(8) J. Lederberg, et al., "Applications of Artificial Intelligence for Chemical Inference. I. The Number of Possible Organic Compounds. Acyclic Structures Containing C, H, O, and N", *J. Am. Chem. Soc.*, **91**, 2973–76 (1969).
(9) M. Milne, et al., "Search of CA Registry (1.25 Million Compounds) with the Topological Screens System", *J. Chem. Doc.*, **12**, 183–9 (1972).
(10) D. Lefkovitz, "The Large Data Base File Structure Dilemma", *J. Chem. Inf. Comput. Sci.*, **15**, 14–9 (1975).
(11) M. F. Lynch, et al., "Computer Handling of Chemical Information", McDonald, London, and American Elsevier, New York, 1971, p 84.
(12) D. J. Gluck, "A Chemical Structure, Storate and Search System Development at DuPont", *J. Chem. Doc.*, **5**, 43–51 (1965).
(13) Reference 11, p 91.
(14) G. W. Adamson, et al., "Strategic Considerations in the Design of a Screening System for Substructure Searches on Chemical Structure Files", *J. Chem. Doc.*, **13**, 153–7 (1973).
(15) C. N. Mooers, "Zatocoding Applied to the Mechanical Organization of Knowledge", *Am. Doc.*, **2**, 20–32 (1951).
(16) If we were dealing with binary variables instead of descriptors, as differentiated in the opening paragraphs, then the optimal incidence would be $1/2$ instead of $1/e$.
(17) D. E. Knuth, "The Art of Computer Programming. Vol. I. Fundamental Algorithms", Addison-Wesley, Reading, Mass., 1968, p 179.

Experimental Algorithmic Generation of Articulated Index Entries from Natural Language Phrases at Chemical Abstracts Service†

STANLEY M. COHEN*, DAVID L. DAYTON, and RICARDO SALVADOR**

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received November 17, 1975

An algorithm was developed which transforms coded, natural language phrases into multiple index entries consisting of headings and articulated subordinate phrases. Coding and dictation procedures were developed to allow document analysts to input phrases and associated data in a format compatible with the Chemical Abstracts Service (CAS) index production system. In an experiment, 13 CAS document analysts generated, coded, and dictated phrases descriptive of the content of documents input to the CAS processing stream. These phrases were then transformed via the articulation algorithm into entries of the type used in *Chemical Abstracts* (CA) volume indexes. In an evaluation of over 20,000 algorithm-articulated entries, 97.2% were judged intelligible and acceptable. The input of phrases required 62.0% of the keystrokes required for the input of individual entries. The major problem was the error level in the analyst-dictated, clerically keyboarded codes for input phrases. On-line interactive processing techniques may essentially eliminate this problem. The phrase input procedures are being adapted for on-line experimentation.

INTRODUCTION

At present production levels, over 2,000,000 index entries are published yearly in the *Chemical Abstracts* (CA) Chemical Substance and General Subject Indexes. To produce these index entries, document analysts extract information from original documents and/or abstracts and then organize this information into a variety of data components used in the Chemical Abstracts Service (CAS) index production system. Two of the components of the CA volume index entries are of primary interest in this study:

1. *Heading*: the primary access term for an index entry and the basis for alphabetical arrangement of the entries within the index
2. *Text modification*: the indented, subordinate phrase modifying the heading.

Figure 1 illustrates a typical CAS index entry consisting of *heading* and *text modification*.

In the present index production system, document analysts identify and dictate all the data components for each index entry individually. Then clerical staff keyboard the dictated entries for input to the system. Most documents indexed require multiple index entries, which are often permutations of the same words and ideas. Figure 2 illustrates such a group of index entries for one document. Notice that the heading portion of one index entry becomes a part of the text modification of other entries.

Armitage and Lynch[1,2] have described a process called articulation which algorithmically generates multiple index entries from a single, descriptive, title-like phrase. Figure 3 illustrates a title-like phrase which, when subjected to an articulation algorithm, could produce the set of index entries in Figure 2. When a single input phrase can be successfully transformed into a set of index entries, one can readily see the potential for reducing the effort in formulating and dictating multiple permutations of the index entry words as well as the potential for reducing the keyboarding effort required for input. Given the magnitude of the CA volume indexes, this potential reduction of input effort represents substantial savings. Reduced input requirements would take on added importance in any future on-line environment where the document analyst would do his own keyboarding at a terminal.

This paper describes a study of phrase input for indexing use at CAS. Each of the four parts discusses one of the four main aspects of the study.

Part I. An articulation *algorithm* capable of transforming analyst-dictated natural language phrases into multiple index entries of the syntactical style of CA volume index entries.

Part II. Procedures for coding natural language phrases suitable for algorithmic articulation. These *coding procedures* were designed to generate articulated index entries and associated data in data element format compatible with the CAS index production system.

Part III. *Testing* the articulation algorithm and the coding procedures in a production-like environment.

Part IV. An evaluation of the *results* of the study.

† Presented before the Division of Chemical Information, 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 27, 1975.

* Author to whom correspondence should be addressed.
** Servicio de Marketing, Compañía Telefónica Nacional de España, Madrid 20, Spain.

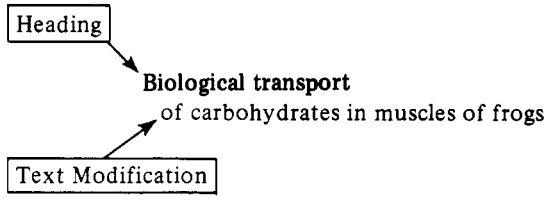

Figure 1. CAS index entry.

Biological transport
of carbohydrates in muscles of frogs

Carbohydrates
biological transport of, in muscles of frogs

Muscles
of frogs, biological transport of carbohydrates in

Frogs
muscles of, biological transport of carbohydrates in

Figure 2. A group of index entries from a single document.

Biological transport of carbohydrates in muscles of frogs

Figure 3. Input descriptive title-like phrase.

X_0 D_0 X_1 D_1 H′ L_1 H L_2 H″ D_2 X_2 | RS_1 RS_2

Figure 4. Model of input phrase.

PART I. THE ARTICULATION ALGORITHM

One of our objectives was to develop the articulation algorithm to operate on title-like input phrases and generate multiple index entries which resembled CA volume index entries to the highest degree possible. As the development progressed, several versions of the algorithm[5] were empirically derived. The one described here is the latest version, and it incorporates experience gained in the testing program. We expect that wider testing will lead to further improvements.

The articulation algorithm is based on a model of an input phrase which describes the content of an indexed document (see Figure 4). The symbols used in the model and elsewhere in this paper are defined as follows:

- H Segment (word or group of words within an input phrase) which is flagged by a document analyst as a heading. The H term will be the heading portion of one of the index entries which result from the articulation of the phrase.
- X Substantive, meaning-bearing segments which are bounded by delimiters.
- D Delimiters (see Table I). These words bound segments within a phrase.
- L Segments adjacent to the H segment with no intervening delimiters. The L segment modifies, or is modified by, H.
- H′ Segment to the left of, and adjacent to, either the H or an L segment. Its rightmost word is *and*.
- H″ Segment to the right of, and adjacent to, either the H or an L segment. Its leftmost word is *and*.
- RS Segments in any position in a phrase which begin or end with one of these repellent terms: *in relation to, in response to, effect on*. These are called repelled segments because the articulation algorithm places them last in a text modification.

Table I. Delimiters

Type A		
about	during	therein
above	following	thereto
according	for	through
after	from	to
against	in	towards
along	into	under
as	near	vs.
below	off	with
between	on	within
beyond	or	without
by	thereby	
due	therefrom	
	Type B	
	of	
	containing	

Table II. Articulation Logic

Condition	Action
D_1 is a Type B delimiter	Structure index entry according to output model 1
Neither D_1 nor D_2 is a Type B delimiter	Structure index entry according to output model 2
D_1 is not a Type B delimiter and D_2 is a Type B delimiter	Structure index entry according to output model 3

In Figure 4 the subscripts for RS segments indicate the order in which the RS segments appear in the phrase. The other subscripts indicate the position of the phrase segments relative to the heading (H). The subscript 1 is assigned to the first X, L, and D to the left of H. Continuing left, the subscript 0 is then assigned. The subscript 2 is assigned to X, L, and D segments to the right of H.

For each heading flagged, the algorithm uses the model to identify segments in the input phrase based on the presence, absence, and type of delimiters and of repellent terms. Most input phrases contain only a subset of the total possible phrase segments identified in the model.

The articulation algorithm is defined by:

1. Articulation logic table (Table II)
2. Output models (Figure 5)
3. Special override conditions:
   a. If both L_1 and L_2 segments are present, and if L_1 ends with a hyphen, change their order to L_1, L_2 in all three output models.
   b. When the H″ segment is present, move the word *and* to the position of the rightmost word of the segment and articulate according to the appropriate output model.
   c. Place an RS defined by the repellent term *in relation to* in the last position in the index entry.
   d. When the designated heading is in the first position in the input phrase and is followed immediately by one of the repellent terms, move the term to the rightmost position of the RS segment. Substitute *in relation to* for the repellent term in the input segment.
   e. Articulate repelled segments in the index entry inversely to the order in which they occur in the input phrase (subject to the special conditions noted in c and d above).
   f. Identify the X_0D_0 segment only when D_1 is a Type B delimiter. In all other cases define everything to the left of D_1 (with the exception of repelled segments) as X_1.

Output Model 1

$L_2, L_1, X_1\ D_1\ H'\ H'', D_2\ X_2, X_0\ D_0, RS_2, RS_1$

Output Model 2

$L_2, L_1, X_1\ D_1\ H'\ H'', D_2\ X_2, RS_2, RS_1$

Output Model 3

$L_2, L_1, D_2\ X_2, X_1\ D_1, H'\ H'', RS_2, RS_1$

Figure 5. Output models.

In all cases it is assumed that the heading portion of the phrase has been designated by the analyst. The program then generates an index entry consisting of the designated heading and a text modification articulated from the remainder of the phrase. The output models in Figure 5 refer to the text modification portion of the index entries.

The three examples in Figure 6 illustrate the operation of the articulation algorithm. Each uses one of the three output models in Figure 5 in the generation of its respective text modification. For simplicity, a single heading was selected in each example. However, more than one heading is normally flagged.

PART II. PHRASE CODING PROCEDURES

Those portions of input phrases selected by the document analysts as headings were so designated by means of dictated codes inserted at appropriate points during the dictation of the input phrase. The phrase codes we developed serve several other functions, all of which are necessary for the articulation program to generate index entries in a data element format compatible with the CAS index production system.

The functions of the phrase codes are:

1. To delimit the beginning and end of analyst-selected headings within the phrase
2. To provide a unique identification of all phrases input for a single document
3. To distinguish chemical substance headings from general subject headings
4. To provide a unique numerical identification for each heading selected for a document
5. To provide a link between generic headings designated within a phrase and other specific headings dictated subsequent to the dictation of the phrase. In such cases the entries generated for the generic term in the input phrase should be identical with the entries generated for the specific headings.
6. To allow the document analysts to suppress the indexing of a generic term flagged within the phrase (when required by indexing practice) while allowing index entries to be generated for specific headings linked to the generic heading (as described in 5 above). The text modifications generated are based upon the position of the suppressed generic heading within the phrase.
7. To allow document analysts to link other modifying data elements used in building the CAS volume index to selected headings.

The coding procedure developed for this study utilizes multicomponent codes to fulfill each of the functions listed above. Coding components serving functions 1, 3, and 4 are used for every selected heading. Coding components serving functions 5–7 are only used with those headings to which they are applicable. Coding components serving function 2 are used each time a new phrase is input.

Figure 7 illustrates an input phrase with the coding which serves function 2 and the three "always-used" functions. Four terms in the input phrase are flagged as headings: aluminum chloride, chlorination, mesitylene, and chloroacetophenones. The dotted lines and the codes delineate the boundaries of the selected headings. The dotted lines also indicate the points in the dictation of the input phrase at which the codes are inserted. PAR and CTH are the coding components serving function 3 above. They are used at CAS to distinguish substances, which are submitted to the nomenclature control system[4] of the CAS Chemical Registry System, from concepts, which are submitted to the general subject vocabulary control system.[3] The digits 1, 2, 3, and 4 provide a unique identification for each heading flagged (function 4) for a document. A combination of PAR and a number or CTH and a number

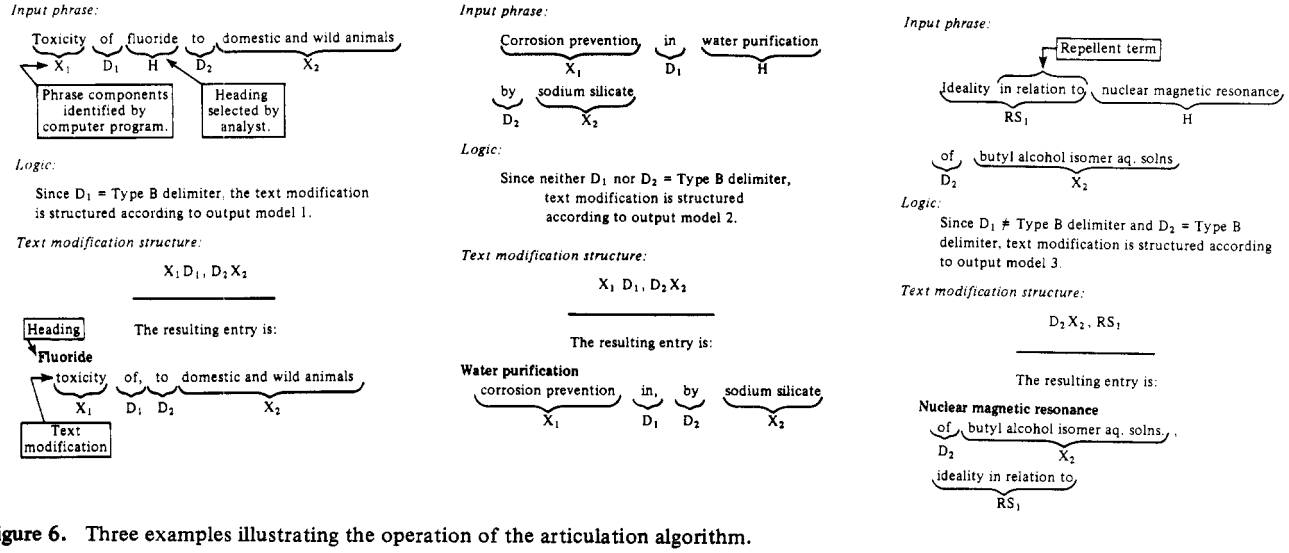

Figure 6. Three examples illustrating the operation of the articulation algorithm.

Coding:	PHS 1	PAR 1		#	CTH 2	#	PAR 3		#	CTH 4		#
Input Phrase:			Aluminum chloride	catalyst for	chlorination	of	mesitylene	by		chloroacetophenones		

Figure 7. Coding of an input phrase.

Dictated input

Coding:	PHS 1	PAR 1	#	CTH 2	#	PAR 3	#	CTH 4-Ø-S	#
Input Phrase:		Aluminum chloride	catalyst for	chlorination	of	mesitylene	by	chloroacetophenones	

Subsequent Dictation: PAR 1 QLF uses and miscellaneous

PAR 4-1 2,2,-trichloro-1-(2,5-dichlorophenyl)ethanone

PAR 4-2 2-chloro-1,2,2-triphenylethanone

Articulated output in data element format

PAR Aluminum chloride QLF uses and miscellaneous
TMD catalyst, for chlorination of mesitylene by chloroacetophenones

CTH Chlorination
TMD of mesitylene by chloroacetophenones, aluminum chloride catalyst for

PAR Benzene SUB 1,3,5-trimethyl-
TMD chlorination of, by chloroacetophenones, aluminum chloride catalyst for

PAR Ethanone SUB 2,2,2-trichloro-1-(2,5-dichlorophenyl)-
TMD aluminum chloride catalyst for chlorination of mesitylene by

PAR Ethanone SUB 2-chloro-1,2,2-triphenyl-
TMD aluminum chloride catalyst for chlorination of mesitylene by

Figure 8. Dictation and articulated output.

delimits the left boundary of a flagged heading (e.g., PAR 1, CTH 2); the pound sign (#) delimits the right boundary. Together they serve the first function. The multicomponent code PHS 1 (function 2) indicates that the phrase shown in Figure 7 is the first phrase entered for the document being indexed. The second phrase would be coded PHS 2, etc.

Functions 5, 6, and 7 of the coding procedure are illustrated in Figure 8. The input phrase is repeated from Figure 7, but additional data have been linked to the fourth flagged heading (chloroacetophenones). The document analyst dictates the additional linked data, the subsequent dictation shown in Figure 8, after dictating the input phrase.

The coding for the heading *chloroacetophenones* includes the basic components used for all headings and two additional components. The zero in the code CTH 4-Ø-S is used to indicate that this is a class heading and that specific members of the hierarchical family headed by the generic term in the phrase will be dictated following the phrase. Figure 8 shows two specific chloroacetophenones dictated subsequent to the phrase. They are linked to the general term in the input phrase through the unique numeric component (4) in their codes.

The code CTH 4-Ø-S illustrates a case in which indexing practice requires that the indexing of a generic heading be suppressed in favor of linked specific headings. Coding the S component of the code suppresses the generation of an index entry for the general heading chloroacetophenones. However, the text modifications of the index entries generated for the two specific compounds linked to the suppressed general heading through the coding are based upon the position of the general heading flagged in the input phrase. For some general headings, indexing practice requires an index entry for the general heading as well as the linked specific headings. This is accomplished by omitting the S component when coding the general heading in the input phrase.

In the CA volume indexes, certain headings are subdivided by qualifying terms such as reactions, properties, uses and miscellaneous, etc. In Figure 8, the subsequent dictation includes "PAR 1 QLF uses and miscellaneous". This links the qualifying term, "uses and miscellaneous", to the heading which is coded PAR 1 in the input phrase. Thus the term "uses and miscellaneous" is associated with the heading "aluminum chloride" in the articulated index entry.

If the phrase input procedures are implemented on a production basis, all flagged headings will undergo the appropriate heading control procedures to extract their preferred forms for printing in the index. Such computer-based control procedures presently operate on standard input as part of the CAS volume index production process. These will operate on phrase input in essentially the same way. The headings selected in this experiment were not subjected to the control procedures. The last three index entries shown in Figure 8 do, however, illustrate the transformation of the three nonpreferred input compound names to the preferred form for publication.

PART III. THE TESTING PROGRAM

Our objectives in setting up the testing program were to determine:

1. The feasibility of the phrase input method at CAS
2. The workability of the articulation algorithm
3. The quality of index entries produced
4. The workability of the phrase input procedures from the viewpoint of the document analyst
5. Any keyboarding savings

The testing program was designed to provide a production-like environment in which document analysts

applied the phrase input procedures to actual CAS material, and the resulting phrases were articulated via the algorithm. A group of 13 experienced document analysts were trained in the coding and dictation procedures for the input of phrases. Using the phrase input procedures, the document analysts indexed documents from subject areas in which they had extensive indexing experience. They later indexed the same documents by the standard CAS procedure (formulating and dictating each index entry individually). A total of 4903 documents, covering 32 of the 80 CAS subject sections, were indexed.

The coded phrases and the standard index entries for the set of documents were keyboarded to separate magnetic tapes; hardcopy records of the input by the two approaches were accumulated. The tapes were used as input to a computer program which made a character-by-character keystroke comparison of phrase input and standard index entry input. The coded input phrases were subjected to the articulation algorithm, and statistical information on certain phrase input parameters was accumulated.

The algorithm-articulated index entries were circulated back to the document analyst who had formulated the input phrases from which they were generated. The analysts evaluated the acceptability of the algorithm-articulated index entries using CAS indexing practice as the standard.

PART IV. RESULTS AND DISCUSSION

Application of the Phrase Coding. All the document analysts who participated in the testing program began indexing by the phrase input method after studying a procedures manual and participating in two three-hour training sessions. Following the completion of the training sessions, the only on-going assistance to analysts in the phrase input techniques was answering occasional questions. The frequency of these questions diminished rapidly after several days of experience with the new technique. Since the phrase input procedures were designed to reflect existing CAS indexing practice, some aspects of the dictation procedure were quite similar to the dictation procedures used in the standard CAS index entry input. Consequently the analysts had to learn only the features which were specific to phrase input.

The document analysts successfully applied the phrase input technique in all 32 subject areas covered by the testing program and had little difficulty coping with the spectrum of indexing situations which occurred.

The foremost problem encountered in the testing program was the error level in the analyst-dictated, clerically keyboarded phrase codes. There was at least one coding error in the phrases input for 24% of the documents indexed.

The reasons for this high coding error level were: (1) inserting the codes during the dictation of a phrase tended to interrupt the analysts' continuity of thought; (2) analysts had difficulty keeping track of numerical components of the codes; (3) codes were omitted by both analysts and keyboarding staff; (4) codes were dictated incorrectly; and (5) codes were keyboarded incorrectly.

We attempted to simplify the coding in order to reduce the coding error level. None of the coding components could be eliminated because each served a vital function in the index production process. We found no way of significantly simplifying the coding procedure while maintaining the capability of building index entries in a format compatible with the CAS index production system. While greater experience with the input of the coded phrases may have somewhat reduced the coding error level, we believe that the on-going error rate would be too high to further consider batch-dictated phrase input.

The phrase input technique is presently being adapted for experimentation with on-line input. On-line interactive processing seems well suited to solving most, if not all, of the coding problems encountered. The advantages of on-line input of phrases relative to the coding error problem are:

1. The phrases will be initially keyboarded without coding. Supplying codes after the phrase is visible on a screen will eliminate the awkwardness of embedding codes within a phrase during dictation.
2. The analysts will be freed from remembering the numerical components of the codes because they will be calculated and supplied by the input program.
3. Since there is a programmable logic to the order in which the codes and components of codes are supplied, interactive processing should greatly reduce errors of omission.
4. Since a computer program will internally generate the actual coding in response to commands of an analyst at a terminal, the coding errors in the dictation-keyboarding cycle will be eliminated.

Table III. Phrase Input Statistics

Documents indexed	4,903
Phrases input	6,609
Headings flagged within phrases	12,959
Phrases per document (average)	1.35
Headings flagged per phrase (average)	1.96
General class headings flagged per phrase (average)	0.6
Specific headings dictated per general class heading flagged (average)	2.62
Algorithm-articulated text modifications	20,546
Text modifications generated per phrase (average)	3.11
Keystrokes for phrase input	954,201
Keystrokes for standard input	1,539,980
Phrase input keystrokes as % of standard input	62.0%

Input Statistics. Statistical data for the input of the coded phrases are summarized in Table III. The document analysts formulated, coded, and dictated a total of 6609 phrases for 4903 documents indexed. A total of 12,959 headings were flagged within the phrases. The average number of headings flagged per phrase was 1.96; i.e., for each of the 6609 index phrases formulated in the testing program, approximately two different index entries were articulated. The figure 1.96 bears some relationship to the potential reduction of keyboarding as a result of phrase input.

It is important to note that the figure 1.96 refers to *distinct* index entries. The *total* number of index entries generated per flagged heading averaged 3.11. This includes different index entries from the same document which have identical text modifications. Distinct entries are those in which the text modifications associated with the headings are different. Figure 8 illustrates this point. A total of five index entries result from the articulation of the phrase; four are distinct; two of the entries have identical text modifications. The present CAS volume index production system allows for the duplication of identical text modifications for different index entries from a single document. Thus, while the heading linkage feature of the phrase input coding procedure retains this capability, by itself it does not create any keyboarding advantage over the standard input system.

The standard input of index entries and the coded phrases for the same set of documents were keyed to separate magnetic tapes. A computer program determined the number of keystrokes required by each input method by performing a character-by-character count on each tape. Comparison indicates that phrase input required 62.0% of the keystrokes required by the standard index entry input. If the keyboarding operation involved only the transcription of phrases, each with an average of about two headings flagged, one would expect phrase input to require one-half as many keystrokes as the standard method. Phrase input keystrokes were actually 62.0% of standard input keystrokes because of the input of subsequent dictation associated with the index phrase. This information

Table IV. Text Modification Evaluation Categories

Category	Definition
1 (Unacceptable)	Ideas are confused or syntax completely wrong
2 (Acceptable)	Text modification is intelligible but not in a CAS-preferred format. These are easily transformed into a more preferred form by rearranging the segments within the text modifications.
3 (Excellent)	Index entry is completely acceptable according to CAS indexing practice.

Table V. Evaluation Summary

Category	No. of text modifications	Percent
Unacceptable	569	2.8
Acceptable	3,804	18.5
Excellent	16,173	78.7

includes specific headings linked to class headings flagged within the index phrases, line formulas, heading-qualifying terms, and other associated data required by the CAS volume index production system. The simple keystroke comparison is important, but it is only one factor to be considered in any economic evaluation of the relative keyboarding requirements of the two input methods.

Evaluation of Text Modifications. A total of 20,546 algorithm-articulated index entries were generated in the testing program. A sampling is given in the Appendix. The document analyst who formulated the phrase for input to the articulation algorithm evaluated the resulting text modifications in accordance with CAS indexing practice. Each of the text modifications was placed in one of the categories defined in Table IV, and the results of the evaluation were tabulated (see Table V).

The text modifications placed in categories 2 and 3 were all completely intelligible, but the order of the substantive words in the category-2 text modifications was not the most preferred order according to CAS indexing practice.

In the CA volume indexes, text modifications are sorted alphabetically within headings based upon the first substantive word. Document analysts have traditionally placed considerable emphasis on placing what they consider the most important idea first in the text modification so that the user can perform a coordinate search between the most important idea and the heading. The subjective nature of choosing the most important idea in a text modification is reflected in the evaluations.

Categories 2 and 3 totaled 97.2% of all the algorithm-generated text modifications. The low fraction of text modifications placed in the unacceptable category (2.8%) was quite encouraging. Most of the adjustments made to the articulation algorithm as a result of the testing were directed toward decreasing this number. It is expected that future tests with the upgraded algorithm will show a significant increase in the percentage of algorithm-articulated text modifications judged excellent or acceptable.

CONCLUSIONS

Document analysts can learn and apply the phrase input procedures after receiving about six hours of formal instruction. Difficulties experienced by the analysts in switching from the familiar individual index entry procedure to the phrase input techniques were not of major consequence.

Documents indexed by the phrase input method encompassed a wide range of the subject areas covered by the CA volume indexes. No major problems were encountered in applying the phrase input procedure over the various subject areas.

Table VI

Index entries	Rating
Abrasion-resistant materials nylon-contg. aluminum stearate as	3
Accommodation coefficients of gases on platinum, thermal conductivity in relation to	3
Acidosis treatment of, lactulose-contg. pharmaceuticals for	2
Alkaline earth fluoride electrolytes, in thermodynamic studies of inorg. compds.	3
Anemia treatment of, sustained-release hematinic tablets for	2
Antigens viral hepatitis, vaccines of	3
Catalysts cobalt ferrite, activity of, topochemical effects in	3
Cell division mitosis, in kidneys and intestines of goldfish after adaptation in temp.	3
Cell nucleus liver, protein metabolism by	2
Dental materials copper-gold alloy, hardening of	3
Electric charge prevention of, on propane polymers by treatment with steam	3
Electron beam irradiated ethylene-vinyl acetate polymer, for paraffin wax compds.	1
Heat transfer relativistic thermodynamics, in between comoving bodies	3
Hydrogen peroxide rotation of, electronic properties and potential barriers to	1
Kaolinite in mudstone of Ciechanowice area, Poland	2
Kinetics of oxidation of sodium formate by triode in aq. soln.	3
Lithium chloride reaction of, with vanadium oxide, reaction mechanism of	3
Liver damage, enzymes of blood serum in response to	3
Magnetic substances superconductors contg., heat capacity of	3
Magnetic domains decompn.-structure replication of, in iron-silicon alloys	3
Malt detn. of color development in	2
Minerals aluminofluoride, of Werainian S.S.R.	3
Mole control, Sevin as rodenticide in	3
Muscle relaxants extracts of anemone hepatica as	2
Muscles proteins of, of ascarid, collagens in relation to	3
Silica coating materials, colloidal, for plastic films for agricultural purposes	3
Skin dermatitis, lubricants in yarn manuf. effect on	3
Stomach tachyphylaxis of, acid secretion by, to secretagogues, cyclic AMP effect	1

A total of 20,546 index entries were generated from coded phrases via the articulation algorithm. It was quite encouraging that 97.2% of the entries were completely understandable. Adjustments made to the articulation algorithm as a result of the testing should effect an even greater percentage of intelligible entries.

Phrase input affords a significant reduction in the number of keystrokes required for the input of index entries, but the phrase codes were not well suited to dictating input for batch

processing. It does appear, however, that the phrase coding and input procedure lends itself quite advantageously to on-line interactive processing. The procedure is presently being adapted for on-line testing.

ACKNOWLEDGMENT

CAS is pleased to acknowledge the partial support of this work from the National Science Foundation (Contract C656).

APPENDIX. ALGORITHM-ARTICULATED INDEX ENTRY SAMPLE

Listed in Table VI is a sampling of the index entries generated from coded phrases via the articulation algorithm during the course of the testing. The index entry ratings assigned by the document analysts are also shown. The purpose of this listing is only to illustrate the quality of text modifications generated by the algorithm. The headings in this listing are the *as-dictated* headings which have *not* been subjected to the CAS heading terminology control systems. Therefore, in some cases, the headings shown are not the headings which would appear in the CAS volume indexes. See Table IV for meaning of ratings.

LITERATURE CITED

(1) J. E. Armitage and M. F. Lynch, "Articulation in the Generation of Subject Indexes by Computer", *J. Chem. Doc.*, **7**, 170–8 (1967).
(2) J. E. Armitage and M. F. Lynch, "Some Structural Characteristics of Articulated Subject Indexes", *Inf. Storage Retr.*, **4**, 101–11 (1968).
(3) R. D. Nelson, W. E. Hensel, D. N. Baron, and A. J. Beach, "Computer Editing of General Subject Heading Data for *Chemical Abstracts* Volume Indexes", *J. Chem. Inf. Comput. Sci.*, **15**, 85–94 (1975).
(4) R. J. Rowlett and F. A. Tate, "A Computer-Based System for Handling Chemical Nomenclature and Structural Representations", *J. Chem. Doc.*, **12**, 125–28 (1972).
(5) R. Salvador, "Automatic Abstracting and Indexing", Masters Thesis, The Ohio State University, Columbus, Ohio, 1969.

ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques

A. J. STUPER and P. C. JURS*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 11, 1975

An interactive computer system has been developed for the convenient application of pattern recognition techniques to chemical problems. The system includes subsystems to perform the following functions: graphical input of molecular structures for the generation of files of structures stored as connection tables; file handling and revision; development of descriptors from molecular structure connection tables for pattern recognition analysis; prior feature selection and preprocessing; discriminant development; feedback feature selection. The system is modular, allowing its execution on a relatively small computer and allowing the user to add or delete routines easily.

The introduction of the digital computer into the chemical laboratory offers the chemist a new and exciting tool. A multitude of tasks including literally hundreds of numerical integrations can be handled by the device with a speed and accuracy unapproachable a mere decade ago. Not all problems faced by the chemist, however, lend themselves to such exacting solution: frequently, equations describing processes of interest are difficult or impossible to obtain, and a host of problems have not yielded to a satisfactory or usable theoretical explanation. In the absence of theoretically based solutions, empirically derived methods will often suffice to yield useful and practical solutions to complex problems.

Standard approaches to the extraction of information from complex data forms have included linear optimization, information theory, and a plethora of statistical analysis techniques. Since the early 1950's pattern recognition methods have also been applied to a variety of data interpretation problems and have paralleled the computer's growth in speed and sophistication with a corresponding expansion in scope and capacity. Pattern recognition techniques have found application in such varied fields as computer and information science, engineering, statistics, biology, physics, medicine, and physiology. Each of these disciplines has adapted the basic methods of pattern recognition to its own specific requirements.

Analyses using pattern recognition have also encompassed a number of chemical problems in areas including mass spectrometry,[1-5] infrared spectrometry,[6-8] stationary electrode polarography,[9] material production problems,[10] NMR spectroscopy,[11,12] and gas chromatography.[13] Recently, several papers have appeared which indicate the utility of these techniques in the search for correlations between molecular structure and biological activity.[14-22] Two pattern recognition program packages have been described and offered to potential users.[23,24]

Pattern recognition methods are uniquely suited to a variety of studies because of several novel attributes. No mathematical model is used, but rather relationships are sought which provide definitions of similarity between diverse groups of data. Pattern recognition techniques are able to deal with high-dimensional data (data for which more than three measurements are used to represent each object). Such high-dimensional data cannot be directly visualized or displayed. In addition, pattern recognition techniques can deal with multisource data or data in which the relationships are discontinuous. In multisource data each measurement can be the result of an independent generating algorithm or experiment, and each can have a different scale, origin, distribution, etc., from all the other measurements. Therefore, there need be no direct functional relationship between the measurements in multisource data as there must be, for example, in an absorbance vs. concentration plot. For many chemical problems, and especially for those providing multisource data, it is difficult to know in advance whether an appropriate set of measurements has been generated to effect a satisfactory solution. The generation of sufficiently informative multisource measurements can become in itself a major part of the