Patent Office
Research and Development
Reports . . . . . No. 9

NEW YORK STATE LIBRARY
OCT 14 1970
GOVERNMENT DOCUMENTS

Newman, Simon M

# LINGUISTIC PROBLEMS IN MECHANIZATION OF PATENT SEARCHING

029.9608
U6H6
73-2757
no. 9
Pam

*Prepared by*

**Simon M. Newman**
***Staff Member***

**Office of Research and Development**
**Patent Office**

**December 17, 1957**
**Reprinted January 15, 1959**

NEW YORK
STATE LIBRARY
ALBANY

**U. S. DEPARTMENT OF COMMERCE**
**Lewis L. Strauss, *Secretary***

**PATENT OFFICE**
**Robert C. Watson, *Commissioner***

For sale by the U. S. Department of Commerce, Washington 25, D. C. Price 10 cents.

# Index


| | Page |
|---|---|
| Foreword | 4 |
| Introduction | 5 |
| Coding Terminology | 5 |
| Unambiguous language | 5 |
| Drawings and Illustrations | 5 |
| Static Structures | 5 |
| Functions | 6 |
| Apparatus | 6 |
| Chemical Compounds | 7 |
| Other Word Classes | 7 |
| Linguistic Aspects of Encoding | 7 |
| Categorizing for Generic Searching | 7 |
| Ambiguities | 8 |
| Implied Conceptual Facets of Terms | 8 |
| Mechanized Encoding | 9 |
| References | 9 |
| Bibliography | 9 |




A73027E3A

## Foreword

The substance of this report has been developed through consultation with staff members from the offices of Research and Development and Classifications Operations of the Patent Office, and from the Data Processing Systems Division of the National Bureau of Standards. For the helpful criticism of many contributors, then, the author is grateful. Particular appreciation is due the Staff Director for his advice and sympathetic encouragement.

The responsibility for the organization and presentation of these controversial materials, however, is the author's alone. Accordingly, he welcomes illuminating comment from any source in this or related fields of interest.

*Simon M. Newman*

# LINGUISTIC PROBLEMS IN MECHANIZATION OF PATENT SEARCHING

## INTRODUCTION[1]

Linguists have taken an interest in the mechanization of Patent Office searching. It is clear that we will need their talents and their methodology if mechanization is to be achieved. In order that we may utilize their skills efficiently, our peculiar linguistics must be projected for analysis. It also seems advisable to collect in one place the many references which have been made to our language problems and to supplement these references with additional material which may guide the interested reader.[2]

## CODING TERMINOLOGY

### Unambiguous Language

The necessity for stating unambiguously the "notions" or phenomena present in patent disclosures has been reviewed, and one proposed system has been described in some detail.[3]

Any system for the mechanization of searching will require the formulation of unambiguous terms. Each term will serve either as a code itself or as a designation of an unambiguous code. *Coding* is thus the process of creating specific unambiguous terms as substitutes for "notions." *Encoding* is the general process of preparing a document for storage by the application of these code terms.

### Drawings and Illustrations

A patent disclosure always includes a verbal specification, followed by a series of claims which define the limits of the invention. Where the nature of the invention permits, the disclosure must include a drawing to which the verbal specification refers. The majority of patents include drawings. Usually they illustrate the structure of the device being claimed; however, they may include or may be a graph, a flow sheet, or a circuit diagram.

If no drawing is included, a patent searcher must read the specification, and his search, though much more detailed, is similar to any other literature search. However, when the patent includes a drawing, a search, whether generic or specific, almost invariably involves a visual examination of the drawing, rather than an examination of the printed specification. In such a search, spot reading of the specification is usually restricted to resolving ambiguities raised in the searcher's mind about some detail of the drawing, or to determining the uses to which the disclosed structure may be put.

The use of drawings rather than the verbal specifications creates one of the basic language problems in the mechanization of searching. The adage that "a picture is worth ten thousand words" particularly applies to the use of patent drawings in manual searching. Complex shapes, interrelations of the functions of their constituent elements, and details of their topological orientation often can be comprehended at a single glance. Information from graphs can usually be assimilated faster than that from the equations which generate them. Circuitry, both electrical and hydraulic, can be followed quickly and accurately when shown in line drawings. Flow sheets--though technically not drawings--serve to abstract the essence of complex processes and hence often can eliminate a tedious and unnecessary study of a specification.

Some drawings have details of shapes and their interrelations which are clear and unambiguous, even though they are not described in the text of the specification. These illustrated details are as valid for Patent Office search purposes as a verbal text describing them. Any coding system must provide for the formulation of terminology embracing all such details. Although it may be possible to store drawings, as such, in a machine memory, any search request might be made either in linguistic terms or in the form of illustrations. Therefore any disclosure so stored also must be encoded in linguistic terms.

### Static Structures

Static structures can be encoded solely in terms of the size, shape, and topological (orientational) relationship of their parts. The terminology drawn from Geometry and Trigonometry will serve to code such factors. Possibly other terms, such as *fillet, joint, lamina,* can be defined unambiguously. Most names now commonly employed to describe objects, however, are not helpful in uniquely describing their structure, since they are usually either functional or descriptive of some incidental property of the structure.[4]

One complication in structural encoding occurs if two or more well-recognized organizations of parts have one element in common. The occurrence of this one part in two separate organizations requires some encoding principle which will allow retrieval of this part in either organization or as a common part of both.

The problem of coding interrelations has been exhaustively analyzed, and one solution has been proposed.[3]

**Functions**

As implied above, a new terminology of functions (uses) of structures is also needed. These terms must be directed *not* to the disclosed *accidental* use, but to some expression of *basic* use defining what has been called the *necessary* or *proximate function.*[5] Terminology derived from proximate function can best be illustrated by one of the very few situations in which such concepts already have been defined. For example, in an analysis of a series of patents directed to methods of shaping devices from metal pieces, it was determined that there were only three basic mutually exclusive methods:

(1) *Assembly.*—The addition of some extraneous material to a single unitary structure; e.g., riveting or welding two girders together.

(2) *Parting.*—The removal of some material from a single unitary structure; e.g., cutting, punching, drilling, turning, etching or sawing.

(3) *Reshaping.*—The change of physical dimensions of one unit without assembly or parting; e.g., rolling, forging, coining or bending.

Shaping of this sort comprises one form of manufacturing.[6] By the addition of two other mutually exclusive methods, it would appear that the entire field of present manufacturing can be encompassed. These two groups are:

(4) *Quantum fluctuation.*—The so-called "changes of state" of matter among gases, liquids, and solids; e.g., melting, condensing or sublimation.

(5) *Generation.*—The creation of new things by atomic or sub-atomic recombination; e.g., chemical reactions resulting in precipitation, or atomic fission and fusion.

In such a broadened set of categories, one would also include in *assembly*, the filling of a mattress with felted cotton; in *parting*, the tearing of the end of a cigarette package; and in *reshaping*, the molding of clay. Other manufacturing processes, of course, may include combinations of these classes. For example, the baling of hay is both assembly and reshaping.

These five classes of manufacturing do not, of course, exhaust function terminology. Many other processes, including measuring, testing, transporting, transmitting of electrical energy, modifying of conditions of pressure and temperature, and projection of optical images must be analyzed likewise.

**Apparatus**

Apparatus illustrated in patent drawings include (1) those consisting solely of static structural parts, (2) those which include one or more parts which may be removed and reassembled from another part, (3) those including one or more incidentally movable but independently operable parts, (4) those which include one or more series of interconnected and usually intercontrolled movable parts, and (5) combinations of one or more of these four groups.

An example of a purely static structural apparatus of group (1) is the conventional core-type automotive radiator. This same radiator may have as an adjunct a cap that may be removed for filling the core and then be reassembled, thus exemplifying a combination of groups (1) and (2). This radiator might also disclose a simple plug valve at its lower portion, movable to draining position for emptying the core. This disclosure then would illustrate a combination of groups (1) and (3). A disclosure of the body portion of a fountain pen would include interrelated elements which when operated constitute the filling mechanism. This is an example of groups (1) and (4) in combination.

The coding of static structures has already been considered. But the creation of additional terminology for similar encoding of both the independently removable and the interrelated moving parts with each other and with static structures will be necessary. In the solution to this part of our problem, it is likely that some coding principle can be evolved which is analogous to *proximate function*, the principle previously suggested as governing one choice of manufacturing process terms. However, the proximate function principle itself definitely is *not* applicable to apparatus, for although the processes of forging and rolling are closely similar, a forging press and a rolling stand are entirely different apparatus; they are correctly described only in terms of the organization of their parts, and *not* by what function these parts may accidentally perform.

For some time it has been clear that both the nineteenth century basis of classification—that of material worked on[7] and the later used basis of an accidental function of the apparatus[8] have proven ineffective in segregating into classes those disclosures which are pertinent to normal search requests. It has been suggested recently that one basis for categorizing manufacturing apparatus is the relative movement of tool (or its holder) with the work (or its holder). This approach is now being utilized in a reclassification project involving cutting machines.[9] Some of the proposed first line (unindented) titles[10] of subclasses, arranged in the order of decreasing complexity, are found in Figure 1. It seems clear that apparatus having tools other than cutting tools can be similarly categorized, and that machines falling into such categories have similar characteristics. It is possible that this approach will offer a key to the solution of one part of this problem.

All apparatus, of course, does not relate to manufacture, nor does all manufacturing apparatus utilize tools. Terminology is also needed for a host of other devices, such as computing, projection, and transport apparatus.

Tool engaged work during dwell of intermittent work feed

Cutting motion of tool has component in direction of moving work

Transverse cutter with motion normal to running length work

Interrelated tool feed and work guide moving means

Interrelated tool feed means and means to actuate work immobilizer

Interrelated tool and work feeding means

CUTTING APPARATUS CATEGORIES

Figure 1

It has been postulated that apparatus terminology should not be derived from either the process performed or the material worked upon. In the process of encoding, however, this conclusion does not preclude the use of additional descriptors of process performed or material worked upon, provided that these descriptors are drawn from the unambiguous terms previously created to describe the material or the process.

### Chemical Compounds

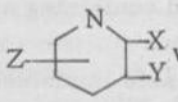
In the field of chemical compounds several sets of terms and rules for their application are in use. Although agreement on these rules has not been achieved, there is substantial agreement about the basic source from which terminology is drawn. The resolution of an unambiguous terminology for chemical compounds rests primarily on standardization. However, some or all of the definition problems previously discussed are present in chemical structures. For example, in complex organic compounds we find that one or more elements may be common to two or more parts of a compound, e.g., the common carbon atoms in two fused rings. Occasionally one or more atoms may resonate between two separate points of attachment. Solutions to some of these problems in one field may well result in their resolution in the other field.

In addition, there are many instances in which it is desirable to designate a class for which there is no existing term. The patent profession has resorted to a logical artifice in postulating this class as a restricted form of a more inclusive class. This artifice is known as a *Markush*[11] expression. For example, it may be necessary to refer to a class which includes less than all known acids, e.g.:

An acid selected from the group consisting of carboxylic acids and sulfonic acids.

In this example the two specified types are *not* intended as alternative species of the more inclusive genus *acid*. Rather, the complex of all the characteristics common to these two species constitute the characteristics of that class for which no single term exists.

A more complex Markush expression designates the desired class in terms of a structural formula, with artificial class designations of some or all of its constitutents, e.g.:

A compound of the formula  wherein

X is a member selected from the group consisting of -OH, -Cl, and $-CH_3$; Y is a member selected from the group consisting of -H, $-CH_3$, and $-CH_2-CH_3$, and Z is a member selected from the group consisting of -SH, $-SO_3H$, and -CNS.

If this were the form of a retrieval question it should be answered by any one of the 27 members of the class stated in the question. In the development of an encoding scheme for recording such artificial genera, both the inclusiveness of the defined class and the identity of the members constituting that class must be preserved.

### Other Word Classes

In addition to those general classes of words which have been discussed, there are many other classes which occur in expository prose which must be considered in creating a comprehensive encoding procedure. There are qualifiers and quantifiers, both of which have been elsewhere considered[12].

At present such entities as size, time, mass, and temperature, are each measured by diverse standards. For each form of measurement there must be concurrence in a single, unambiguous standard. Distances of .01 millimeter and of 10,000 light years (1 light year = $6 \times 10^{12}$ miles), for example, may both occur in a single disclosure relating to astronomy. Color is sometimes described by hue and sometimes by wavelength. Ratios of numbers and ranges of values occur frequently in disclosures. Other classes of terms (many with specialized problems) undoubtedly will be noted by specialists in other fields.

## LINGUISTIC ASPECTS OF ENCODING

### Categorization for Generic Searcning

Any notion or phenomenon may be described by one or more facets of its occurrence. The allegory of the blind men and the elephant illustrates this concept. A particular chemical compound may be correctly described by the term "alkane" or by the expression "saturated hydrocarbon." A brief case is a "leather article," a "piece of luggage" and a "compartmented receptacle." Any chosen term may be subsumed under myriad classes. A few of the diverse classes under which the term "pencil" might be subsumed and some of the other members of these classes, for instance, are listed in Fig. 2.

| OTHER CLASS | OTHER CLASS MEMBERS |
|---|---|
| Writing equipment | Typewriter<br>Fountain pen |
| Wood containing articles | Desk<br>Hammer handle |
| Portable implements | Hoe<br>Eyeglasses |
| Artist's equipment | Sketch pad<br>Pallet |
| Coating applicators | Brush<br>Salt-shaker |
| Envelope openers | Knife<br>Scissors |
| Indicators (i.e., pointers) | Flash light<br>Finger |
| Pocket tearers | Silver dollar<br>Key |

PENCIL

Figure 2

The effective use of coined unambiguous terms in an information retrieval process requires that they be categorized in a plurality of overlapping hierarchies, both for the purpose of encoding a disclosure specifically and generically, and of allowing flexibility in formulating a retrieval request in the same manner. When one has a pencil before him, it is not difficult to perceive many of the generic classes to which a pencil may belong. Nor is it more difficult, given a series of such broad generic terms, to determine other members in each of the series.

At the present time an examiner may formulate a retrieval request from a claimed disclosure of an article in one of these "other" classes. If the claim is drawn in terms broad enough to recite generically the characteristics of the "other" class, ingenuity and imagination are required to discover other subsumed terms which are materially different from the term for the disclosed object. For example, if the examiner has a disclosure of a salt-shaker before him, and the claim is drawn in terms of a coating applicator, he of course searches salt-shakers. If he cannot find a salt-shaker which anticipates the claim, he usually thinks of talc dispensers, clothes sprinklers, track sanders and the like, but seldom of pencils. For some strange psychological reason, the presence of the originally disclosed object, the salt-shaker, throws a mental shadow which tends to hide other objects defined by the claim if they are unlike the disclosed object in appearance or operation. This situation is faced almost every time an inquiry is framed for manual searching. A categorical scheme suited to the logic of a mechanized search will eliminate such psychological interference.[13]

### Ambiguities

The resolution of ambiguities either in single terms or whole phrases or sentences may require evaluation of the context, either in the same sentence, in other sentences, or even in the drawings. For example, in an analysis of qualifying language used in a U. S. Patent, the following expression was found which referred to a joint between two separate beads: "The resilience of the material permits the head on one bead being forced through the mouth into the socket of another bead." Interpreted in context, this statement was found to mean that "there must be enough resiliency to yield without breaking or tearing . . . and . . . without allowing the joint to separate . . . ."[14]

Homonyms constitute a specific form of ambiguity, and are numerous in the expository prose of patent usage, since the *jargon* of specialized fields often appropriates terms from other fields. In appropriation, they may be given either a narrower meaning, e.g., *force* (in physics); a broader meaning, e.g., *light* (when used to include infrared radiation); a meaning suggested by their shape or function, e.g., a *coat* (of paint); or a meaning which is purely arbitrary, e.g., a *frog* (of a railroad track).

Two words may be synonyms in one context, but not in another. For example, *deep* and *dark* when used to modify *blue* might be considered virtually synonymous; but when used to modify *chamber*, they definitely have distinct meanings. In describing new techniques, scientific writers undoubtedly will use the existing vocabulary in other special senses and will coin new terms which would not appear in any existing lexicons.

### Implied Conceptual Facets of Terms

An encoding problem closely related to the categorization of broader class terms is raised by the implied conceptual facets of a single term. The designation of the material from which an article is made implies all the known properties of that material. The disclosure of an electrically-actuated device implies a source of current to operate it. Similarly, a term for a disease may imply (1) the cause, (2) the part of parts of the body affected, (3) the symptoms of and/or tests for its presence, (4) the drug or medication used, and its method of and apparatus for administration, (5) other methods of treatment, and possibly (6) the medical uses to which the disease may be put; i.e., a patient may be purposely infected with the disease as inoculation or as a treatment for a different body malfunctioning.

Before patents can be encoded, all the pertinent implied concepts of each explicit term will have to be coded. The choice of the pertinent from the plethora which may be conjured up is the perplexing task of the encoder.

## MECHANIZED ENCODING

Linguistic problems once solved, practicable coding logics formulated, and the searching process mechanized, there remains the staggering task of encoding the more than three million United States patents, the five or more million foreign patents, and other disclosures now in the files. The United States Patent Office alone is issuing new patents at the rate of 50,000 a year. It has accordingly been proposed that plans be made to encode by mechanical means. Some progress has already been made in mechanized pattern recognition. It may safely be assumed that within a few years a practical print reader will be available which will recognize the printed word. Some aspects of research in mechanized translation of language give promise that eventually the meaning of a sentence, in context, will be mechanically extracted from the printed page, "translated" into a simple, unambiguous language, and stored in a machine memory for use in an information retrieval system. Hopefully, as pattern recognition procedures are developed, the direct encoding and storing of standardized charts or drawings will be possible. How non-standard charts and drawings could be encoded for such storage without pre-editing (i.e., remaking them in standard form) is not apparent at this time.

Furthermore, it is well understood that the transformation of complicated manual encoding procedures into a form presentable to a machine will entail elaborate programming. Theories for the inductive inference programming of data processing equipment and their accompanying mathematics have been postulated.[15] Other empirical heuristic logics for the programming of data-processing machines are presently under experiment.[16] Perhaps the key to mechanical encoding lies amid these nascent theories. The research efforts of the Patent Office must persist in this, as well as in the linguistic phase of its general program so that when a mechanical search method has been realized, manpower will not be needlessly wasted in manually encoding and continuously updating the vast files.

## REFERENCES

1. Familiarity with the contents of those Patent Office Research and Development Reports listed in the bibliography is assumed, particularly: Newman, *Problems in Mechanizing the Search.*

2. The annotations in the bibliography specifically refer to linguistic problems involved in mechanization.

3. Andrews and Newman, *Storage and Retrieval, Preliminary Report;* and Newman, *Storage and Retrieval, First Supplementary Report.* For those who may be interested in a broad statement of this system without the details, Newman, "Linguistics and Information Retrieval," summarizes these two reports.

4. Andrews, and Newman, *Storage and Retrieval, Preliminary Report,* p. 3.

5. Bailey, "History of the Classification of Patents," p. 548. In a revised reprint of this article (now out of print) this same reference occurs on p. 57.

6. Manufacturing is used here in its broadest sense, which includes any process of creating, constructing, fabricating, machining, working, shaping, assembling, disassembling and repairing a thing.

7. Cf. *Manual of Classification,* Class 29, Metal Working, pp. 29-1 ff.; Class 144, Woodworking, pp. 144-1 ff.

8. *Ibid.* Class 82, Turning, pp. 82-1 ff.; Class 77, Boring and Drilling, pp. 77-1 ff.

9. Currently under revision is Class 164, presently titled "Cutting and Punching Sheets and Bars." The proposed new class will probably have a title such as "Cutting Machines."

10. "How to Read Subclass Titles and Definitions." *Manual of Classification,* p. I.

11. Markush was the applicant in an early litigated case in which this form of definition of an artificial genus was held to be legally acceptable.

12. Newman, *Storage and Retrieval, Supplementary Report,* p. 15, col. 2, and p. 16, col. 1.

13. There are other situations in which it is now virtually impossible for an examiner to locate pertinent disclosures, but which will not exist when the mechanization of searching has been accomplished. For example, if a search for a mechanical movement is unsuccessful in the proper locus of the present classification scheme, the examiner has the hopeless and impossible task of looking at every patent disclosing movable parts.

14. Newman, *Storage and Retrieval, Supplementary* Report, p. 15, col. 2.

15. Ashby, "Design for an Intelligence-Amplifier" and MacKay, "The Epistemological Problem."

16. Newell and Simon, "The Logic Theory Machine" and Newell, Shaw and Simon, "Empirical Exploration."

## BIBLIOGRAPHY

Andrews, Don D. and Simon M. Newman. *Storage and Retrieval of Contents of Technical Literature, Nonchemical Information, Preliminary Report.* Patent Office Research and Development Report [No. 1.] Washington 25, D. C., Department of Commerce, 1956. This entire report concerns linguistic problems.

Ashby, W. Ross. "Design for an Intelligence-Amplifier." *Automata Studies.* Eds. C. E. Shannon and J. McCarthy. Princeton, Princeton University Press, 1956, pp. 215-234.

# Index


| | Page |
|---|---|
| Foreword | 4 |
| Introduction | 5 |
| Coding Terminology | 5 |
| Unambiguous language | 5 |
| Drawings and Illustrations | 5 |
| Static Structures | 5 |
| Functions | 6 |
| Apparatus | 6 |
| Chemical Compounds | 7 |
| Other Word Classes | 7 |
| Linguistic Aspects of Encoding | 7 |
| Categorizing for Generic Searching | 7 |
| Ambiguities | 8 |
| Implied Conceptual Facets of Terms | 8 |
| Mechanized Encoding | 9 |
| References | 9 |
| Bibliography | 9 |




A73027634