A Numeric Semantic Code System for Universal Machine Use

PETER GLICKERT
1325 E Street, N. W., Washington, D. C. 20004

Received June 19, 1967

This paper describes the framework for a machine-usable semantic code not limited to any particular subject field, and the theories behind the code. The code is compared with the Western Reserve University Semantic Code System, and possible uses for the code in present retrieval systems are outlined.

A set of about 1500 code designations for English words and terms, to be used in machine recording and searching of verbal information, has been created. Universality of application is sought by hospitality of the codes to every concept expressed or expressible in the English language, and an extensive expansion of this set of designations is hoped for. The codes, about 1000 of which have been published (*1*), can be of use in certain indexing systems already in operation, as a substitute for present descriptor designations, and role indicators.

The codes were created as part of a possible future mechanized storage and retrieval system in which the full verbal text of any document will be recorded in the form of "microsentences" (*2*, projects 3.3, 3.76, 5.43, 7.32). Such procedures would use a grammar, which is not discussed in this paper, and a code system, the framework of which is presented herein, which makes the grammar possible. These procedures would, of course, also be applicable to recording and searching the abstracts of the documents, and the codes themselves could be used in mechanizing the indexes of the documents.

The codes discussed herein were drawn up using much the same process Perry and Kent used in drawing up the Western Reserve University Semantic Code (*3*), although the author was unaware of the details of their work at the time he constructed the codes. There are, however, major distinctions in the type of vocabulary selected for encoding, in the way the selected vocabulary was processed before encoding, and in the type of code applied to the resulting terms. These distinctions stem from differences in codification theory, the logic of which can be summarized as follows:

> A vocabulary of machine-readable code descriptors must be essentially unrestricted as to its size and the subject matter it covers if all topics of interest to, for example, the chemist are to be recorded for machine searching in a single system.
>
> If such a large vocabulary is employed it must be organized for indexing and searching uses (*4*).
>
> The provision of a semantic or faceted code is the most practical mode of vocabulary organization for machine searching (*3*, p. 242–243).
>
> Semantic codes for specific concepts are in reality definitions of these concepts (*5*, p. 41), the code definition being a combination of smaller and more generic code expressions.
>
> To make generic codes available for the development of specific codes, and for other reasons, codification must consider an ultimate generic level of concepts to be codified.
>
> Most of the concepts regarded in English as describing "things" are derivatives of verb concepts. The semantic relationship between such "things" is best developed by codifying the parent verb concepts semantically and designating the "things" as derivatives of the verbs.

The second and third points of the above summary are explained in the documents referred to. This paper attempts to explain the other principles and reviews the most important steps in the codification work. Italicized expressions herein, followed by an asterisk, will be found, along with their codes, in the Appendix.

NEED TO COVER ALL SUBJECT FIELDS

The information needs of the chemical profession are not confined to strictly chemical information, no matter how broadly or narrowly the term chemical may be interpreted; rather, the needs cover as many topics as are included in the Subject Index of the chemist's fundamental literature search tool, *Chemical Abstracts*. Thus, to include in a single machine-readable system all topics of possible interest to the chemist, the system would have to include machine readable codes for "quantum mechanics, electromagnetic radiation, electronic phenomena, photography, antibiotics, potential barriers, field theories and so on and on" (*6*). As it cannot be predicted what further fields of technology, or even the social sciences, may someday be of enough importance to chemists to justify inclusion in chemical journals, can any subject matter field be arbitrarily dismissed as not needing to find a home in a system developed to record the full information needs of the chemist?

Current projects call for the development of indexing systems in a number of specialized fields, including medicine, the sugar industry, meteorology (*2*, projects 5.24, 5.41, 5.63, 5.66, 5.71, and 5.73) and physics (*7*). Even though information needs in these different systems may vary (*2*, project 5.39), efficiency indicates that a single reference list of code designations would be desirable, from which codes specific to various scientific disciplines may be selected. Use in these systems of compatible machine-

readable code designations may avoid the need for future consolidation efforts (2, project 7.10).

So far as the preparation of documents for machine recording is concerned,

> "What are the advantages that a common descriptor language might bring? ... (a) It might avoid the duplication of document analysis; ... (b) It might avoid the duplication of semantic analysis; ... (c) The training of staff for retrieval systems and their transfer from one system to another would be facilitated if a common language were in use" (5, p. 67).

The WRU system (3) is among the best yet developed and has been widely tested (8, 9). It was designed originally to handle metallurgical information and is said to be applicable to other fields (3, p. 240, 285) although not purposefully designed for interconvertability (10) among all fields of technology. In the author's opinion, however, the recording, in machine searchable form, of all topics of interest to chemists using this or another current system would be beyond the theoretical capacity of the system. For one thing, the role indicators, which present systems use, appear to block application of any one system to a variety of subject matter fields (5, p. 68, 69).

NEED FOR A LARGE VOCABULARY

In a compilation by Vickery (5, p. 30, 31) of the vocabulary size of some operating machine retrieval systems, 41 out of 44 systems studied appear to be concerned with information expressible in terms of words, rather than chemical formulae. It appears to be a general rule of thumb that the larger the number of documents to be recorded in the file, the larger the vocabulary of descriptors used in the system. Vickery finds an almost constant ratio of documents recorded to vocabulary size. In 39 of 41 systems, this ratio is less than 30 documents per descriptor; in 30 of the systems, this ratio is less than 10. These figures tend to reinforce the viewpoint that a system for recording an unlimited number of documents should have an unlimited vocabulary.

Vickery has stated that a retrieval system can fail when certain concepts cannot be expressed, or when certain levels of specificity are absent (5, p. 174). Further, can *Chemical Abstracts*, with its "relatively little concentration of words into descriptors" (5, p. 47), be accurately recorded without a large vocabulary?

A large vocabulary can provide advantages both in the indexing procedure and in one phase of the search operation. In indexing with a limited list of descriptors, topics will often be found in documents which fall between the descriptors provided. The indexer must make a decision as to what available descriptor to apply to the word (5, p. 28). Every such intellectual decision which must be made by the indexer increases the probability of inconsistent choices on the part of a group of indexers. The larger the vocabulary, however, the more likely it will be that any word encountered by an indexer in a document will be a word already in the system. Further, the provision of a large vocabulary can improve relevance in searching, by including fewer concepts within each term of the formal descriptor vocabulary (5, p. 27).

A large vocabulary has certain disadvantages. A searcher may have too many terms to choose from and thus fail to ask for the correct term. The indexer may encounter more than one word in the descriptor list appropriate to a concept contained in a document, thus producing inconsistencies. The author believes that these disadvantages may be overcome more effectively by organizing the vocabulary, than by reducing its size. He proposes that the vocabulary of the retrieval system be as large as the number of concepts expressible in natural language.

VOCABULARY ORGANIZATION—THESAURI

Most present indexing systems use marks in the machine-readable file which are arbitrary—that is, in which a code for any particular term does not bear any relationship to a code for a related term. The system may arrange its index terms alphabetically and apply to each term a number, lower-numbered terms appearing earlier in the alphabetized list. The list of numbers may be nonconsecutive to allow insertion of additional terms into the system in their alphabetical place. Those systems which mark the tally with natural language terms, whether in full or shortened form, also are arbitrary, as there is generally little relationship in meaning between similarly or even identically-spelled words. Generally, such systems—for example, Medlars (11)— provide a thesaurus to inform the searcher (and also the indexer) of related topics to consider in searching and in indexing, thereby improving the recall of the system (4). Sometimes the relationship brought out by the thesaurus is a genus-species relationship and a hierarchical feature is introduced into the system to allow searching on selected levels of generality or specificity of topic. For example, the file can be questioned at a generic level to assure maximum recall (4), and if too many documents for ready review are revealed, further searches of the file or a subfile can be made with a more specific question to improve relevance. With a properly organized thesaurus, an indexer who encounters more than one descriptor suitable to a concept found in a document, can be guided to a generic term which subsumes the descriptors in question.

It is also possible to build thesaurus features into the mechanized programming of a system. In Salton's SMART System (12), where machinery is employed which has an extensive programming capacity, the hierarchical feature of the topics may be taken care of automatically in the machine program; when a specific term is entered into the record, the programming is such as to record on the tally the mark for a generic term as well as the mark for the specific term. Systems which use arbitrary codes, along with printed or programmed thesaurus lists may be considered externally organized.

VOCABULARY ORGANIZATION—SEMANTIC CODES

Semantic codes are internally organized. In a semantic code, related marks are put in the machine file for words having related concepts behind them. The use of a semantic code simplifies programming by allowing a specific

term to be recorded and a search to be conducted on a more generic level by omitting components of the specific code in the search question. A good example of such a situation would be one where a document is concerned with the effect of bromine on a certain organic compound. For bromine, the tally would be marked according to a code which provides a certain designation—for instance, *A*, for a nonmetallic element, a more elaborate designation, *AA*, for a halogen, and a still more elaborate designation, *AAB*, for bromine. Thus the recording for this document would be marked *AAB*, to represent bromine. A later search question asking about the effect on the same organic compound of a nonmetal would reveal this record, as would a question asking the effect on this compound of a halogen or of bromine. A search question specifically asking the effect on this compound of chlorine (code *AAA*), would not reveal this document.

SELECTING THE VOCABULARY

Arguments have been made both for and against the provision in descriptor vocabularies of very general descriptors. In regard to the bromine example given above, the most generic level supplied is nonmetallic element. Should the system supply a code designation for *chemical element**, to subsume the term nonmetallic element? Should there, in turn, be a designation for *chemical substances (entities)** in general?

Although Vickery has stated that "there is no ultimate elementary level which we can adopt, since the level must depend on the purpose of the retrieval system" (*5*, p. 42), the author believes, for a number of reasons, that an ultimate generic level of code designations should be available for use in a system when needed. By "ultimate generic level" of concepts is meant a set of concepts which cannot be subsumed by more general concepts. For one thing, Vickery states that a "system which aims to index the detailed contents of ... documents may need to move to a higher generic level" (*5*, p. 45).

Secondly, although code designations for very generic concepts may not often need to be used in themselves, they can be of inestimable value in providing a basis for drawing up semantic codes for related, more specific concepts. An example is given by Perry and Kent where the code maker is presented with the problem of encoding the term "lattice" (*3*, pp. 246 to 248). To encode this term, the closest existing code term "frame" was chosen as the foundation for extrapolation of a code specific to "lattice." It will be readily understood that a frame in its usual sense is a structure outside another structure, while lattice describes internal structure—in the given situation, of a crystal. The WRU code system, however, does not provide any term generic to both internal and external structures; therefore, a code term inherently including the connotation "external" was forced to serve in the code designation of an internal structure.

An ultimate generic level for English expressions was indicated by John Locke, in 1824, as follows:

> "All things that exist are only particulars ... Leave out the idea of animal, sense and spontaneous motion; and the remaining complex idea, made of the remaining simple ones of body, life and nourishment, becomes a more general one, under the more comprehensive term *vivens*. And not to dwell longer upon this particular, so evident in itself, by the same way the mind proceeds to body, substance and at last to being, thing and such universal terms which stand for any of our ideas whatsoever"(*13*).

Normally, in preparing a vocabulary for an indexing system, the art to which the code is to be applied is studied and the descriptor terms used therein are collected (*5*, p. 43). However, because it contains the most general of words, the author chose the Basic English Word List (*14*) as the starting point for the vocabulary to be used in the present codification. Permission was obtained from the copyright owners to use this list.

AMPLIFYING THE VOCABULARY

It is the author's conviction that the verb concept *exist** is generic or parental to both the "universal" concepts given by Locke in the above quotation. *Being** describes the result of the operation of *exist**, while *thing** is a purely functional description of an operator of *exist**. In the same way, the verb concept *eat** is parental to the noun concept *food**, etc. Such nouns may be considered "frozen" verbs, because in English, "(w)e are constantly reading into nature fictional acting-entities, simply because our verbs must have substantives in front of them" (*15*). This transposition of situations or functions (verbs) into entities (nouns) occurs not only with abstract concepts such as "justice" or "chemistry," but also in the description of many entities, usually thought of as concrete or picturable. For example, the concept of "catalyst" does not at all describe the structure of a material; it may be a solid, a liquid, a paste, a slurry or even a gas. The term "catalyst," therefore, describes only a possible use or function of a structurally undefined material in a certain total situation.

Most of the words of the Basic English Word List are presented as nouns, but to the author they appear to be frozen verbs which have to be thawed for proper organization of the situations or functions they describe before the nouns can be codified semantically. Thus, the words of the list were studied to supply parent verbs, such as *eat**, for nouns such as *food** and these verbs were added to the vocabulary to be codified.

It is conventional in drawing up a list of index terms to combine synonyms (*5*, p. 35)—that is, words which cannot be considered truly distinct from each other in meaning—and when such combining was applied to the Basic English Word List, the number of words in the vocabulary diminished somewhat. This combining blurred many distinctions in parts of speech (*3*, p. 226); in particular, prepositions, qualities, adverbs of manner, and some conjunctions, when combined with the verb "to be" were synonymous with verbs in the list or obvious verbs outside the list; for example, *accompany** is considered synonymous with (*be*) *near*. This discovery thus enabled all relational terms (*5*, p. 58) to be codified as verbs.

The list was also amplified to supply generic terms such as *ingest** for species such as *eat** and *drink**, and *shine**, which is generic to (*be*) *bright** and (*be*) *dark**.

Words were added to supply enough terms to make out detailed schedules for chemical elements, plant and animal species, and geometric shapes, so that the species, shapes, elements, and compounds contained in the list could be codified properly.

Certain words in the Basic English Word List are considered unencodible—namely, the pronouns and demonstratives (the, this, that, and such)—and the inclusion of these concepts in the recording of a document must rely upon the grammar of the retrieval system. These words were removed from the vocabulary, as was the verb "be," the meaning of which appears to depend completely upon the other words with which it is used.

CODE ALPHABET AND COMPUTER METALANGUAGE

In terms of computer technology, the metalanguage of the semantic code is extremely simple; however, a detailed explanation of the code alphabet (5, p. 110) will improve understanding on the part of those interested in how the codes were constructed. It was decided to use an alphabet of 128 binary bit positions—i.e., a field of 128 bit positions would be made available on the tally (unmarked record of the system, 5, p. 179), to accomodate each code word. One or more of these positions would be marked (punched, etc.) to designate a particular code word.

As can be seen from the Appendix, the codes are expressed in terms of the "1's" in this 128 bit position field to be occupied by the code. These code designations may be transliterated into any format suitable for use in the machinery to be employed in creating and searching the file. The codes are presented here in numerical terms, suitable for maximum utilization of storage space in an eventual operating system.

The codes can be visualized in the following way: a frame consisting of 16 columns (characters or bytes) of an eight-channel Flexowriter tape would be theoretically dedicated to each word. On the ordinary IBM card, likewise, 16 columns would be dedicated to each word, the columns being punched in the lower eight holes—i.e., the 2 to 9 holes. In the codes, therefore, a designation such as 3-9 may be visualized as meaning that the number 9 hole in the third column of the dedicated space is to be punched. The code 4-4,5,6,7 means that the holes number 4, 5, 6 and 7 of the fourth column of the space on the card are to be punched.

In terms of computer operation, the codes may be considered as composed of a series of major and minor characteristics which are separated by a hyphen, as 2-3 or 13-5. There may be a string of minors associated with any major. The string relationships mean conjunction (and) and are indicated by a comma. The space is used to denote conjunction between separate majors as 2-2,4 3-9, which means the species 2 and 4 of genus 2 and the species 9 of genus 3. Conditional relationships, or, *vel*, etc. only occur in preparing retrieval statements, not in the specification of a semantic code.

If the codes prove to be feasible, it is realized that a dedication of 16 columns to each code word would still leave the greater part of the tally unpunched. To conserve space in an eventual system, therefore, columns need not be dedicated; rather a column can be marked by a binary code in the upper four positions (A, B, 0, and 1) on the IBM card (or in the preceding column of a Flexowriter tape) with the number of the column to be designated. The four upper positions of the IBM card allow binary designations from 0 to 15 to be made, allowing 128 (16 × 8) bit positions to be used.

Of these positions, the 0- series is reserved for links and designations of other codes, such as line-formula notation system codes or alphanumeric codes. Thus a mechanized system using this code system is hospitable to the inclusion of information presented in any other code system in the same file. The 15- series is reserved for the designation of links, punctuation, symbols for document identification, etc. All simple codes so far constructed have fit within the 128 bit positions, but no theoretical reason has been found for not being able to continue a long code word in a second group of 128 bit positions. Eventually, of course, a vocabulary tape can be supplied so that natural language words and expressions, which have been codified, automatically will be converted to code form for entry on the tally, or for questioning of the record.

In the Appendix, the words themselves are often presented as a group of synonymous expressions, with single words being given first, in alphabetical order, followed by synonymous terms. The word "be" is ignored in alphabetizing. The codes thus are often expressed as being at the intersections of meaning of two or more words or terms and certain concepts are excluded from a particular code by using the device "qv" (which see). In this way, the author hopes to limit the ambiguity, inherent in most natural language words, to proportions which are manageable in machine storage and retrieval of verbal information.

FIVE PRIMORDIAL SITUATIONS

The first words selected from the expanded Basic English Word List used as the vocabulary for codification were the verbs. All verbs could be subsumed under one of five primordial terms: *exist**, *act**, *affect**, *change** and *control**. Since for a first thing or subject to change or control a second thing or object, the subject must affect the object; since to affect an object the subject must act; since to act the subject must exist, it was clear that a hierarchical arrangement of these terms was called for, so that a question such as "does A affect B" would be answered by a statement in the file that "A changes B." The resulting scheme is presented in Table I.

These primordial situations are expressed, as can be seen, in codes containing the 4- series of designations. The development of Table I illustrates the procedure used throughout the codification work. The words in the original verb list were distributed among five lists named for the five primordial levels of activity.

Table I.

Concept	Code
exist	4-4
act, do	4-4,5
affect	4-4,5,6
change	4-4,5,6,7
control	4-4,5,6,8

MOST FREQUENT SUBJECT AREAS

After distribution of the verbs into five lists according to the five primordial levels of activity, each list was studied to find the most general subject areas to which each word therein pertained. No hard and fast rule was adhered to as to whether these subject areas should be considered entities or activities. As can be seen, these areas can sometimes be expressed in verb terms, but usually a substantive (noun) term is employed. Those codes found in the Appendix which contain the 3-9 indicator are nouns which were given their codes after the first verb schedule had been completed. The most frequent areas were energy, psychology, internal relationship, external relationship, *time** and *space**. These areas have the designation 5- in the codes of the Appendix.

As the words in each general area were studied, it appeared that enough words were present in each list to justify creation of subsumed more specific areas. Thus, under energy, *light** and *sound** were involved in enough words to justify creation of special categories; under psychology, *emote**, *decide**, *know**, and *communicate** were given special designations; under internal relationship, *amount**, *composition**, *surface**, and *live** were designated; while *have** became a species under external relationship.

Words in each of the five lists were given like designations for like subject areas. For example, *have** and *acquire** were both marked with the designations for *have**, showing that *have** is to *exist** as *acquire** is to *change**. Thus, the question, "does A have B?" would be answered by a statement in the file that "A acquires B." Also, this marks a break from a strictly hierarchical designation in which a term like *have**, subsumed under *exist**, would only by accident have a code related to that for *acquire**, subsumed under *change**, even though the mode of subsumption in each situation is alike. It was later found that this study is essentially that described by Vickery as a study of categories (*5*, p. 45).

Upon further consideration, many verb concepts covered in the Basic English List, while well-related to named subcategories had a relationship to the main category which was remote, tenuous, or confusing. In other instances the inclusion of the generic designation was superfluous or restricting. For example, the term *observe** involves *light** (originally subsumed by energy), *know** (subsumed by psychology), *live** (subsumed by internal relationship), and external relationship. It was decided that the indicators specific to *light**, *know**, *live**, and psychology were sufficient to distinguish this term from others and that the connection of the term *observe** with generic categories other than psychology was tenuous. As another example, certain kinds of communication have nothing to do with psychology, etc. It was decided, therefore, to give independent, correlative designations to these subcategories and down-grade the original major categories. The original subcategories appear in the 6- and 7- series of the codes in the Appendix.

To these, the category *direct**, *orient** was later added, along with a designation for *accompany**. This latter designation is a composite of the designations for external relationship, *time**, and *space**. Further along in the work, it was found that the terms *line** and *hollow** are useful designators and these categories may in the future be removed from subsumption under the genus *construction** The original generic category designations are maintained in codes only when considered clearly pertinent to the definition of the word encoded or when necessary to distinguish two somewhat related terms; for example, *construction** is a species of *composition** which is confined to spatial matters. Therefore, it is codified with the designators for *space** and *composition**. Likewise, *chemical element** is distinguishable from *chemical compound** by the use of the old internal relationship designator in the former and the old external relationship designator in the latter. The generic term for *chemical entity** lacks both these designators. It remains to be seen by experimentation whether the elimination of these generic categories detracts from the usefulness of the code system.

It was originally intended to let the 8- series of bit positions be used for arbitrary species designations, comparable to the Arabic numerals in the Western Reserve University semantic code (*3*, p. 233), the 9- series being used for subspecies, the 10- series for still further subspecies, etc. In practice, some designators in the 8- and 10- series have acquired semantic significance, while all of the 9- series except 9-2 and 9-3 have been assigned definite meanings.

Many words such as *abet**, *help*, and *guide** carry a connotation of good (favorable) or bad (unfavorable). Others have also found a need for such designations (*2*, project 3.85). Two types of negative are used: a simple absence or lack, and a direct opposite—compare *have**, *lack**, *acquire**, and *lose**. Also, many words are extreme versions of other words; thus, *coincide**, *be together* is an extreme or emphatic form of *accompany**, *be near*, while *own** is an extreme form of *have**. Also, a need was found for reflexive or correlative designations—for example, *become** is a reflexive form of *change**, while *hang** and *rest on** are correlative forms of *bear**. Code designations for each of these factors are provided in the present 9-series.

After codification of the verbs, nouns found to be frozen verbs were codified as derivatives of the verbs. Those entities considered to have an existence independent of any situation, such as *light**, *space**, *time**, *composition**, were codified without verb designations, but using the same spectrum of code designators.

After nouns in the list had been codified, it was found possible to refine the codes applied to some verbs. Thus, *eat** and *drink**, which had been arbitrarily distinguished as species of *ingest**, were recodified, including the codes for solid and liquid in their codes. Quantities, including numbers up to *nine** were codified semantically in a separate but interlocking schedule. Not every meaning of every word in the vocabulary has been codified, but this will be accomplished as the work progresses.

MECHANICAL SORTING

After all of the words in the vocabulary had been codified, the words and codes were transferred to specially designed McBee cards and the edges of the cards were notched to conform to the code thereon. A card was

used for each code word contained in any code expression.

After the notching of the cards, they were sorted according to each code designator to which a definite meaning had been assigned, and subsorts according to combinations of these designators were made. In this way, thesaurus groups and subgroups were created for any of the conceptual threads evident in the codes. The sorting revealed a number of duplicate codes and such were revised to assure unique conceptualization for each code. After revision, the cards were prepared for printing in the alphabetical order of the words. The design of the notched card permitted photographic reproduction directly from the cards themselves.

THE WHEN-TO-STOP PROBLEM

As mentioned above, a semantic code is considered the most practical organized code system for machine use and a specific code in such a system amounts to a definition--that is, a combination of generic codes, representative of the words in the definition of the specific concept. In drawing up such definitions however, there is the problem of deciding when to stop (*5*, p. 42); that is, how many of the properties or qualities associable with a concept should be referred to in the code? Even in making a word definition of a concept this problem appears (*16, 17*). Newman (*18*) has outlined a large body of concepts which can be associated with the concept "dog." How many of these concepts should be used in making up a code definition for the word?

One principle which can be here applied is to restrict the code definition to those designations which do not increase the information content of the word or code unjustifiably (*5*, p. 150). For example, bromine can be assigned to many different classes, depending on the special interests of index users: oxidizing agent, radio opaquing agent, monovalent ion, gaseous element, etc. Would the inclusion of designations for all of these properties increase the information content of the code unjustifiably?

As my guide for when to stop, I used, rightly or wrongly, Aristotelian notions of substance and accident. Those properties which a thing or situation must necessarily have in order to be what it is (its substance) are included as designators in the code description; whatever else may be associated with the concept (accidents) depends upon its synthetic relationship (*3*, p. 101) in the document to be recorded. Thus the code for bromine should include the idea of actual or potential monovalence because this is one of the distinguishing features of this material; it helps to define the substance of the material. The other properties recited above depend on circumstances external to the bromine: its state depends upon the temperature and pressure; whether it is an oxidizing or reducing agent depends upon the other materials it is near; its opaquing properties depend upon the type of radiation employed.

The total information system should allow for recording of particular circumstances recited in a document; however, the author believes that such recording should be a function of the grammar of the system, not of the code designations, just as in English the recounting of particular circumstances is generally a burden upon the sentence and other grammatical features of the system, rather than upon the root word itself (*19*). As for the inclusion in the recording of a document of information not explicitly recited, this is believed to be an editorial function beyond the scope of the mere retrieval system (*3*, p. 97). Whether a mechanized system will ever be able to read between the lines and whether a true retrieval system can operate without such reading are topics worthy of investigation. However, I do not believe that all progress must wait upon definitive answers to these questions; rather, I believe that full recording of explicit information given in a document is a worthy goal.

USES FOR CODES IN PRESENT INDEXING SYSTEMS

Perhaps the most important contribution which the code system can make toward indexing systems presently used in areas of interest to the chemist is the provision of the verb codes as substitutes for presently used role indicators and other relational terms (*5*, p. 58-60). Tests appear to prove that role indicators are responsible for many of the problems of current indexing systems (*4, 9, 20*). I feel that many such errors may stem from the smallness of the role indicator vocabulary in many systems and from the failure of this vocabulary to offer generic-specific choices in searching.

I believe also that substitution of codes from my system for presently used descriptor codes can provide a number of advantages. Of course, it can make a file of arbitrary codes into a file of semantic codes and thereby simplify the questioning of the file and make it possible to conduct more generic searches.

Where a semantic code is presently used, I believe my codes may be more complete and may provide greater recall. The unfortunate tendency of English to categorize entities in terms of their function may have misled code-makers into considering the functional word as a complete description of the entity, ignoring structural properties in making up the code. This appears to be the reason why a search for information about the manufacture of gears, using the WRU code, did not reveal a document concerned with the manufacture of sprockets (*8*), although the two devices are intimately related, structurally. Preferably a code system should designate both structure and function and be able to distinguish between the two. By providing verbs and by expressing functional terms as derivatives of verbs, the verb codes may lead to codes having clearer meanings. Also, the resolution of presently used descriptors into codes which designate structure as well as function is believed to be a key in avoiding mistakes due to obsolescent terminology (*21*).

CONCLUSION

When the codification work described above is compared with the steps outlined by Vickery (*5*, p. 60, 61) for vocabulary control in descriptor languages, the construction of the codes in question followed steps 4, 6, 7, 8, 9, 11, and 13. Steps 3, 5, and 10 of Vickery were substituted for by providing for inflections of words, as also provided for in the Western Reserve University Semantic

Code. Vickery's steps 14 to 17 can be taken care of in the construction of the proposed grammar for the machine language.

The construction of the codes required the making of many specific decisions as to what particular English words mean, what other words they are related to in meaning, the degree and importance of such relationships, and how the important relationships can best be portrayed in a mechanical system which cannot have any understanding of the terms comparable to human understanding. The mechanical sorting procedure described above served as a test to verify these decisions and I believe that the mechanical creation of the thesaurus lists proves most of the decisions to have been correct.

The work was undertaken partly to see whether a semantic code could be constructed for a very general vocabulary of English words not restricted to a particular field of technology, or to technology at all. This has been proved and a framework can be established within which codes for all more specific words can be included, thereby providing a code vocabulary for machine recording and searching of information in any subject field. Such a vocabulary would, of course, cover all subject fields of interest to chemists.

APPENDIX I

abet, aid, help				4-4,5,6	9-4
accompany, be close, be near *qv* abut, about, bring, communicate				4-4	5-6,7,8
acquire, get, obtain, receive, take possession of				4-4,5,6,7	7-2
act, do					4-4,5
affect, be instrumental					4-4,5,6
amount, degree, quantity *qv* dimension, level number, some				3-9	6-3
bear, support			4-4,5,6	5-6,7,8	6-9
become, change into				4-4,5,6,7	9-9
being, existence				3-8,9	4-4
beverage, drink, liquid (which can be ingested		3-9	4-4	5-8	6-4 9-6,7
	4-3,4,5,6	5-6	6-3,9	7-3	10-8
(be) bright			4-4,5	6-7	9-7
change					4-4,5,6,7
chemical, chemical entity			3-9	6-4	8-7
chemical compound		3-9	5-6	6-4	8-7
chemical element		3-9	5-5	6-4	8-7
coincide, be together, be with, be at			4-4	5-6,7,8	9-7
communicate			4-4	5-6,7,8	7-7
communicate, inform, tell		4-4,5,6	5-4,6	7-5,7	
composition, identity, nature				3-9	6-4
construction, shape, structure			3-9	5-8	6-4
control, direct, operate *qv* orient					4-4,5,6,8
controller, guide, manager			3-9	4-4,5,6,8	
(be) dark			4-4,5	6-7	9-6
decide				4-4,5	7-6
direct, orient				4-4,5,6	6-9
drink, ingest liquid	4-4,5,6	5-6	6-3,9	7-3	9-9
	10-8	3-9	4-4	5-8	6-4 9-6,7
eat, ingest solid	4-4,5,6	5-6	6-3,9	7-3	9-9
	10-8	3-9	4-4	5-8	6-4 8-5
emote			4-4,5	5-4	7-4
exist					4-4
food, solid (which) can be ingested			3-9	4-4	5-8 6-4
	8-5	4-3,4,5,6	5-6	6-3,9	7-3 10-8
guide, manage *qv* lead				4-4,5,6,8	9-4
hang, depend from	4-4,5,6	5-6,7,8	6-9	8-3	9-3,9
have, possess				4-4	7-2
(be) hollow *qv* empty			4-4	5-8	6-4 10-8
ingest	4-4,5,6	5-6	6-3,9	7-3	9-9 10-8
know				4-4	7-5
lack, be without, not have, be free of				4-4	7-2 9-6
light				3-9	6-7
line, fiber, straight line *qv* direction		3-9	5-8	6-4	8-5 10-7
live, be alive *qv* abide				4-4,5	7-3
lose			4-4,5,6,7	7-2	9-8
nine				2-4,7,9	3-6
observe, see, view, watch		4-4,5	5-4	6-7	7-3,5 9-9
own				4-4	7-2 9-7
rest on, be on	4-4,5,6	5-6,7,8	6-9	8-3	9-2,9
shine				4-4,5	6-7
sound				3-9	6-6
sound, give off sound waves				4-4,5	6-6
space				3-9	5-8
surface				3-9	6-5
thing				3-9	4-4
time				3-9	5-7

LITERATURE CITED

(1) Glickert, P., "A Codification of English Words," Washington, D. C., 1966.

(2) "Current Research and Development in Scientific Documentation," No. 14, National Science Foundation, Washington, D. C., 1966.

(3) Perry, J. W., and A. Kent, "Tools for Machine Literature Searching," Interscience, New York, 1958.

(4) Montague, B. A., "Testing, Comparison, and Evaluation of Recall, Relevance and Cost of Coordinate Indexing with Links and Roles," *Am. Document* **16**, 201 (1965).

(5) Vickery, B. C., "On Retrieval System Theory," 2nd ed., Butterworths, London, 1965.

(6) Bassett, L. G., S. C. Bunce, A. E. Carter, H. M. Clark, and H. B. Hollinger, "Principles of Chemistry," p. 4, Prentice-Hall, Englewood Cliffs, N. J., 1966.

(7) American Institute of Physics, Documentation Newsletter **7** (3) December, 1966.

(8) Rees, A. M., "The Cleverdon-WRU Experiment: Search Results," in "Information Retrieval in Action," A. Kent, Ed., The Press of Western Reserve University, Cleveland, 1963.

(9) Herner, S., F. W. Lancaster, and W. F. Johanningsmeier, J. CHEM. DOC. **5**, 92 (1965).

(10) Vickery, B. C., "Coding for Interconvertability," p. 1167, "Information Retrieval and Machine Translation," A. Kent, Ed., Interscience, New York, 1961.

(11) "Medical Subject Headings," *Index Medicus*, **6**, No. 1, Part 2 (1965).

(12) Salton, G., "The Evaluation of Automatic Retrieval Procedures, Selected Test Results Using the SMART System," *Am. Document* **16**, 209 (1965).

(13) Locke, J., "Of General Terms," pp. 176-77 in"The Language of Wisdom and Folly," I. J. Lee, Ed., Harper, New York, 1949.

(14) Richards, I. A., "Basic English and Its Uses," Norton, New York, 1943.

(15) Whorf, B. L., "Languages and Logic," in I. J. Lee, Ed., *op. cit.*, p. 283.

(16) Pascal, B., "On Definition," in I. J. Lee, Ed., *op. cit.*, p. 138

(17) Korzybski, A., "Fate and Freedom," in I. J. Lee, Ed., *op. cit.*, p. 345.

(18) Newman, S. M., "Storage and Retrieval of Contents of Technical Literature, Nonchemical Information," p. 6, Second Supplementary Report, Patent Office Research and Development Reports No. 12, United States Department of Commerce, 1958.
(19) Richards, I. A., and C. Gibson, "Interpretation," in I. J. Lee, Ed., *op. cit.*, p. 162.
(20) Cohen, S. M., C. M. Lauer, and B. C. Schwartz, "An Evaluation of Links and Roles as Retrieval Tools," J. CHEM. DOC. 5, 118 (1965).
(21) Riddles, A. J., "Computer Based Concept Searching of United States Patent Claims," IBM Technical Report, ITIRC-004, Yorktown Heights, N. Y.: Thomas J. Watson Research Center, 1965.

Computer Assisted Primary Index Preparation

C. J. MALONEY, S. BRYAN, and M. EPSTEIN
National Institutes of Health, Bethesda, Md. 20014

Received June 28, 1966, and September 21, 1967 (Revised)

A man-machine primary indexing system employing a special "indexing language" called BICEPT is fully described. The system is applicable for computer-processed or non-machine-readable text. It differs from most other machine indexing projects in that the specification of index terms is largely manual. The method was applied to a journal article in an operational test.

PRIMARY *vs.* SECONDARY INDEXING

The use of various forms of indexes to find or to relocate desired information is of long standing and wide application. It is not generally appreciated, however, that indexes are employed in two quite distinct circumstances which we are here distinguishing by the terms "primary" and "secondary." A primary index is an index to a single document or series of documents issued from a single source (or cooperating group of sources) and normally bound with or at least an item in the collection indexed. It will, of course, be indexed at the source, and there will be no acquisition or language problem, even to the extent of technical jargon or drift in meaning of terms over time. The index in the back of any book, or the index volume of a set of encyclopedias would be examples. By secondary index we mean an index normally providing access to a multiplicity of documents issued from a variety of sources and over an extended period of time, usually planned to extend into the future. The index to *Chemical Abstracts* and the *Grants Index* of the National Institutes of Health are two examples. An annual or decennial index to a newspaper or journal could fall in either class, but is usually intermediate between them.

A secondary index answers the question: "which (if any) documents discuss the subject of....?" A primary index answers the question: "where (if at all) in *this* document is a statement, however brief, disguised, and/or trivial, made on the subject of....?"

A number of manuals and one journal (*1*) are specifically devoted to preparation of primary indexes, though these are more likely to stress similarities than differences between the two classes of indexes. A fairly full comparison of these two related but distinct problems is given elsewhere (*2*). A most important distinction, however, which makes discussion of computer assisted preparation of primary indexes especially timely is that, increasingly, the full text of the item to be indexed is available in machine-readable form. For example, the American Chemical Society plans to produce all of its journals by computer by the early seventies (*3*). The full impact of this now well-established trend is being appreciated only slowly. Thus Baxendale (*4*) in her otherwise excellent review of the 1965 status of "content analysis, specification, and control" includes "the current cost of providing computer-readable input" in the cost of primary indexing.

Another distinction relates to the fullness of indexing in the two situations. Judgments concerning the desirability of full indexing for one purpose cannot be successfully based on considerations arising in the uses for the other, yet until the two applications are distinguished there is a not unnatural tendency to do so. Judged by standards applicable to a secondary index, a primary index will appear unduly prolix. Its merits can only be appreciated by one who has spent hours in attempting to relocate a buried item in a book, journal, or newspaper, which he is sure he read but wishes to relocate. *The authors feel it is quite safe to assert that, when very full indexing is technically feasible to supply, the demand for doing so will be found to exist, as was true when computers made extensive computation possible, or when the aeroplane became efficient for long distance travel.* Indeed, certain authors (*5*, *6*) have considered it worthwhile to study the profitability of storing full text and then scanning the total store in answer to every query. Such an approach may be justified in case the total potential of the file must be exhausted upon enquiry—say, in a tactical military situation. It is quite possible that, in certain other