Searching Legal Literature Electronically: Results of a Test Program

Jessica S. Melton

Robert C. Bensing

Follow this and additional works at: https://scholarship.law.umn.edu/mlr

Part of the Law Commons

Recommended Citation


https://scholarship.law.umn.edu/mlr/1606

This Article is brought to you for free and open access by the University of Minnesota Law School. It has been accepted for inclusion in Minnesota Law Review collection by an authorized administrator of the Scholarship Repository. For more information, please contact lenzx009@umn.edu.
Searching Legal Literature Electronically:

Results of a Test Program

The replacement of doormen with electric-eye mechanisms and of bank clerks with computers has probably troubled most of us at one time or another, but we have always been able to return to the comforting thought that it can't happen to lawyers—at least not yet. To the extent that we think of the lawyer as an intelligent instrument of research, we have been deluding ourselves: the machine age is here. The subject of this Article is an ingenious system of indexing and encoding legal information for electronic retrieval from a modern computer. The system is so conceived that the computer will supply answers (in the form of citations) to questions ranging from a request for cases containing certain specific words to a request for cases concerning basic legal concepts. In the following pages, the authors describe the already well-established use of the system in the scientific literature, explain the logical analysis involved in processing source materials and questions for the computer, and venture an opinion as to the implications of the system for legal research in the years ahead.

Jessica S. Melton*

Robert C. Bensing**

INTRODUCTION

In a 1959 address entitled "Progress and New Developments in Electronic Research for the Lawyer,"1 Vincent P. Biunno, Coun-

*Assistant Director for Technical Operations, Center for Documentation and Communication Research, School of Library Science, Western Reserve University.

**Professor of Law, School of Law, Western Reserve University.

1. Presented in August, 1959, at a meeting of the American Bar Association in Miami, Florida.
sel to the Governor of New Jersey, made this comment on the state of legal documentation:

With each passing year, we pile up decision on statute on rule on regulation and then construct large and cumbersome digests, compendiums, indexes and other archeological devices which we hope will help us find what we want in the evergrowing mound. And as we do it we can feel in our bones that we are taking two steps forward and sliding three steps back. Ask any legislator how many bills he has had passed to eliminate a law from the statute books. Ask him for a full description of all the ways in which the bill he introduced will alter existing law. Ask him what provisions he had in mind when he wrote: "All laws or parts of laws inconsistent herewith are hereby repealed." Ask any judge writing an opinion, or a lawyer writing a brief, whether he can really say he had the time to look for, find, analyze and apply every precedent relevant to a point at issue. There is strong suspicion that the mountain of precedents has grown to such size that legal research ordinarily consists of no more than snatching the first bit of relevant material that can be found and then flying by the seat of the pants. Let us not delude ourselves. Our legal system depends on precedent to insure that we have a government of laws and not of men, but in practice we rely more on generalized experience, on the lawyer's "feel" based on vague personal recollections of precedent, rather than on precedent itself.

There is no doubt that the punch card and the computer can lead us out of this morass. The primitive experiments carried on mostly by individuals in various parts of the country have clearly shown that these devices are a natural tool for the law. But, and this "but" is a big one, it will take a very serious effort, a great deal of time and quite a bit of money before any of us can push a few buttons and have a printed slip with citations pop out of a slot on the desk. The real questions are: Should we get involved in this? Will it be worth it? How do we go about getting started?

This Article describes one experiment—an attempt to apply to the literature of the law the same documentation techniques developed for and successfully applied to the literature of science and technology. The investigation has been a co-operative effort of the School of Law and the Center for Documentation and Communication Research of Western Reserve University, Cleveland, Ohio.

A. PREVIOUS APPLICATIONS OF ELECTRONIC SEARCHING

The Center for Documentation and Communication Research was established in 1955 at Western Reserve University as a research arm of the School of Library Science. A long-range goal of the Center has been the development and application of equipment and techniques to facilitate the communication and utilization of recorded knowledge in the way that well-equipped laboratories contribute by experimentation to the generation of new knowledge.
One of the first tasks undertaken in 1955 was an investigation of the feasibility of searching metallurgical literature by machine. The American Society for Metals sponsored the experimental and pilot phases of this program. By January, 1960, the program reached an operational stage of sizeable proportions. Approximately 150 documents (journal articles, reports, books, patents, etc.) are processed each day for future machine retrieval. Fifty of these daily documents fulfill the Center’s contract with the American Society for Metals, which calls for the analysis and machine encoding of 12,000 metallurgical documents during 1960. The remainder of the daily quota (about 100 documents) is in fulfillment of a contract grant from the National Science Foundation to extend the scope of coverage of the literature from pure metallurgy to peripheral fields such as physics, nuclear science, mechanics, physical chemistry, and geology. It is estimated that by the end of 1960, a file of machine-readable indexes for 35,000 to 40,000 documents will have been processed.

These files are used by subscribers to the Machine Documentation Service, which is operated by the American Society for Metals. Still in the pilot stage are two projects for the recording of medical information. Both of these projects are entering their second year. By the end of 1961, each will have been tested and evaluated and will be ready to enter an operational state analogous to that of the metallurgical literature service.

B. Scope of the Law Project

To investigate the problems connected with the application of advanced computer techniques to the searching and correlating of information of interest to the lawyer, the project personnel decided to analyze the Sales provisions of the Uniform Commercial Code and the cases interpreting this part of the Code. The code-and-cases combination was adopted in order to test the Western Reserve University (WRU) system in both these areas. The Uniform Commercial Code was selected because of its potential national importance. Expediency was the determinant in the decision to index only one part of the Code, and the selection of the one part—Sales—was, on the whole, arbitrary.

It was believed that if these statutes and cases could be indexed and cross references provided, so that the machine system would respond to questions by directing the questioner to all the relevant code passages and cases, this would demonstrate the feasibility of applying the WRU system to larger portions of the law.

The product of the WRU pilot project was two reels of encoded, machine-readable tape, comprising two sets of encoded indexes ready for machine searching. The first reel contains encoded versions of article 2 (Sales) of the Uniform Commercial Code. The second contains the encoded contents of a number of cases wherein provisions of article 2 had been construed. To test the usefulness of the indexing, questions were formulated, the answers to which were within the contents of the analyzed and encoded material.

A typical question was:

Buyer and seller entered into a written agreement, whereby seller was to supply buyer with specified machine components. Seller also submitted a sample which was approved by buyer. The components delivered to the buyer were found defective and not fit for the purpose for which the buyer intended to use them. The seller knew the purpose for which the components were intended to be used. The buyer rejected the components.

QUESTION: The buyer wishes to know if he has a cause of action against the seller.

This question was analyzed by the WRU project personnel, encoded, and stated in the type of formal logic which a machine used for searching can recognize. The searching machine was then programmed, or instructed, to identify information relevant to the question. The files of encoded information were scanned. Three answers were identified and supplied by the machine in the form of the following citations:

1. U.C.C. §2-313
2. U.C.C. §2-315

That the answers are relevant to the question can be seen by reading the case and statutes to which they refer.3

3. Section 2–313 of the Uniform Commercial Code is set out in the text infra at p. 240. Section 2–315 provides:

Implied Warranty: Fitness for Particular Purpose
Where the seller at the time of contracting has reason to know any particular purpose for which the goods are required and that the buyer is relying on the seller's skill or judgment to select or furnish suitable goods, there is unless excluded or modified under the next section an implied warranty that the goods shall be fit for such purpose.

And a brief of the Smith's Sons case discloses the following:

Suit by plaintiff buyer to recover price paid for defective castings. Defendant seller counterclaims for value of castings delivered and which were rejected, but for which defendant claims he was not promptly notified of rejection.
Plaintiff, a New York corporation, contracted to buy castings from defendant, a Pennsylvania corporation. A sample was submitted by de-
I. BASIC THEORY OF THE WRU INDEXING SYSTEM

There are several techniques by which the statutes and cases could have been processed for machine searching. One way would have been to translate every word of the source material into machine-readable form. The person conducting the search could then ask the machine to identify passages containing specific words he felt would be contained in the authority pertinent to his problem. This method, while providing a complete "index," has the disadvantage of yielding a high probability of "no" answers and false answers, for the questioner could seldom anticipate the exact words or grammatical constructions in the authority for which he was searching. Thus, if his question specified "dog" and the only pertinent authority contained the word "collie" rather than "dog," he would not get an answer from the machine. Similarly, if his question specified "man bites dog" and the only authority containing the same words was a case in which "dog bites man," the machine would obviously produce a false answer. The latter situation, called a "false drop" in the field of information retrieval, can theoretically be avoided by a complex syntactical analysis. However, such machine programs would increase the cost of searching far beyond its practical value at the present time.

Another way of applying machine methods to indexing statutory and case materials would be to follow conventional indexing rules of the type the West Publishing Company has formulated, and then to translate this index into machine-readable form. This method would have the advantage of ordering and systematizing the subject matter involved, and also would enable the index to be machine-read very rapidly. The disadvantages of this "topic" and "subtopic" approach are two-fold: (1) the indexer would have to

[The Smith's Sons case also contained a procedural question. For the purpose of this pilot project, however, non-Sales questions were not encoded.]


[For the purpose of the example involved here, citation is made only to the Uniform Commercial Code. However, the Smith's Sons case occurred in Pennsylvania; consequently, the actual citation was to PENN STAT. ANN. tit. 12A, §§ 2–313, 2–315 (1954). Further, although the Uniform Commercial Code sections were cited by the court, those sections technically were not applicable, since the cause of action accrued prior to the effective date of adoption in Pennsylvania of the Uniform Commercial Code. The case was actually governed by comparable sections of the Uniform Sales Act which the court also cited in support of its holding, but which, for the sake of simplicity, have not been included in the sample brief set forth here.]
anticipate future questions and provide access by means of a key word or two—an almost impossible task, and (2) because the index would be conceptual rather than specific, it would be impossible to confine a search to cases containing a specific, descriptive word which is a vital element in the problem confronting the person conducting the search.

The WRU indexing system is a compromise between the extremes of specificity and abstraction just described. Under the WRU approach, only "significant words" which convey the important concepts in the passage are selected for indexing. The restriction of indexing to "significant words" (which may be compared to the words which would be listed in the West Publishing Company's Descriptive Word Index) permits a sharp reduction in the total volume of stored data without a material sacrifice in the specific content of that data. The converse advantage of conceptual analysis is obtained through the logic inherent in the WRU system of encoding words, for under this logical system an encoded word is comprised of the several generic concepts expressed by that word. Thus the WRU system of indexing is designed to permit searches either for specific words or for generic concepts—or for both simultaneously.

A. Semantic Code

The heart of the WRU indexing system is the semantic code into which significant words are converted. In establishing this code, the WRU analysts have used much the same logic as is used in the parlor game known as "twenty questions"—except, of course, that in the encoding process the analyst is concerned with deriving the logical elements of a known word, rather than with establishing an unknown word from its logical elements.

Thus in encoding the word blueprint, the analysts reasoned that a blueprint is a form of information, that it is used to regulate, and that it acts upon construction. They then set up four-digit combinations of letters called "semantic factors," which symbolize the individual logical elements of the word to be encoded. Thus the logical element information is symbolized by the incomplete semantic factor D[ ]CM, the logical element regulation is symbolized by the incomplete semantic factor R[ ]GL, and the logical element construction is symbolized by the incomplete semantic factor C[ ]NS.

It will be noted that logical analysis of the word blueprint is not completely represented by reducing that word to the semantic factors indicated, for the representations of information, regulation, and construction are insufficient to convey the concept that a blueprint is a form of information, used to regulate, and acting upon
construction. This further logical analysis of significant words is accomplished in the semantic code by supplying a single letter called an "infix," in the second digit of each semantic factor. Thus in the semantic factor DACM (which symbolizes the logical element *a form of information*) the basic concept *information* is symbolized by the semantic factor D[|]CM, while the insertion of the infix A symbolizes the particular application *a form of*. Different applications of the basic concept *information* (D[|]CM) are expressed by inserting other infixes (such as W, U, V, and Y), in the second digit of the four-letter semantic factor.

We have traced the analysis of the word *blueprint* from the ultimate level of abstraction, as symbolized by the semantic factor D[|]CM (*information*), R[|]GL (*regulation*), and C[|]NS (*construction*), into an intermediate level of abstraction, as symbolized by the infixed semantic factors DACM (*a form of information*), RUGL (*used to regulate*), and CVNS (*acting upon construction*). It remains to explore the manner in which specificity is obtained in the semantic code.

Synonyms or extremely close approximations of the word *blueprint* necessarily contain the same logical elements found in that word, but the semantic code must permit distinguishing the significant word *blueprint* from similar significant words. This goal is attained by appending a four-digit numerical factor to the semantic factors contained in a given word. Thus the word *blueprint* is encoded as CVNS DACM RUGL 3002, while the logically similar word *specifications* is encoded as CVNS DACM RUGL 3001.

In addition to providing for specificity as to significant words, the semantic code also provides for specificity at the ultimate level of abstraction. As already noted, we can vary the application of a *basic concept* by varying the second letter of the semantic factor based upon that concept. Similarly, by inserting a four-digit numerical factor among the semantic factors of an encoded word, so that the numerical factor immediately follows the particular semantic factor which it is to modify, we have a device for specifying a *type of the basic concept* upon which the semantic factor is based. For example, the word *model* is encoded as CANS 1021 RYGL 2001. The interior numerical factor 1021 indicates that the type of construction involved is specifically *reproduction* (and of course the exterior numerical factor 2001, in combination with the three preceding factors, identifies the specific word *model*). Perhaps, strictly speaking, the interior numerical factor is more properly a tool of abstraction (similar to the device of varying the second letter of a semantic factor) than a device for specificity, but it is easier to keep in mind the relative roles of the semantic and numerical factors if the former are considered as devices of abstraction and the latter are thought of solely as specifying devices.
With this brief theoretical background, one may begin to appreciate the extreme flexibility afforded by the semantic code. Thus the encoding of the word *model* as CANS 1021 RYGL 2001 makes it possible to locate a passage in which *model* is a significant word by a number of logical routes, ranging from a search for the specific word *model* (for which the entire code would be specified by the searcher) to a search for the basic concept *construction* (C[ INS]) or *regulation* (R[ ]GL).

The full range of research possibilities opened by the semantic code is suggested by the following search patterns to which the encoded word *model* (CANS 1021 RYGL 2001) might respond:

1. C[ INS (construction)
2. CANS (a form of construction)
3. C[ INS 1021 (construction of the specific type known as reproduction)
4. CANS 1021 (a form of reproduction)
5. R[ ]GL (regulation)
6. RYGL (characterized by regulation)
7. C[ INS R[ ]GL (construction and regulation)
8. C[ INS RYGL (construction and characterized by regulation)
9. CANS R[ ]GL (a form of construction and regulation)
10. CANS RYGL (a form of construction characterized by regulation)
11. C[ INS R[ ]GL 2001 (any aspect of model)
12. C[ INS RYGL 2001 (some aspect of model and characterized by regulation)
13. CANS R[ ]GL 2001 (some aspect of model and a form of construction)
14. CANS RYGL 2001 (some aspect of model, a form of construction characterized by regulation)
15. C[ INS 1021 R[ ]GL (construction, specifically reproduction, and regulation)
16. C[ INS 1021 RYGL (construction, specifically reproduction, and characterized by regulation)
17. C[ INS 1021 R[ ]GL 2001 (some aspect of model and construction, specifically reproduction and regulation)
18. C[ INS 1021 RYGL 2001 (some aspect of model and construction, specifically reproduction, characterized by regulation)
19. CANS 1020 R[ ]GL (a form of construction, specifically a form of reproduction, and regulation)
20. CANS 1021 RYGL (a form of construction, specifically a form of reproduction, characterized by regulation)
21. CANS 1021 R[ ]GL 2001 (some aspect of model, a
form of construction, specifically a form of reproduction, and regulation)

(22) CANS 1021 RYGL 2001 (the specific word model)

It will be noted in the foregoing list that the addition of the exterior numerical factor to a semantic code not completely supplied with letter-symbol infixes or interior numerical factors has the effect of indicating that the code in question is one for some aspect of the specific concept implied by a complete code bearing that exterior numerical factor. This is not particularly important with such a word as model; hence it may be clearer if the example of the words swim, swimming, and swam is used. Each of these words contains the same semantic factors, M[ ]TN (motion) and H[ ]DR (water), and each has the same exterior numerical factor, 2001. The words, however, are differentiated by the infixes, thus:

swim, MATN HVDR 2001 (a form of motion acting upon water)
swimming, MUTN HVDR 2001 (used for motion acting upon water)
swam, MWTN HVDR 2001 (acted upon by motion acting upon water)

Thus the incomplete code M[ ]TN H[ ]DR 2001 would indicate that the concept encoded has something to do with swimming, but the particular application is not indicated.

In a long list of encoded terminology, the foregoing 22 search patterns could detect hundreds of significant words whose meanings are related on various conceptual levels. The multi-level conceptual relationships of words are graphically illustrated by a list of the encoded versions of five words (including model) which might be significant in a case arising under a provision of the Uniform Commercial Code:

(1) CVNS DACM RUGL 3002 (blueprint)
(2) CVNS DACM RUGL 3001 (specifications)
(3) CANS 1021 RYGL 2001 (model)
(4) DACM RUGL 2002 (terms)
(5) DWCM 1086 (written)

It is because of this feature of conceptual relationship that semantic factors may be likened to “cluster points” for terminology. In this respect, semantic factors are analogous to the West Publishing Company’s Key Numbers, which are described as “cluster points around which certain cases naturally gravitate.” If the analogy to West’s topical approach is continued, a glance back at the encoded version of blueprint shows that the semantic code puts that word simultaneously under the topics of “things having to do

with regulation" (R[ ]GL), of "things having to do with information" (D[ ]CM), and of "things having to do with construction" (C[ ]NS).

In West's Law Finder, each of the Key Number topics is described as a panel of letter boxes in a post office; within each panel there are many pigeonholes. Similarly, the semantic factor C[ ]NS (construction) may be thought of as a "panel" in which there are "pigeonholes" for the significant words blueprint, specifications, and model, as reference back to the encoded versions of those words will confirm. In the "panel" D[ ]CM (information), there are "pigeonholes" for blueprint, specifications, terms, and written. In the "panel" R[ ]GL (regulation), there are "pigeonholes" for blueprint, specifications, model and terms.

In printed indexes, a term can be put under more than one heading only by repeating the term or by cross-referencing to it, both of which lengthen the index and add to printing costs. Thus, because of space limitations and economic considerations, potentially useful information often is omitted from printed indexes. Under the WRU indexing system, however, there is no corresponding omission of useful information, for the semantic code automatically makes each encoded word accessible to almost every conceivable search, whether the search be specifically or abstractly oriented.

B. SYMBOLIC GRAMMAR

The WRU analysts have constructed a highly simplified symbolic grammar, so that the variety of English syntactical expression may be transformed into a uniform, consistent pattern. This symbolic grammar consists of three-digit combinations of letters (e.g., KEJ, KAM, KVJ) called "role-indicators." The role-indicators are used to relate each word or phrase to the other words and phrases in the passage to be encoded. By this means, the actual usage of a word in the passage is made clear.

Although the encoding of a lengthy passage may require the selection of a large number of significant words in order to characterize adequately the informational content of the passage, only a limited number of role-indicators are required. For purposes of the law project, the WRU analysts formulated role-indicators with the aim of designating fundamental relationships which are basic to the law. These role-indicators may be described as having the same analytical function as the following five elements which the West Publishing Company lists as being "common to every case": (1) parties, (2) places and things, (3) basis of action or issue,
(4) defense, and (5) relief sought. The WRU analysts have formulated role-indicators to express these and other relationships, both legal and nonlegal. The function of role-indicators is explained further in the section of this Article dealing with operational procedures.

C. SYMBOLIC PUNCTUATION

The WRU indexing system employs punctuation symbols which perform the same functions as do conventional punctuation marks, in addition to performing certain other functions for which there are no counterparts in conventional punctuation. Because punctuation is perhaps best understood through its use, more detailed discussion of the punctuation symbols is reserved for the following material on operational procedures. In general, however, the punctuation symbols simply provide further clarification of the relationships between the significant words and their role-indicators.

II. OPERATIONAL PROCEDURES

The foregoing discussion has been concerned with the theoretical bases of the WRU indexing system. What follows is an explanation of how one goes about operating such a system—how source materials are encoded for electronic storage, and how questions are formulated so as to extract information pertinent to the questioner's problem.

A. PROCESSING SOURCE MATERIALS FOR ELECTRONIC STORAGE

The processing of source materials requires a very small number of men and machines. The machines are standard punch-card equipment, including a key-punch unit, a punch-card file, and an automatic encoding unit. The requisite personnel consist of an abstracter (also known as an "indexer" or "analyst") and an operator for the punch-card equipment. Of these two persons, only the former need be "learned in the law."

Very briefly, the WRU system of processing source materials operates in the following manner:

(1) The abstracter analyzes the passage to be encoded, writes down the significant words (in their ordinary form), assigns symbolic role-indicators to those words, adds punctuation symbols, and hands the written abstract to the operator of the punch-card equipment. (Note that the abstracter does not translate the significant words into the semantic code.)

(2) The machine operator copies the abstract on the key-punch unit.

(3) The punch-card equipment automatically checks the punch-card file for the semantic-code equivalents of the significant words, fits the encoded words in with their concomitant role-indicators and punctuation, and transcribes the wholly symbolic abstract on a reel of magnetic tape. (The tape is then stored for later use in a computer, by means of which searching is accomplished.)

A more complete understanding of the mechanics of the WRU pilot law project may be obtained through an examination of a concrete problem. Therefore, the following discussion retraces the steps which were actually taken in processing section 2-313 of the Uniform Commercial Code, which reads as follows:

Express Warranties by Affirmation, Promise, Description, Sample.

(1) Express warranties by the seller are created as follows:

(a) Any affirmation of fact or promise made by the seller to the buyer which relates to the goods and becomes part of the basis of the bargain creates an express warranty that the goods shall conform to the affirmation or promise.

(b) Any description of the goods which is made part of the basis of the bargain creates an express warranty that the goods shall conform to the description.

(c) Any sample or model which is made part of the basis of bargain creates an express warranty that the whole of the goods shall conform to the sample or model.

(2) It is not necessary to the creation of an express warranty that the seller use formal words such as "warrant" or "guarantee" or that he have a specific intention to make a warranty, but an affirmation merely of the value of the goods or a statement purporting to be merely the seller's opinion or commendation of the goods does not create a warranty.

The analysts initially decided that the following words of section 2-313 were the most important for indexing: express warranty, affirmation, promise, description, sample, seller, buyer, creation, goods, bargain, and conform. Next, role-indicators were assigned to each significant word and punctuation symbols were added. At this point it became apparent that the role-indicators adequately conveyed the information which the words creation and conform expressed, and those words were deleted from the abstract. Similarly, it became clear that an additional word, selling, was necessary if the abstract was to convey all of the important concepts contained in the section.

6. It should be noted that in the WRU system of indexing for machine retrieval, the index can be made as detailed as is useful to the reference problems of the legal profession. Any aspect of the subject matter which is contained in (or is implicit in) the subject matter may be included.
The result of the foregoing analysis was a written abstract which appeared in the telegraphic form indicated in Table I.\(^7\) (Please note: *Only the material contained in the first column of Table I appeared in the telegraphic abstract; the numbers at the left and the material at the right of Table I are included solely for purposes of explanation.)*

**Table I**

<table>
<thead>
<tr>
<th>Punctuation, Role-Indicators and Significant Words</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) .KEJ, goods</td>
<td>A thing or person is the object of a process or transaction (a relationship symbolized by the role-indicator KEJ), namely, <em>goods</em>.</td>
</tr>
<tr>
<td>(2) ..KAM, selling</td>
<td>A process or transaction occurs (KAM), namely, <em>selling</em>, which is</td>
</tr>
<tr>
<td>(3) .KQJ, seller</td>
<td>by or from (with the idea of instrumentality) (KQJ) a <em>seller</em>, and which is</td>
</tr>
<tr>
<td>(4) .KVJ, buyer</td>
<td>to or toward (with the idea of recipient) (KVJ) a <em>buyer</em>.</td>
</tr>
<tr>
<td>(5) ..KDF KAM KWJ, express warranty</td>
<td>A thing is defined (KDF), namely, <em>express warranty</em>; a thing results, is made, or is created (KWJ), namely, <em>express warranty</em>; a process occurs (KAM), namely, <em>express warranty</em>; and</td>
</tr>
<tr>
<td>(6) .KAM, bargain</td>
<td>a process occurs (KAM), namely, <em>bargain</em>; both the express warranty (in each of its aspects) and the bargain are</td>
</tr>
<tr>
<td>(7) .KQJ, seller</td>
<td>by or from (KQJ) a <em>seller</em></td>
</tr>
<tr>
<td>(8) .KVJ, buyer</td>
<td>to or toward (KVJ) a <em>buyer</em> and are</td>
</tr>
<tr>
<td>(9) .KAH, affirmation</td>
<td>conditional upon (KAH) an <em>affirmation</em>,</td>
</tr>
<tr>
<td>(10) .KAH, promise</td>
<td>conditional upon (KAH) a <em>promise</em>,</td>
</tr>
</tbody>
</table>

\(^7\) In the pilot project, information contained in the comments following each provision of the Uniform Commercial Code was also included in the telegraphic abstract. For the sake of simplification, such information was excluded from the abstract shown in Table I.
Punctuation, Role-Indicators and Significant Words

<table>
<thead>
<tr>
<th>Number</th>
<th>Role-Indicator</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>.KAH, description</td>
<td>conditional upon (KAH) a description,</td>
</tr>
<tr>
<td>12</td>
<td>.KAH KEQ, sample</td>
<td>conditional upon (KAH) a sample, and may be equated to (KEQ) a sample.</td>
</tr>
<tr>
<td>13</td>
<td>.KEQ, goods</td>
<td>There is an equation of and to (KEQ) a thing or person, namely, goods.</td>
</tr>
<tr>
<td>14</td>
<td>../UCC § 2–313</td>
<td>The source of the foregoing information is UCC § 2–313.</td>
</tr>
</tbody>
</table>

Although the reader may tend to disagree at this point, the function of the punctuation symbols utilized in the foregoing abstract is to avoid confusion. One or more periods (.) placed in front of one or more role-indicators (KAH, KEQ, etc.) links the role-indicator to the term which follows it. Line five of the foregoing abstract illustrates the usefulness of this device. For example, the role-indicator KDF (indicating that something is defined) is linked with express warranty, and not with any other word in the entire telegraphic abstract. Consequently, in a machine search for “statutes defining affirmation, goods, and seller,” section 2–313 would not respond to the search, even though that section does contain the words affirmation, goods, and seller.

Two periods (..) placed in front of one or more role-indicators indicate that one “complex” or “phrase” of information is complete at that point and that subsequent punctuation, role-indicators, and words pertain to a succeeding “phrase.” Thus, in the foregoing abstract, lines two through four form one such “phrase” and lines five through 12 form another. This device prevents “cross-talk,” or false associations, between units of information within different “phrases.” On the other hand, the units of information within a phrase all interrelate. Thus, in lines five through 12, the analyst has indicated that both the express warranty and the bargain are made by the seller and that they are conditional upon the terms listed in lines nine through 12. However, the selling (listed in line two) is not conditional upon these terms; therefore, it has been placed in a separate “phrase.”

A comma (,) placed after one or more role-indicators indicates the end of role-indicator information and the beginning of “word” information. During subsequent processing a comma will also be
placed after the code equivalent of each word,\textsuperscript{8} to signify the end of the word information. A slash (/) is an instruction to the searching machine to print out the section number whenever the section so designated answers a question.

After the telegraphic abstract of section 2–313 was written by the analyst, the role-indicators, punctuation, and words were transferred to punch-cards by a key-punch operator. Punch-card equipment then automatically matched the newly punched cards to a punch-card file, in order to translate the words into the semantic code. The punch-card file may be compared to a bilingual dictionary, in that it contains the semantic-code equivalent of each English word used by the analyst to characterize the subject matter of the Uniform Commercial Code. There are no “grammar” words in this dictionary, since in this system grammatical connections are conveyed by role-indicators. The words in this dictionary are all legal terms or qualifying terms. Thus the process of encoding the English words selected for indexing is essentially a “dictionary look-up” operation to translate each English word into its semantic-code counterpart. However, in the WRU system, machines perform this dictionary look-up process in its entirety by a series of matching and reproducing operations. The fully encoded information is then automatically transferred to tape.

In the processing of section 2–313, therefore, the end result was a completely encoded abstract, recorded on tape and ready for machine searching. This abstract would print out from the tape as follows:

\texttt{.KEJ, BANN 1032,... KAM, SULL 1001,, KQJ, SWLL 1001,}
\texttt{.KVJ, PQPR 1001,, KDF KAM KWJ, RYPR 1098, GARM SUFT 2002,, KAM, GARM 1003,, KQJ, SWLL 1001,}
\texttt{.KVJ, PQPR 1001,, KAH, DACM 1082,, KAH, GARM 1009,}
\texttt{.KAH, DUSP 1003,, KAH KEQ, LAMN MYTT RUGL 3001,, KEQ, BANN 1032,, /UCC § 2–313}

\textbf{B. Processing Questions for Electronic Search}

In order for the machine (computer) to identify passages which are relevant to a question, it must be instructed to look for the pieces of code in combinations which will constitute an answer. When a question is to be prepared for machine search, the first step toward instructing the machine to identify pertinent authorities is to analyze the question in the same way in which the Uniform Commercial Code sections and cases were analyzed.

To illustrate, let us return to the question outlined at the beginning of the article, relating to the unfortunate situation of a pur-

---

\textsuperscript{8} See the completely encoded abstract set out in the text \textit{infra} at the end of this section of the Article.
chaser of defective machine components. Analysis of that question led to the conclusion that to be an "answer," a statute or case must contain: (1) mention of an object of a transaction or process, (2) a process or transaction, namely buying or selling, and (3) a written agreement, or an agreement in which specifications or samples were supplied by the seller. In regard to the first of the three preceding elements, the analyst reasoned that statutory law would probably generalize this concept, and would be unlikely to contain the word machine components. Therefore, in searching the files containing the statute, only the role-indicator KEJ (a thing or person which is the object of a transaction or process) was specified. However, in the search for precedents in the file of cases, the word machine components was specified in the hope of finding an exact precedent. If it developed that there were no exact precedents, the search could then be broadened gradually so that such concepts as "fabricated parts," "fabricated objects," and "goods" in general would be identified.

When the question had been analyzed, it was next encoded by the analyst, who looked up each English word in the semantic code dictionary. It was at this point that the analyst defined the scope of the search. Remembering that the semantic code allows the search to be conducted on many levels of generality or specificity, an analyst, by choice of codes and pieces of codes, can frame a question which will detect only the statutes and decisions exactly in point, or the search can be extended to any degree believed profitable. In fact, several versions of a question can be searched simultaneously, thus furnishing the questioner with a set of answers graded as to the degree of probable pertinency to the question.

The sample question with which we have been dealing was first encoded as follows:

(1) KEJ (an object is "transacted" or processed)
(2) KAM (transaction or process)
(3) PUCR 1001 (buying)
(4) SULL 1001 (selling)
(5) DWCM 1086 GARM 1001 (written agreement)
(6) GARM 1001 (agreement)
(7) CWNS DACM RUGL 3001 (specifications)
(8) LAMN MITT RUGL 3001 (sample)
(9) KQJ (by or from)
(10) SWLL 1001 (seller)
The analyst next explored the dictionary of semantic codes, which, as was mentioned earlier, is arranged to act as a thesaurus of terminology. He found, for example, that the codes for the words *specifications* and *sample* led to an array of conceptually similar words, such as *blueprint* and *terms*. He also saw that this array had one semantic factor in common, RUGL; meaning that all these words shared in the meaning “used to regulate.” Next, he turned to the array of words containing D[ ]CM in their codes. Here he found the following kinds of words:

- rights: DACM DYMN 2002
- reason to know: DACM 1041 LYGL 2006
- knowledge: DACM 1041
- obligations: DACM DYMN 2003
- offer: DYCM 1066 LYGL 2005
- parol: DWCM SYND 1006 2002
- patent: DYCM 1002 PUTT 2002

After searching all the chains of conceptual associations displayed by the semantic-code thesaurus, the analyst revised the coding of the question to read as shown in Table II.

### Table II

<table>
<thead>
<tr>
<th>Encoded Abstract</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) KEJ</td>
<td>an object is “transacted” or processed</td>
</tr>
<tr>
<td>(2) KAM</td>
<td>transaction or process</td>
</tr>
<tr>
<td>(3) PUCR 1001</td>
<td>buying</td>
</tr>
<tr>
<td>(4) SULL 1001</td>
<td>selling</td>
</tr>
<tr>
<td>(5) D[ ]CM</td>
<td>information</td>
</tr>
<tr>
<td>(6) GARM</td>
<td>a form of agreement</td>
</tr>
<tr>
<td>(7) RUGL</td>
<td>used to regulate</td>
</tr>
<tr>
<td>(8) KQJ</td>
<td>by or from</td>
</tr>
<tr>
<td>(9) SWLL</td>
<td>acted upon by selling</td>
</tr>
</tbody>
</table>

The reason for the revision was to enable the machine to identify concepts synonymous with or closely related to those conveyed by the English words given in the original question. For example,
the codes for "specifications" and "sample" each include RUGL (used to regulate); specifications is encoded as CWNS DACM RUGL 3001, sample as LAMN MYTT RUGL 3006. By generalizing, or "relaxing," the search requirement to RUGL only, instead of the complete codes for "specifications" and "samples," the searcher enabled the machine to detect other words whose codes contain RUGL, such as "blueprint," and "terms," which would be equally relevant. It is the analyst's task to state these complete codes and/or separate semantic factors in the relationships necessary to meet the total informational requirements.

In this instance, both complete codes and separate semantic factors were adopted to complete the search specification. In its full form, the search specification might have been stated as follows:

An object is "transacted" or processed in one element of the document. In the same document, but in another element, either buying or selling is shown as a transaction or process; a form of agreement is shown either as information or as something used to regulate; and something is shown as acted upon by selling, with its role in the element being by or from.

In order to express the preceding statement of the search specification in the more concise, symbolic language that a computer "understands," however, a logical notation was used to assemble the encoded elements of the abstract shown in Table II. Thus the encoded question, expressed in this logical notation, appeared as follows:

KEJ·*{KAM·(PUCR 1001,+SULL 1001,)·GARM,
         *(D[ ]CM,+RUGL,)·KQJ·SWLL...)}

The reader will note that, in addition to the symbolic punctuation already explained, three new symbols have been introduced in the foregoing notation. The period above the line (·) indicates conjunction, or "together with." The plus (+) symbolizes disjunction, or "or." The parentheses are used in the same manner as in a simple mathematical expression; thus, in the foregoing notation, they indicate that all of the informational elements following the first conjunctive symbol are to be taken as a group, while the informational elements between the second and third conjunctive symbols form one sub-group and the elements between the fourth and fifth conjunctive symbols form another sub-group.

With the encoded question expressed in the logical notation indicated, the computer was then programmed to identify any ab-

9. See text at p. 237 supra for a list of five words with similar logical elements.
10. See textual discussion supra following Table I.
CONCLUSION

The indexing system for legal literature described in this Article is by no means a total departure from traditional indexing, in that it depends upon a person skilled in the law to select and record significant aspects of subject matter for future retrieval. Yet this system achieves a depth and flexibility far exceeding the practical possibilities of printed indexes. The WRU pilot project represents a new approach to the application of computers to the problem of literature searching in the law.

The potential existence of magnetic tapes containing detailed indexes to statutes and cases opens up the prospect of a valuable adjunct to law libraries, both public and private. Copies of such tapes could be acquired by any firm or institution having access to a computer, or copies of these tapes could be maintained at various regional centers throughout the country. The lawyer desiring a search could take or send his question to a programmer, who would be either on the staff of his firm or on the staff of an information retrieval center. The programmer would translate the lawyer's question into "machine language," and the computer, after scanning the tape indexes, would print out the references which the lawyer should investigate in order to prepare his case.

What of the cost of mechanized literature searching? How would it compare with the cost of manual searching?

It appears that lawyers and law offices do not as a rule figure the exact time and cost of literature searching. Rather they compute their bills to clients on the total time required for a case. Therefore, a statement of the comparative costs of searching by mechanized or conventional means could be only a meaningless guess. The question of how much undiscovered information costs is not capable of a direct answer. In the area of scientific research, the magnitude of difficulty of the conventional literature search can be illustrated by the fact that one corporation estimated that if a project budget was less than $100,000, it was cheaper to run all experiments over rather than manually to search the literature for precedents. The lawyer does not have even this way out. He cannot experiment. Nor can he afford an exhaustive search. Therefore, at the present time, in most cases a lawyer must be satisfied with a search which will have a reasonable probability of being just a hairline more extensive than that performed by his opponent.

This hit-or-miss approach to precedent searches does, of course, have an economic basis; only a limited amount of funds can be
devoted to searches, regardless of how affluent the client might be. Also, it places a premium on the lawyer's skill in deciding where to stop in a search, rather than permitting him to review all of the material that may be pertinent and then to expend his time and professional skills in evaluating and adapting that material to suit the needs of a particular case.

The computer performs repetitive, routine tasks more thoroughly, at lower cost, and faster than human beings. Computers therefore can relieve the human being of such tasks and allow him to devote his full energies and time to the reasoning tasks which he, of course, performs far better than a computer.