys of imperialism

223:isr and syr exchange fire

121:isr says that syr opened fire

223:isr reported that 6 migs from syr were shot down by isr jets

223:syr opened fire on isr border settlement

SAMPLE KEDS PHRASES

00: SOCCER
00: TENNIS
01: BACKIN DOWN FROM
01: BOWED TO PRESSU
01: LIFTED *8 CLOSUR ORDERS
03: ARRIVE IN
03: CONCLU *5 VISIT
03: DISCUS
03: HAS DISCUS WITH
03: HAS INVITE *9 VISIT
03: HELD *8 TALKS
03: HELD TALKS *8 WITH
03: HOSTED *8 MEETIN
03: LEFT *5 TODAY
03: LEFT *5 YESTER
03: LEFT HERE
03: LEFT ON
03: MET *5 WITH
03: MET HERE
03: WILL *1 VISIT
03: WILL LEAVE ON
03: WILL *4 MEET *8 IN
03: WILL *4 VISIT *2 SOON
04: HAS APPROV
04: REASSU
04: SUPPOR CALLS
04: WELCOM THE RESIGN
04: WELCOM
04: WOULD SUPPOR
05: ARE TO SIGN *8 PACT
05: HAS PLEDGE
06: CAN TAKE PART
06: COULD TAKE PART IN
06: EXPRES REGRET
06: HAS GRANTE
06: REOPEN ITS EMBASS
06: RESUME DIPLOM TIES
06: SET UP CONSUL TIES
06: SIGNED A PROTOC
06: WOULD PERMIT
07: APPROV *5 SALES
07: APPROV *8 AID
07: ARE TO GET
07: GET *9 AID
07: GIVE *8 DOLLAR
07: IS CONTRI *5 DOLLAR
07: LOAN TO FINANC
08: AGREES TO
08: SIGNED AN AGREEM
09: APPEAL TO
09: ASKED *5 TO
09: CALLED ON
09: HAS APPEAL TO
09: HAS CALLED FOR
09: URGED *7 RECONC
11: DISMIS
11: HAS RULED OUT
11: HAVE CONDEM
11: NEVER AGREE
11: REJECT
11: SAYS *3 OPPOSE
11: TURNED DOWN
11: UNABLE TO AGREE
12: ACCUSE
12: CHARGE
12: CONDEM
12: CRITIC
12: HAS BLAMED
12: POSED A REAL THREAT
13: DELIVE A *5 REBUFF

```
13: HAVE   EXPRES CONCER
13: LODGIN A    PROTES
14: DENIED
15: DEMAND
16: WARNED
17: ORDERE ATTACK ON
17: THREAT *5   TO   KILL
17: WOULD  PROVOK RETALI
18: STAGED *5   STRIKE
18: WENT   ON   *5   STRIKE
20: HAS  DEPORT
21: HAVE   ARREST
21: HAVE   DETAIN
21: IMPOSE AN   EMBARG
21: TROOPS CONFIN
21: TROOPS SEALED OFF
22: ACTIVI *8   KILLED
22: AIRCRA BOMBED
22: ALLIES KILLED
22: BEATEN BY   *3   POLICE
22: BEING  SHOT
22: FORCES KILLED
22: GUNSHI ATTACK
22: HACKED TO   DEATH
22: JETS  *5  BLASTE
22: JETS  *5  DESTRO
22: JETS  *5   RAIDED
22: PLANES ATTACK
22: SHOT   AT
22: SHOT   DEAD
22: WARPLA ATTACK
22: WERE   KILLED
22: WOUNDE
22: WOUNDI
```