[Proceedings of the *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, Hamilton, New York, August 14-16, 1985]

KNOWLEDGE RESOURCE TOOLS FOR ACCESSING LARGE TEXT FILES

Donald E. Walker
Artificial Intelligence and Information Science Research
Bell Communications Research
435 South Street MRE 2A379
Morristown, New Jersey 07960

ABSTRACT

This paper provides an overview of a research program just being defined at Bellcore. The objective is to develop facilities for working with large document collections that provide more refined access to the information contained in these "source" materials than is possible through current information retrieval procedures. The tools being used for this purpose are machine-readable dictionaries, encyclopedias, and related "resources" that provide geographical, biographical, and other kinds of specialized knowledge. A major feature of the research program is the exploitation of the reciprocal relationships between *sources* and *resources.* These interactions between texts and tools are intended to support experts who organize and use information in a workstation environment. Two systems under development will be described to illustrate the approach: one providing capabilities for full-text subject assessment; the other for concept elaboration while reading text. Progress in the research depends critically on developments in artificial intelligence, computational linguistics, and information science to provide a scientific base, and on software engineering, database management, and distributed systems to provide the technology.

1. INTRODUCTION

This paper provides an overview of a research program just being defined at Bellcore.[1] The objective is to develop facilities for working with large document collections that provide more refined access to the information contained in these “source” materials than is possible through existing information retrieval procedures and yet stop short of the processing required to identify their full meaning. The current technology supports search strategies that depend on matching index terms, which constitute a general characterization of the subject matter of a document, or on the application of pattern templates, which match specific sequences of words in the text. In contrast, most long range research efforts in computational linguistics and artificial intelligence are trying to understand the meaning of individual sentences by analyzing their syntactic, semantic, and discourse structures. Our own efforts can be described most simply as addressing an intermediate goal based on capturing the semantics of words and phrases in the documents.

The tools being used for this purpose are machine-readable dictionaries, encyclopedias, and related “resources” that provide geographical, biographical, and other kinds of specialized knowledge. An increasing number of these kinds of materials are becoming available in machine-readable form, primarily as a byproduct of the use of computer-driven photocomposition techniques. In our research, we are identifying and extracting the information contained in these *resources* so we can apply them

[1] This research is actually a continuation of work begun at SRI International in collaboration with Robert A. Amsler, who has also joined Bellcore to participate in this effort. His contributions—and often his words as well—are reflected in this paper.

more effectively to text. However, it is clear that in their present form they are not wholly adequate for our objective. As an example, consider the use of a current dictionary for understanding a newspaper story—a relationship to which we will return below. The central elements of such stories are people, places, institutions, and events, few of which are included in the dictionary. Because of this kind of mismatch, a central element of our research program is exploring the interactions between *sources* and *resources,* that is between the structural features of texts and organized bodies of knowledge. While other knowledge *resources* can supplement the dictionary entries, analyses of the texts themselves can provide additional data, for example, candidate terms for proper nouns and recurrent phrases. In addition, dictionary information can be used to help identify the topical focus of a text, and, given that focus, the text can be analyzed to reveal features that should be incorporated into the dictionary to increase its coverage.

The techniques we are developing are intended for use in a workstation environment by experts who organize and use information from documents in the course of their work. In contrast to the research in artificial intelligence on "expert systems" in which the goal is to extract the knowledge from an expert and embody it in a system, our concern might be said to be with "systems for experts," that is, with systems that support specialists in their search for knowledge. Two systems under development will illustrate our approach: one providing capabilities for full-text subject assessment; the other for concept elaboration while reading text.

In the next two sections, we consider some of the *sources* and *resources* we have been gathering. Following that, we present the two systems we have been building. In the conclusion we consider some of the directions for our work and their implications for information retrieval.

2. SOURCES

Over the years we have been acquiring a variety of materials for our research. The following list indicates the range and variety of *sources* so far collected. However, with the increasing use of computer-generated phototypesetting and word-processing, we will be limited only by our capacity for storing items and our ability to get permission to use them.

- **The Brown University Corpus** 1,000,000 words of text gathered in 1963 and 1964 in samples of 2,000 words and intended to provide a representative cross section of reading material from 15 different subject areas [Kucera and Francis, 19671 The collection was motivated by an interest in getting frequency statistics for English, the results of which are referenced in the next section.
- **New York Times News Service** Stories over 70,000,000 words collected over a period of more than two years, beginning in 1983. The Times also contains a broad range of subject matter and a variety of different prose styles.
- **Associated Press News Service** Releases we receive this material at a rate of 1,500,000 bytes a day, and have just begun to accumulate a sample for research.
- **The Handbook of Artificial Intelligence:** a comprehensive survey of the field compiled during the period 1975 to 1980 and issued in three volumes edited by Barr, Feigenbaum, Cohen [1981, 1982]. In addition to its broad coverage of the field, the **Handbook** provides an example of text formatted in TeX, a new language designed for use with new computer-based printers.
- **Understanding Expository Text** a scientific monograph by Britton and Black [1985] that both provides an opportunity to examine a full-length book and is at the same time a "handbook for

[1] Some publishers, in recognition of the possible market value of their documents in digital form, have made it more difficult to gain access to them for research purposes. We are particularly grateful for those companies that have made their products available to us.

analyzing explanatory text."

- **Unix Manuals**: five volumes of manuals for using, managing, and programming in Unix [1979, 1984]. In addition to text, they contain a wide variety of material in many different formats.
- **Moby Dick** and **Pride and Prejudice**: two novels that illustrate two quite different English prose styles.
- **LATA Switching Systems Generic Requirements**: a 5,000,000 byte file that contains the official definition of the telephone central office switch.

3. RESOURCES

The *resources* we have collected can be grouped in five categories: word frequency lists, machine-readable dictionaries, derivative dictionary data, reference works, and databases. Each will be considered in turn.

Word Frequency Lists: as the name indicates, they contain information on the relative frequency of occurrence of words in text:

- **Computational Analysis of Present-Day American English**: statistics on the 1,000,000 word Brown University Corpus described in the previous section [Kucera and Francis, 1967]. Over 50,000 different graphic word types are present, arranged both in order of frequency and alphabetically. For each entry, information is provided about its relative occurrence in the 15 different subject areas and 500 samples. A wide range of distributional analyses have been made of the data.
- **The American Heritage Word Frequency Book**: statistics on 5,000,000 words of text gathered in 500 word samples from the kinds of reading children in grades 3 through 9 were likely to encounter in the late 1960's [Carroll, Davies, and Richman, 1971]. The 87,000 graphic word types were derived from 22 different kinds of materials, mostly different curriculum levels. They are grouped both alphabetically and by frequency, distinguishing both grade level and category.
- **New York Times News Service Corpus**: statistics on an 8,300,000 word sample of stories corresponding to three months of material distributed in the last three months of 1983 [Walker and Amsler, 1985]. The database in which this material has been stored has been designed to accumulate information about case and punctuation as well as by frequency according to day, month, and year.

Machine-Readable Dictionaries: Each contains a large variety of different kinds of information for an entry: spelling form, syllabification, pronunciation, part-of-speech, inflections, etymology, sense distinctions, definitions, usage notes, example sentences, and synonyms. We have at various times been working with the following machine-readable dictionaries:

- **Merriam-Webster New Collegiate Dictionary** (Seventh Edition): one of the classic collegiate dictionaries and probably the first available in machine-readable form [G&C Merriam, 1963]. It contains over 70,000 entries and represents more than 15.5 million bytes of data. It has been widely used in research because of its extensive distribution following its computerization in the mid-1960's.
- **Merriam-Webster New Pocket Dictionary**: an abridged version of the 'Seventh' [G&C Merriam, 1964]. It contains almost 23,000 entries and consumes 4.5 million bytes of storage.
- **Longman Dictionary of Contemporary English**: a dictionary designed for people learning English as a second language [Proctor, 1978]. It contains over 55,000 entries and occupies 14 million bytes of storage. It is distinctive in a number of respects: the defining vocabulary is limited to 2,000 words; it has a very refined set of grammatical codes; and the computer tape provides detailed semantic information and a set of subject codes to characterize specialized senses of word.

- **Oxford Advanced Learner's Dictionary**: another "Learner's" dictionary [Hornby, Gatenby, and Wakefield, 19631 it contains 35,000 entries and occupies 6.5 million bytes of storage. It is distinctive in containing an unusually large number of example sentences.

Derivative Dictionary Data: Two sets are available; they represent the taxonomic relationships among the 24,000 different noun senses and the 11,000 verb senses in the Merriam-Webster Pocket Dictionary [Amsler, 1980, 1981]. The kernel terms in the definitions of each noun and verb sense were identified and disambiguated with respect to the appropriate word sense in their definitions. Relationships among the resulting pairs of words were calculated and a set of tangled hierarchies generated. Processing of this kind is intended to provide a more comprehensive lexical semantic classification of the language, demonstrating the complex interrelationships among words.

Reference Works: The first three listed below are currently available; the fourth is expected in the near future.

- **The World Almanac & Book of Facts 1985**: one of the classic reference works for factual information [Newspaper Enterprise Association, 1984]. It contains text, tables, illustrations, and time lines to summarize encyclopedic information about science, history, geography, biography, and related categories of information. The source file, still being extracted from its phototypesetting matrix, occupies 10 million bytes of storage.
- **INSPEC** citation data for 6 months of the Computer and Control Section from 1980; about 10,000 items and 2 million bytes of storage.
- **Psychological Abstracts**: citation data plus abstracts for 2 years of material published in 1982 and 1983; 5-10 million bytes of storage.
- **Academic American Encyclopedia**: one of the first encyclopedias to be available in machine-readable form [Grolier, 1984]. It has more than 29,000 articles and tables and incorporates material online that is not in the printed edition.

Databases: Material under this heading is not in text form. The set below is primarily geographical in emphasis, but that reflects in part an interest in maps in the group. A large variety of other types of databases would be easy to acquire.

- ZIP-Code List (5-digit): contains all the towns in the United States with ZIP-code assignments, including street name and address breakdowns for cities with multiple ZIP-codes.
- Names and Address List: for 1.25 million people living in Kansas; about 60 million bytes of data.
- Telephone Yellow Pages for two major US area codes from 1980 and 1984; approximately 6 million bytes.

4. FORCE4, A SYSTEM FOR FULL-TEXT CONTENT ASSESSMENT

The first system we developed to demonstrate the interaction between *sources* and *resources* is **FORCE4**, a procedure for full-text content assessment.[2] It makes use of a set of subject codes assigned to specialized word senses in the **Longman Dictionary of Contemporary English (LDOCE)**, applying them to stories from the **New York Times News Service (NYTNS)**. The result is an identification of the primary subject content of those stories.

[2] This work was actually done at SRI International; the system was designed by Robert Amsler, and much of this description can be found in Walker and Amsler [1985]. The system was implemented on a DEC/20 computer. The name **FORCE4** reflects the currency of *Star Wars* when we began the project.

4.1 How FORCE4 Processes Text

To understand how **FORCE4** works we first need to consider the **LDOCE** subject codes in more detail. There are 120 two-letter field codes that mark, for example, areas like medicine (MD) and political science (PL). These field codes are divided into 212 subfield categories; for example, physiology is represented as MDZP and diplomacy as PLZD, the Z being used in the third position exclusively as an indicator of subcategorization. The field codes can also be combined so that the designation for meteorology (ML) together with the one for building (CO), that is MLCO, is used to mark the entry *lightning conductor.* Similarly, MLGO (meteorology plus geography) marks *temperate* and *torrid,* while GOML (the same fields in the reverse order) marks *permafrost* and *drift ice.* In addition, there are 38 locality codes that identify major geographical areas and countries or distinguish areas within them. Thus, U represents Europe and F represents France; combined with meteorology, the code MLUF is applied to *mistral,* a distinctive wind that is characteristic of southern France. The word *typhoon* is marked MLX, meteorology and Asia. There are over 2600 realized combinations of two, three, and four-letter codes. Out of the 55,000 entries in the dictionary, 18,000 are marked as having specialized subject senses.

The following weather report will be used to illustrate the operation of **FORCE4**.

> Heavy rainfall and high winds clobbered the California coast early today, while a storm system in the Southeast dampened the Atlantic Seaboard from Florida to Virginia.
>
> Traveler's advisories warned of snow in California's northern mountains and northwestern Nevada. Rain and snow fell in the Dakotas, northern Minnesota and upper Michigan.
>
> Skies were cloudy from Tennessee through the Ohio Valley into New England, but generally clear from Texas into the mid-Mississippi Valley.

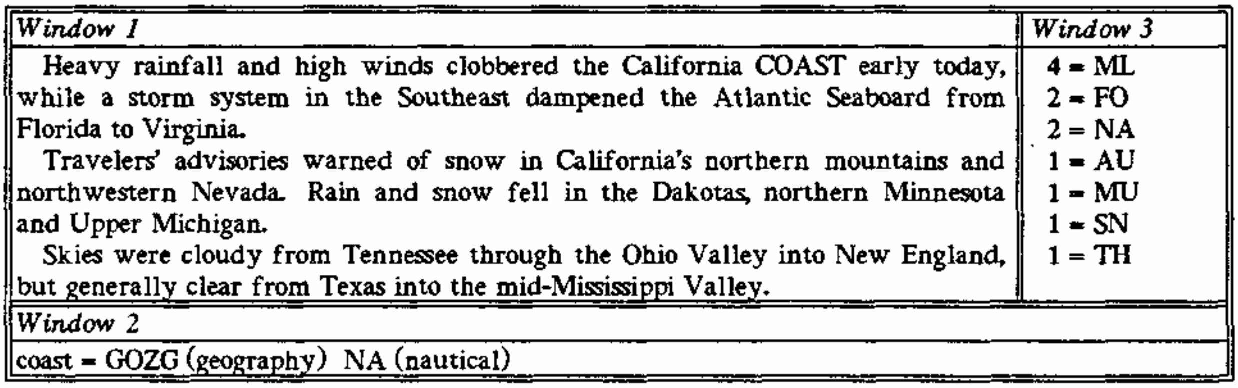
The **LDOCE** subject codes for the first four content words in the report are as follows:

- *heavy:* FO—food, ML—meteorology, TH—theatre
- *rainfall:* ML—meteorology
- *high:* AU—motor vehicles, DGXX—drugs and drug experiences, FO—food, ML—meteorology, RLXX—religion, SN—sounds
- *wind*: HFZH—hunting, MDZP—physiology, ML—meteorology, MU—music, NA—nautical

FORCE4 works by applying the **LDOCE** codes to each successive word in the text. The frequency of occurrence of these codes is cumulated, and the codes themselves are arranged in order of frequency. We developed a display program to show how this process provides an assessment of stories. The program employs a full-screen display format with three windows that contain, respectively: (l) the text being processed; (2) the program's intermediate inferences regarding the syntactic and semantic properties of each content-bearing word in the text; and (3) a running tally of the frequencies of the top subject assignments made to the document on the basis of the cumulative set of content-bearing words that have been analyzed.

The text appears in *Window 1,* which occupies the major part of the screen, beginning at the upper left corner. For each content-bearing word in the text, the subject codes are looked up in the **LDOCE**. If the word fails to have a subject code, it is analyzed to determine whether it is the inflected form of some word with a set of subject codes. If subject codes are found, they are displayed together with their English descriptions in *Window 2* at the bottom of the screen. The subject codes identified for a word are merged into the set of subject codes established for the text so far, and the resulting array of revised frequencies is sorted and displayed in high-to-low order in *Window 3* at the upper right of the screen.

Figure 1 shows **FORCE4** in the course of processing a weather report. The text has been analyzed through the word *coast* (capitalized here, but in inverse video in actual operation). The most frequent subject code is meteorology (ML) with 4; food (FO) and nautical (NA) both have 2; the rest all have the value 1. Completing the processing of the text yields the results shown in Figure 2. Meteorology (ML) is still the most frequent code with 10; geographical terms (GOZG) and drugs and drug experiences (DGXX) have 4; nautical (NA) has 3; and food (FO) and military (MI) have 2.



Window 1

Heavy rainfall and high winds clobbered the California COAST early today, while a storm system in the Southeast dampened the Atlantic Seaboard from Florida to Virginia.

Travelers' advisories warned of snow in California's northern mountains and northwestern Nevada. Rain and snow fell in the Dakotas, northern Minnesota and Upper Michigan.

Skies were cloudy from Tennessee through the Ohio Valley into New England, but generally clear from Texas into the mid-Mississippi Valley.

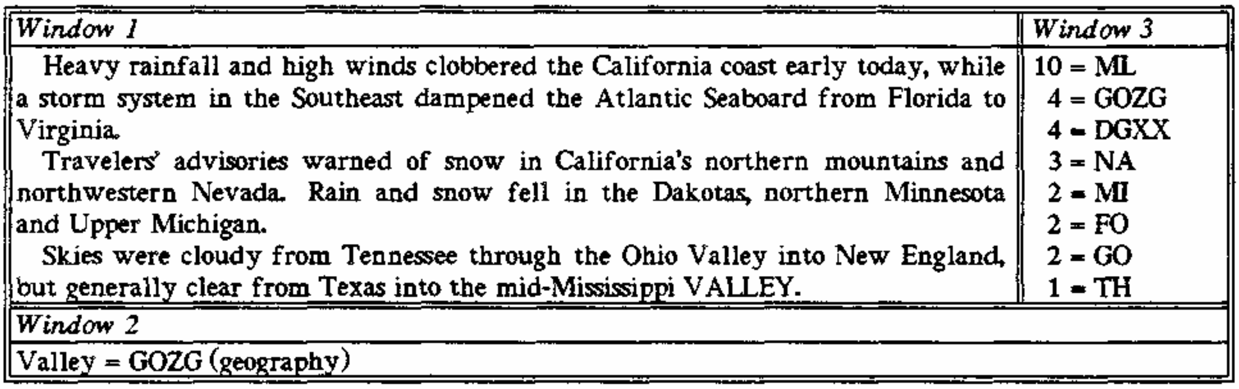
Window 3

4 = ML
2 = FO
2 = NA
1 = AU
1 = MU
1 = SN
1 = TH

Window 2

coast = GOZG (geography) NA (nautical)

Figure 1. Example of **FORCE4** processing, through the word *coast*.



Window 1

Heavy rainfall and high winds clobbered the California coast early today, while a storm system in the Southeast dampened the Atlantic Seaboard from Florida to Virginia.

Travelers' advisories warned of snow in California's northern mountains and northwestern Nevada. Rain and snow fell in the Dakotas, northern Minnesota and Upper Michigan.

Skies were cloudy from Tennessee through the Ohio Valley into New England, but generally clear from Texas into the mid-Mississippi VALLEY.

Window 3

10 = ML
4 = GOZG
4 = DGXX
3 = NA
2 = MI
2 = FO
2 = GO
1 = TH

Window 2

Valley = GOZG (geography)

Figure 2. Example of **FORCE4** processing, through the word *Valley*.

A set of more than 100 **NYTNS** stories (a 24-hour sample) was processed against the **LDOCE** codes to determine subject content. The results were remarkably good; **FORCE4** works well over a variety of subjects—law, military, sports, radio and television—and several different formats—text, tables, and even recipes.

4.2 A Discussion of the FORCE4 Approach

The **FORCE4** approach definitely merits further development. However, it is worthwhile to consider the weather example in a little more detail in order to point out some of the current limitations and, in general, to illustrate problems encountered in using machine-readable dictionaries.

First, it is appropriate to consider the words in the text that were marked with subject codes: *heavy, rainfall, high, winds, coast, storm, Southeast, Seaboard, Virginia, snow, mountains, northwestern, rain, snow, fell, Upper, skies, cloudy, Valley, New, clear, Valley.* Two of the words that were marked for meteorology, *heavy* and *high* were actually being used as adjectives modifying the following nouns *(rainfall* and *winds,* respectively) and not in the coded sense. *Seaboard* should have been treated as part of a compound, *Atlantic Seaboard,* and not as a separate word, as should *Upper* with respect to

Michigan and *New* with respect to *England.* The two *Valleys* may also be parts of compound expressions, since they denote areas quite different from the states they modify.

The following words did not have subject codes (function words and other general vocabulary items are excluded): *clobbered, California, early, today, system, dampened, Atlantic, Florida, Travelers', advisories, warned, California's, northern, Nevada, Dakotas, Michigan, Tennessee, Ohio, England, generally, Texas, mid-Mississippi.* The obvious thing to note is the predominance of state names and other regional designations. The **LDOCE**, like most dictionaries, does not include many proper nouns. *Virginia* is actually coded for its tobacco sense. The rest of this set of terms probably do not have specialized subject senses, although it is a little puzzling for *Southeast* and *northwestern* to be included, while *northern* is not.

There are, of course, a number of conditions that have to be satisfied in order for FORCE4 to work well, even at its present, early state of development:

1. The content-bearing words of the text must have entries in the LDOCE.
2. The subject-codes for the content-bearing words must include the sense in which the word is being used in the text.
3. The content-bearing words must not also be common function words, e.g., as *in* is in the sentence, "The tide is in."
4. A sufficient quantity of text must be examined for the topmost subject-code assignments to stabilize (typically more than a sentence, but often less than two paragraphs).
5. The text must be about a single topic, rather than a collection of different topics such as is found in a news summary of major headline stories.

The procedures that we have described for **FORCE4** constitute only the first step in the development of its capabilities for content assessment. One extremely desirable extension entails *pruning* the spurious sense entries in the text. If we accept the principle that a word with multiple senses, and thus several subject codes, is likely to be used in the text only in one of these senses, then the following procedure can be applied. Take the code with the highest frequency—that would be meteorology (ML) in Figure 2—and, for all the words so marked, eliminate from the cumulative frequency list all of the other codes they contained. The ordering there showed meteorology (ML) as the most frequent code with 10; geographical terms (GOZG) and drugs and drug experiences (DGXX) had 4; nautical (NA) 3; and food (FO) and military (MI) 2. After pruning, ML would still be 10, of course; GOZG would remain at 4; DGXX would be reduced to 1; NA would be reduced to 2; FO would be eliminated; and MI would become 1. Note that, if a slightly more radical pruning strategy were invoked—that is, removing a subject code in the list if any instances of it were eliminated—DGZZ, NA, and MI would all be 0, leaving only ML and GOZG, which is, of course, the desired *profile* for this weather report It should be obvious that much more experimentation is needed to explore these issues.

To explore the significance of the discrepancy between words in text and dictionary entries, we compared our 8 million word sample of the **NYTNS** text with the **Webster's New Collegiate Dictionary (W7)**.[3] Of the 119,630 different word forms present in the **NYTNS** sample and in the **W7**, 27,837 (23%) occurred in both; 42,695 (36%) occurred only in the **W7**; and 48,828 (41%) occurred only in the **NYTNS**. The fact that almost two-thirds of the words in the dictionary (61%) did not appear in the text is not surprising; dictionaries contain many words that are not in common use. That almost two-thirds of the words in the text were not in the dictionary (64%) is more

[3] The **W7** was chosen for this study because it has a larger vocabulary, similar results would be expected for the **LDOCE.**

problematic. A preliminary analysis of a sample of the **NYTNS** forms that were not in the **W7** reveals the following breakdown (expressing the values in fractional form is intended to show their approximate character): one-fourth were inflected forms; one-fourth were proper nouns; one-sixth were hyphenated forms; one-twelfth were misspellings; and one-fourth were not yet resolved, but some are likely to be new words occurring since the dictionary was published.

The inflected forms can be accommodated by performing a morphological analysis on the text entries. Hyphenated forms can also be handled, although some of the instances found in the **NYTNS** are more difficult to deal with; they also broach the issue of noun-noun compounds and phrases, a critical problem that will be discussed at more length below. Misspellings are less easy to detect and correct, but recent developments in spelling correction algorithms suggest that some progress is being made in this area [Durham et al., 1983]. The missing proper nouns constitute a more serious problem. As noted in the analysis of the weather text, most dictionaries do not contain the names of people, places, institutions, trade-names, and similar items. Yet these entries are key features of newspaper text and essential for almost any class of documents. Geographical and biographical dictionaries can provide an initial base from which to develop such entries, but it would, of course, be necessary to assign subject codes to them and perhaps to establish new subject codes for them.[4]

Specialized technical dictionaries are one source for the additional vocabulary required. It is also possible to acquire entries from the texts themselves, although obviously at this stage of our understanding it would necessarily be a machine-aided operation. We have done some preliminary work toward this objective. Using the simple criterion of selecting all the words in the **NYTNS** sample that occur at least five times with initial capitals and never occur only in lower case, we were able to create a large file of entries that, while far from exhaustive, certainly contain many candidates for inclusion as proper nouns. Taking advantage of patterns in the text, we made some progress toward identifying cities in the United States as capitalized words preceding state names and followed by a comma. Similarly, the introduction of a name in a story is often followed by some explanatory information set off by commas. Consider some recent examples from the Associated Press:

- Edwin M. Joyce, president of CBS News,...
- Jessie Helms, R-N.C.,...
- Winfield, I1l., about 20 miles west of Chicago, ...
- Maryland Toleration Act, passed in 1649,...
- Prime Minister Margaret Thatcher, a staunch supporter of self-starting capitalists, ...

In these remarks so far, we have been ignoring one of the major problems in using dictionary entries for content assessment. That is the relative scarcity of multiword entries in a dictionary. The problems of performing semantic analyses of noun-noun compounds are, of course, well known [Rhyne, 1976; McDonald, 1982]. Equally difficult is the identification of aggregates of terms that should be treated as units in a "phrasal lexicon" [Becker, 1975; Smith et al., 1982]. As noted in discussing the weather text, *Atlantic Seaboard, Upper Michigan,* and *New England* should all have been treated as single entries. Working with the **NYTNS** text, we have made a beginning in attacking this problem by identifying groups of frequently recurring words, specifically those bounded by function words and sentence boundaries. The most frequent entries in the resulting lists are candidates for inclusion as multiword "dictionary" entries.

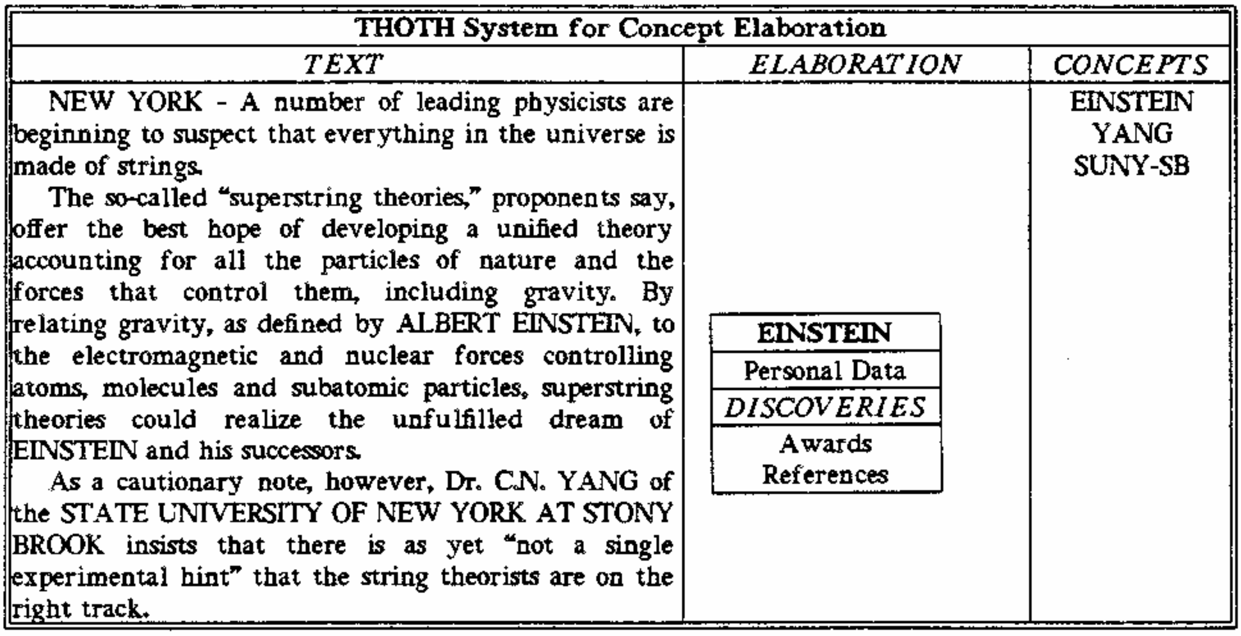
With regard to augmentation of the current **LDOCE** subject codes, it is worth remarking that even those that currently exist may need refinement. Certainly, if the **FORCE4** procedure is to be

[4] It should be noted that both people and places can have sense distinctions that may be important to separate for given contexts. For example, a particular person may be noted as both actor and political figure; a state may be considered in relation to its location, its wine, and its ethos.

extended to more technical text or even to more homogeneous text where greater discrimination is desired than that provided by the current coding, parameters will need to be developed that motivate the creation and assignment of additional categories. One potential source for them is the kind of taxonomic analysis entailed in the creation of the derivative dictionary data described in the *resource* section above [Amsler, 1980, 1981]. These taxonomies establish relations among terms that can be exploited in making further dimensional analyses. For example, recognizing that *vehicle* includes *automobile, bicycle, carriage, locomotive, sled, tractor, truck,* and *wagon,* and noting that the definitions refer to parameters reflecting 'motive force,' 'objects transported,' 'surface medium,' and the like, one can begin to organize terms along these lines. The work by Evens et al. [1980] on semantic-relations provides another direction for expansion. On a more technical level, it should be possible to take classification systems and thesauruses collected for a particular field of study and use them to help structure the vocabulary in that field.

5. THOTH, A SYSTEM FOR CONCEPT ELABORATION

The second system we are developing is intended to provide a user with a means of elucidating text during the course of reading it. In its current implementation, **THOTH** identifies concepts in stories from the **New York Times News Service**, the **Associated Press**, and similar *sources,* and then displays elaborations of them based on entries in the World Almanac and related *resources* to a user. The system makes significant use of the kind of workstation environment provided by LISP machines, in our case, the *Symbolics 3670.*[5]



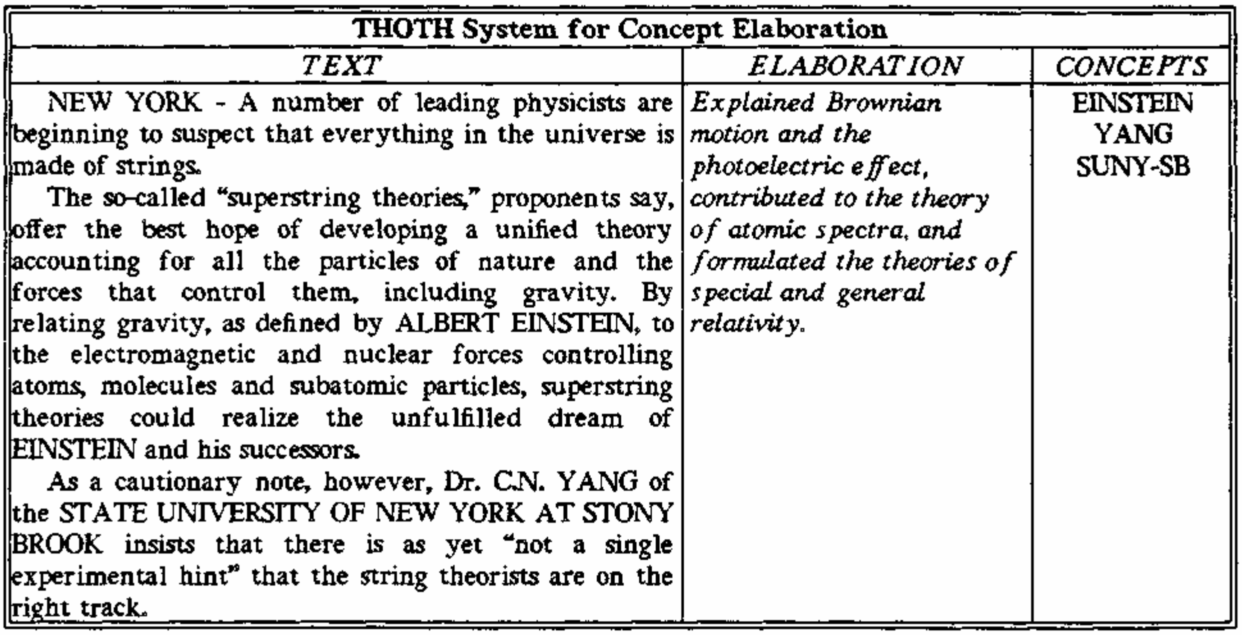
THOTH System for Concept Elaboration

TEXT	*ELABORATION*	*CONCEPTS*
NEW YORK - A number of leading physicists are beginning to suspect that everything in the universe is made of strings. The so-called "superstring theories," proponents say, offer the best hope of developing a unified theory accounting for all the particles of nature and the forces that control them, including gravity. By relating gravity, as defined by ALBERT EINSTEIN, to the electromagnetic and nuclear forces controlling atoms, molecules and subatomic particles, superstring theories could realize the unfulfilled dream of EINSTEIN and his successors. As a cautionary note, however, Dr. C.N. YANG of the STATE UNIVERSITY OF NEW YORK AT STONY BROOK insists that there is as yet "not a single experimental hint" that the string theorists are on the right track.	**EINSTEIN** Personal Data *DISCOVERIES* Awards References	EINSTEIN YANG SUNY-SB

Figure 3. Illustration of **THOTH** with a concept menu.

[5] George Collier was responsible for the design and implementation of THOTH; Robert Amsler has been processing the World Almanac to derive the material for concept elaboration. *Thoth* was the Egyptian god of wisdom and learning; the choice of that term for the system was influenced by our use of ancient Assyrian names for our *Symbolics* computers: Eridu, Nippur, Kish, and Uruk (Samuel Epstein is responsible for that venue); unfortunately, we did not find an Assyrian name that so nicely conveyed the intent of the system and at the same time provided an allusion to the more common use (except in AI circles) of the word "lisp".

Setting up **THOTH** entailed the following steps (l) a set of windows was created on the display to provide space for scrolling the *text* and for displaying the *concepts* and the *elaboration;* (2) a set of stories from one of the news services was identified; (3) concepts considered relevant for consideration were selected; and (4) appropriate information from the **World Almanac** and other relevant *resources* was extracted and encoded. Using the system in its current fledgling form is relatively simple. A person reads text on the screen; concepts that the system recognizes in the story are both highlighted in the text and listed in the *concepts* window; by moving the mouse to position the cursor on a concept, the user can determine what options are available and select from them. This operation is illustrated in Figures 3 and 4, which show how the information is displayed on the *Symbolics 3670* screen. The "Einstein" menu in the *elaboration* window of Figure 3 is the result of selecting that concept from the text Figure 4 presents the information associated with *Discoveries.*



THOTH System for Concept Elaboration

TEXT	*ELABORATION*	*CONCEPTS*
NEW YORK - A number of leading physicists are beginning to suspect that everything in the universe is made of strings. The so-called "superstring theories," proponents say, offer the best hope of developing a unified theory accounting for all the particles of nature and the forces that control them, including gravity. By relating gravity, as defined by ALBERT EINSTEIN, to the electromagnetic and nuclear forces controlling atoms, molecules and subatomic particles, superstring theories could realize the unfulfilled dream of EINSTEIN and his successors. As a cautionary note, however, Dr. C.N. YANG of the STATE UNIVERSITY OF NEW YORK AT STONY BROOK insists that there is as yet "not a single experimental hint" that the string theorists are on the right track.	*Explained Brownian motion and the photoelectric effect, contributed to the theory of atomic spectra, and formulated the theories of special and general relativity.*	EINSTEIN YANG SUNY-SB

Figure 4. Illustrating the result of menu selection on THOTH.

THOTH provides a context within which we can begin to examine how concepts associated with particular terms can be elaborated. The first research issue that we are addressing concerns how to extract and encode material contained in a *resource so* that it constitutes an appropriate elaboration for a concept. We are working primarily with the **World Almanac** at this time. A major problem is determining how the attributes and values are expressed in the format patterns contained in the text. For example, in the section on "Nations of the World" each country begins with a centered boldface header containing the name in a larger font size. The major attributes, like *People, Geography, Government,* and *Economy,* begin paragraphs and are printed in boldface followed by a colon. Subordinate attributes within the paragraph are also in boldface followed by colons: for example, *People* contains *Population, Age distrib., Pop. density, Urban, Language, Ethnic groups,* and *Religions,* The values for these latter attributes—in roman font—are variously expressed as numbers, percentages, text, or some combination, often with dates in parentheses to indicate the time that data item was collected. To complicate matters further, the colons themselves may be in bold face or roman, without regard for the hierarchical level. Consequently, to distinguish major from subordinate attributes, it is necessary to know whether one begins a paragraph or not. A brief history of the country usually concludes that section, most often without a specific heading bearing that name; however, for more complex entries, like the United Kingdom and the USSR (with their several constituencies), that attribute may appear in boldface, too, but usually followed by a period.

As the preceding characterization is meant to imply, the different attributes and their values are not neatly compartmentalized. In addition, not every country contains every attribute; nor are all the values consistently rendered. As a result, this first extraction is requiring a substantial amount of "hand" editing for coherent results. One consequence, though, is that we should be in a position to advise the editors of the **World Almanac** on how to create a database from which entries in the printed version could be more systematically generated. In any case, the attribute/value structures do lend themselves well to realization in **THOTH,** and we are making substantial progress in identifying the ones in the **World Almanac** and organizing them for appropriate use.

6. DISCUSSION

Content assessment and *concept elaboration* are only two of the possible benefits flowing from our explorations of the interactions between document *sources* and reference *resources.* Our goal is to provide a more general set of tools that can be used by knowledge specialists for working with large text files. However, as the discussions of **FORCE4** and **THOTH** suggest, one of the steps is creating an appropriate set of workstation environments for our own research in order to cope with the vast amount of data that is entailed.

It is clear that progress will depend critically on developments in artificial intelligence, computational linguistics, and information science. Areas of particular interest are knowledge representation and knowledge acquisition; the lexicon, parsing, and text generation; and document analysis, abstracting, and indexing. However, effective realization of these expansions of the scientific base also depends on developments in technology. We need new software for handling both the sequential and structural features of texts, which can be viewed both as strings and as trees or tangled hierarchies. Storage and access for billions of words of text and for dictionaries that contain hundreds of thousands of entries require advances in database management techniques. The range and variety of the knowledge *resources* suggest a distributed systems design embodying server concepts. The massive movements of data entailed demand access through wideband networks, as does the accommodation of images and other "nontextual" features that must eventually be included in a documentary data base.

Projecting the results of our efforts is clearly premature, but it is appropriate to note that, if we are successful, the "systems for experts" that we are working toward will create an entirely new environment for knowledge workers, increasing their access to vast document collections and simplifying the procedures for working with the information they contain.

7. ACKNOWLEDGMENTS

I wish to express my appreciation in particular to Robert Amsler for his many contributions to the work reported here. George Collier deserves special mention for his work on **THOTH,** I am also indebted to my colleagues in **21233**, especially Robert Allen, David Copp, Stephen Hanson, and Patricia Gildea, and to Michael Lesk from Bellcore and George Miller from Princeton for their support and consultation.

It is appropriate to thank the people and organizations who have provided the *sources* and *resources* that we are using in our research. Special recognition is due the suppliers of the materials that have figured most prominently in our current activities the Longman Group for the **Longman Dictionary of Contemporary English**, G.&C. Merriam Company for the **Webster's Seventh New Collegiate Dictionary,** the Newspaper Enterprise Association for **The World Almanac,** the New York Times Syndication Sales Corporation for the **New York Times News Service,** and the Associated Press for the **Associated Press News Service.**

The research on **FORCE4** was supported in part by grants IST-8208578, IST-8209346, and IST-8300940 from the National Science Foundation, grant LM 03611 from the National Library of Medicine, a contract funding SRI International Project 5383, and Internal Research and Development funds from SRI International.

8. REFERENCES

Amsler, R.A. 1980. The Structure of the Merriam-Webster Pocket Dictionary. Ph.D. Dissertation, University of Texas at Austin, Texas.

Amsler, R.A. 1981. “A Taxonomy for English Nouns and Verbs.” In **Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics.** Menlo Park, California; Association for Computational Linguistics. Pp. 133-138.

Barr, A.; Feigenbaum, E.A.: and Cohen, P.R. (Eds.) 1981, 1982. **The Handbook of Artificial Intelligence** (3 volumes). Stanford, California: HeurisTech Press; Los Altos, California: William Kaufmann.

Becker, J. 1975. “The Phrasal Lexicon.” In **Theoretical Issues in Natural Language Processing.** R. Schank and B.L. Nash-Webber, eds. Menlo Park, California: Association for Computational Linguistics. Pp. 60-63.

Britton, RK.; and Black, J.B. 1985. **Understanding Expository Text.** Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Carroll, J.B.; Davies, P.; and Richman, B. 1971. **The American Heritage Word Frequency Book.** Boston, Massachusetts: Houghton Mifflin.

Durham, I.; Lamb, D.A.; and Saxe, J.B. 1983. “Spelling Correction in User Interfaces.” **Communications of the ACM,** 26:10, 764-773.

Evens, M.W.; Litowitz, B.E.; Markowitz, J.A.; Smith, R.N.; and Werner, O. 1980. **Lexical Semantic Relations: A Comparative Survey.** Carbondale, Illinois, and Edmonton, Canada: Linguistic Research, Inc.

Grolier Electronic Publishing. 1984. **Academic American Encyclopedia.** New York, New York: Grolier Electronic Publishing.

Hornby, A.S.; Gatenby, E.V.; and Wakefield, H. 1963. **The Advanced Learner’s Dictionary of Current English. London,** England: Oxford University Press.

G.&C. Merriam Company. 1964. **The New Merriam-Webster Pocket Dictionary.** Springfield, Massachusetts G. & C Merriam Company.

G.&C. Merriam Company. 1963. **Webster’s New Collegiate Dictionary** (Seventh Edition). Springfield, Massachusetts: G. & C. Merriam Company.

Kucera, H.; and Francis, W.N. 1967. **Computational Analysis of Present-Day American English.** Providence, Rhode Island: Brown University Press.

McDonald, D.B. 1982. Understanding Noun Compounds. Ph.D. Dissertation, Carnegie-Mellon University, Pittsburgh, Pennsylvania. [CMU Technical Report CMU-CS-82-102.]

Newspaper Enterprise Association. 1984. **The World Almanac & Book of Facts 1985.** New York, New York: Newspaper Enterprise Association.

Procter, P. (Ed.). 1978. **Longman Dictionary of Contemporary English.** Harlow and London, England: Longman Group Limited.

Rhyne, J.R. 1976. Lexical Rules and Structures in a Computer Model of Nominal Compounding. Ph.D. Dissertation, University of Texas, Austin, Texas.

Smith, R.N.; Bienstock, D.; and Housman, E. 1982. “A Collocational Model of Information Transfer.” In **Information Interaction: Proceedings of the 45th ASIS Annual Meeting, Volume 19.** A.E. Petrarca, C.I. Taylor, and R.S. Kohn, eds. White Plains, New York: Knowledge Industry Publications.

Unix. 1979, 1984. **Unix User’s Manual** (2 volumes), **Unix Programmer’s Manual** (2 volumes), **Unix System Manager’s Manual.** Berkeley, California: University of California.

Walker, D.E.; and Amsler, R.A. 1985. “The Use of Machine-Readable Dictionaries in Sublanguage Analysis.” In **Sublanguage: Description and Processing.** R. Grishman and R. Kittredge, eds. Hillsdale, New Jersey: Lawrence Erlbaum Associates,