assistance necessary. The relative ease with which one can learn the Wiswesser system is in part due to its use of known chemical symbols and the use of mnemonics. It is also due in part to the efforts of Dr. Smith in improving the original notation. Assisting Dr. Smith in evaluating and improving the Wiswesser Notation are a group of users known as the Chemical Notation Association. These 14 chemical workers all have encoded at least 5000 structures and are constantly using and improving the notation. Newsletters are circulated and formal meetings are held to ensure that the notation will keep pace with chemical advances and that problems which arise will be promptly discussed and solved. At least 16 structure files coded according to the Wiswesser Line Notation are being maintained. These files contain over one-half million compounds.

ACKNOWLEDGMENT

The author is indebted to Mr. D. A. Kerr who prepared the SRI Molecular Formula and Accession Number Indexes and to Mr. Wm. S. Duvall of the SRI computer section who wrote the ALGOL program.

LITERATURE CITED

(1) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., and Metcalf, E. A., J. CHEM. DOC. 4, 56 (1964).
(2) Granito, C. E., Gelberg, A., Schultz, J. E., Gibson, G. W., and Metcalf, E. A., *ibid.*, 5, 52 (1965).
(3) Granito, C. E., Schultz, J. E., Gibson, G. W., Gelberg, A., Williams, R. J., and Metcalf, E. A., *ibid.*, 5, p. 229.
(4) Gelberg, A., *ibid.*, 6, 60 (1966).
(5) "Survey of Chemical Notation Systems," National Academy of Sciences–National Research Council Publication 1150, Washington, D. C. 1964.
(6) Wiswesser, W. J., "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954.
(7) Personal communication with Mr. Ernest Hyde of the Central Research Laboratory of Canadian Industries Ltd., McMasterville, Quebec, has shown that all substructures and spatial relationships can now be searched by computer-generated fragments of the Wiswesser Notation.
(8) Wiswesser, W. J., "A Line-Formula Chemical Notation," revised by E. G. Smith, McGraw-Hill, to be published, 1967.
(9) The very fast ALGOL sort routine was not available when this program was written. The new routine will perform the sorting operations in approximately half the time.
(10) Personal communication with Dr. C. M. Bowman of the Dow Chemical Co., Midland, Mich.

Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures*

J. M. MULLEN
Shell Development Company, Emeryville, California 94608

Received November 11, 1966

Novel features have been added to a paper tape typewriter having a removable typing element: A symbol set has been devised which requires only nine characters for typing common chemical structures. The typewriter has an uncoded "INDEX" key which advances the paper without carriage return. A companion key, "BACK INDEX," was provided which directly retracts the paper. Both have been coded. A tape record containing information sufficient for a computer to calculate an atom-bond connection table for a chemical structure is obtained by typing the structure in any order solely from the keyboard or by use of the reader with prepunched tapes containing frequently occurring substructures. Cost was about one-fourth that of earlier paper tape chemical typewriters.

Only A. P. Feldman (*1*) and his colleagues at Walter Reed Army Institute of Research are known to have made successful prior efforts to use a paper tape typewriter as a computer input device for chemical structures. Their current machine, the Army Chemical Typewriter (ACT), uses a three-shift keyboard; one shift contains a set of 39 chemical symbols, modified slightly from those of Miller and Fletcher (*2*) at American Cyanamid. The tape of the ACT also records the X and Y coordinates of characters as they are recorded on paper. The ACT is technically attractive and seems quite appropriate to the Army's large problem, but its price of $18,000 is questionable for an industrial laboratory.

* Based on a paper presented before the Division of Chemical Literature, 150th National Meeting of the American Chemical Society, Atlantic City, N. J., Sept. 13, 1965.

FEATURES OF THIS NEW MACHINE

The typewriter described in this paper (Figure 1) is distinguished by having an abbreviated character set, modification to provide movement of the paper in all four directions under keyboard or tape control, capability for use of edge-punched cards to facilitate typing of frequently occurring substructures, and a replaceable typing element. Its projected cost on a commercial basis is under $5000. Additional cost information on the pilot model is given in the Appendix.

Context. The typewriter is being used to prepare unit records for an existing manual chemical structure system and to provide concurrently an input tape for computer processing. The tape is used to establish a file of chemical structures which can be searched for identity or for the presence of one or more desired structural fragments. The unit record for the system described here includes the items shown in Table I.

Table I. Elements of a Chemical Compound Information System

Registry Number
Chemical Structure
Molecular Formula
Nomenclature
Preparation: Date, Chemist, Reactants, etc.
Properties: Melting Point, Boiling Point, Refractive Index, etc.
Other Textual Information

Symbols for both text and structure were obviously required, and, of course, both upper and lower case alphabets are desirable.

Special Character Set. A special character set was designed, using half-line spacing, which requires only nine basic symbols to type the structures of most chemical compounds. Figure 2 shows the nine basic symbols in the upper case positions of the numeral keys; the most frequently used symbols were placed on keys under the strong index and middle fingers. The "fat dot" is one of H. P. Luhn's (*3*) last contributions to documentation before his untimely death in 1964. It is used to represent carbon atoms (and their hydrogens) in a ring system. Luhn observed that this single nondirectional mark replaces 24 corner symbols as used by Miller and Fletcher (*2*) and by the Army Chemical Typewriter. Two elongated diagonals (keys 0 and 39), suggested by Maxwell Gordon (*4*) have been provided for typing stereo configurations. The alphabetic and numeric keys are in their conventional positions. Figure 3 shows that the numerals are lower case size, a concession to their frequent use as subscripts. When used as locants or multipliers they are not misleading; half a line high they provide conventional superscripts.

Compatibility. Under the auspices of the Subcommittee on Compatibility (*6*) of the NAS-NRC Committee on Modern Methods for Handling Chemical Information, discussions were held with representatives of the Army which led to geometric compatibility between the nine basic symbols used in the work under discussion and the larger number on the ACT. (The term "geometric compatibility"

Figure 1. Shell's Chemical Typewriter.

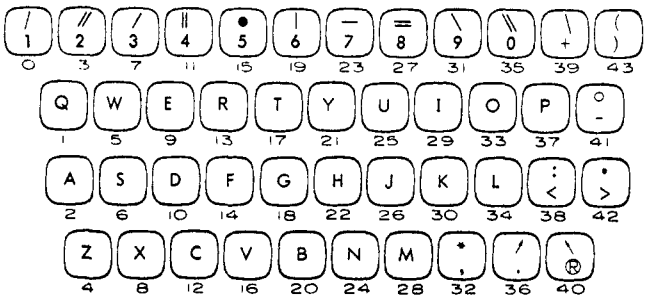


Figure 2. Chemical keyboard.

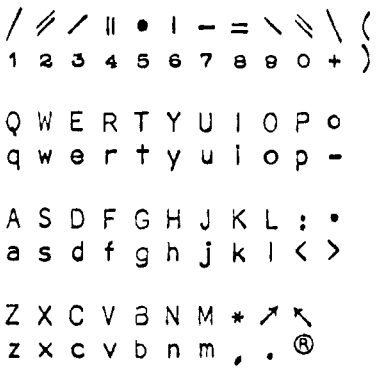


Figure 3. Typeout of chemical keyboard (5).

is meant to imply that a structure could be copied from either system into the other using the same number of lines and spaces with the line segments of the two structures being interchangeable one for one.) Eli Lilly and Co. (7) working with IBM have found the entire character set suitable for a high-speed chain printer, excepting only that the elongated diagonals must be shortened slightly so that they do not quite reach from corner to corner of the type space. The Appendix of this paper contains further comments on compatibility and on the selection and placement of the symbols.

Figure 4 shows a multicyclic compound typed by the Army Chemical Typewriter and by ours. Figure 5 shows a variety of compounds including stereo configurations and metallic complexes. Early plans do not include handling stereo concepts in the computer, but at least partial provision has been made for their graphic representation with a view to the future.

Back Index. The symbols discussed above are equally usable on any typewriter equipped to handle paper tape. The methods described by Waldo and DeBacker (8) or as modified by Horowitz and Crane (9) can be adapted to a tape typewriter with or without the special symbol set—*i.e.*, the structures can be typed on a line-by-line basis after careful layout by a worker familiar with chemical structures. Returning to Figure 1, Shell's Chemical Typewriter is distinguished from others by having coded keys for INDEX and BACK INDEX which respectively move the paper one half line up or one half line down without carriage return. (INDEX and BACK INDEX are also known as LINE FEED and REVERSE LINE FEED.) The INDEX key is standard on the typewriter as purchased, but is not coded. The BACK INDEX and the two codes have been added. The options for manual

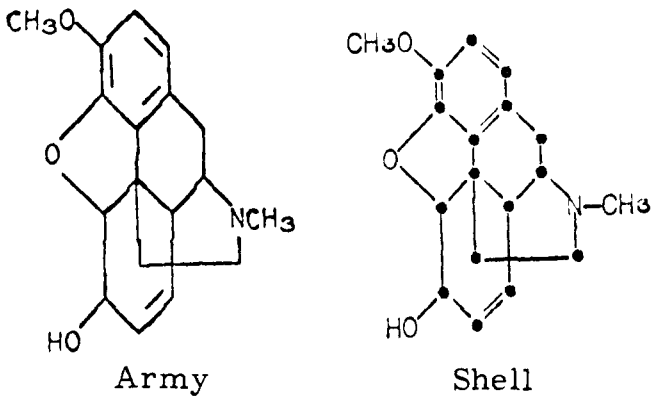


Figure 4. Structural diagrams of codeine.

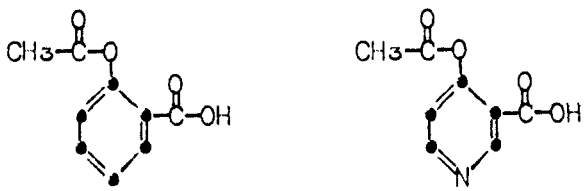
adjustment of the line spacing caused by a carriage return are still available—that is, full space, space and a half, or double space. The so called "soft platen" feature is also undisturbed. The chemical structure input program uses full-line spacing for carriage return and half-line spacing on both indexes.

The typist is instructed to signal the beginning of a structural formula, then copy it in *any* convenient order, and signal the end of the structure. She may encode the structure entirely from the keyboard or she may use the reader to copy parts of structures previously encoded. She must *not* move the platen or carriage by hand while typing the structure. If these instructions are followed and if no machine errors occur it follows that the coordinates of each symbol can be calculated with reference to the original starting position. The typewriter tape may be used in its original form to retype the structure; no intermediate computer processing is necessary (as is the case for the ACT). The rest is done by the computer. The transformation of typewriter codes to a connection table is easily visualized by some but only with difficulty by other equally intelligent chemists. The latter may appreciate the explanation given in the Appendix as applied to a simplified example.

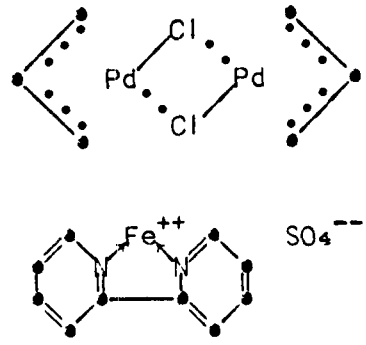
Removable Typing Element (5). Like the symbol set, the INDEX-BACK INDEX feature will work on any typewriter. The reasoning which led to choosing the particular machine (10) is as follows: Typing a structure had been reported (1, 11) and since confirmed by experience to require about 2 minutes. A week of steady work produces about 1000 structures. Thus, structure typing is about 20% of full time for each 10,000 structures typed annually. Assuming that the backlog of structures has already been entered, 10,000 structures per year is projected to be the right order of magnitude for an industrial laboratory. Choosing a typewriter with a fully interchangeable typing element made it unnecessary to commit a typewriter exclusively to chemical information work. Once particular attention was paid to the typing element machine, other benefits were observed. While it types only marginally faster from the keyboard, its rate from tape is about 175 words per minute as compared to about 100 words per minute for conventional machines with hammerlike type bars. So if an appreciable fraction of work could be done using the reader, operating costs would be lower.

Cis and *Trans* 2-Butene

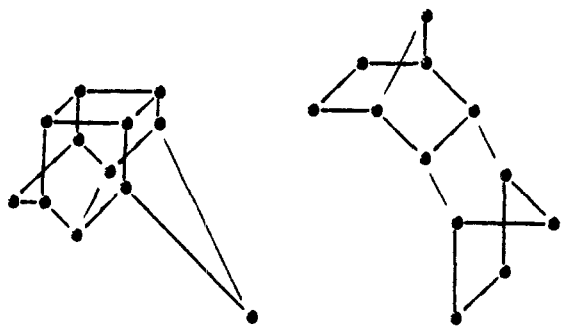
Chair and Boat Forms of Cyclohexane



Aspirin and an Analog



Metallic Complexes



Stereo Configurations

Figure 5. Sample structures.

Edge-Punched Card Attachments. The use of the reader to reduce keyboard typing with its inherent inaccuracy has already been mentioned. Early attempts to make use of this inviting feature were not encouraging in that handling the short tapes required too much of ordinary human dexterity. Furthermore, identification of the prepunched tapes increased the frustration. Drawing a complex substructure on a tape only 1 inch wide presents problems, as do storing and retrieving the tapes for use when needed. Happily, the manufacturer offers an inexpensive option on both the punch and reader for handling cards of any reasonable width interchangeably with tape. The usual card, a continuous fanfold of 3 inch × 7 inch segments with sprocket holes prepunched near the lower edge, is shown in Figure 6. (Note that there is plenty of room to show a much larger substructure than the one illustrated.) A continuous belt moves a blank card into punching position. After a substructure is punched the card is ejected by depression of the FEED button. Another belt moves a prepunched card into and through the reader, starting and stopping the reader by signals from the card. Manual switch buttons enable the typist to stop the reader at any point. In this way she can, for example, keyboard a hetero atom into a carbon skeleton or change a bond from single to double (see aspirin and its analog in Figure 5). Provision is made for skipping over the corresponding code in the reader.

Other Potential Uses. The INDEX-BACK INDEX feature is potentially applicable for computer input from any graphic presentation for which symbols can be established and self-consistent rules written. Among these are:

Electronic circuit diagrams (*12*)
Mathematical equations (*13*)
Music (*14, 15*)
Typesetting

A less obvious "graphic presentation" is a business form. Use of the INDEX-BACK INDEX feature, again with "hands off" the platen and carriage, permits a typist to fill in the form in any convenient order. The computer can be programmed to identify the nature of information by its calculated location on the page rather than by its occurrence in a fixed sequence. The concept would be applicable where it is more difficult to redesign a form than to program the computer. Forms used by industry but prescribed by governmental agencies fall in this class.

CONCLUSION

The chemical typewriter discussed in this paper successfully meets the limited objective of a computer input device for chemical structures whose topology is completely described. It provides the same capability as the Army Chemical Typewrtier at a price only nominally greater than that of a standard tape typewriter and with an acceptable reduction in redundancy when compared to the ACT, where codes for coordinates are actually punched. Still to be verified is the promise of successful computer input of the more complex stereo isomeric compounds; capability for *cis* and *trans* isomerism whether in chains or rings is almost surely attainable. Adequate

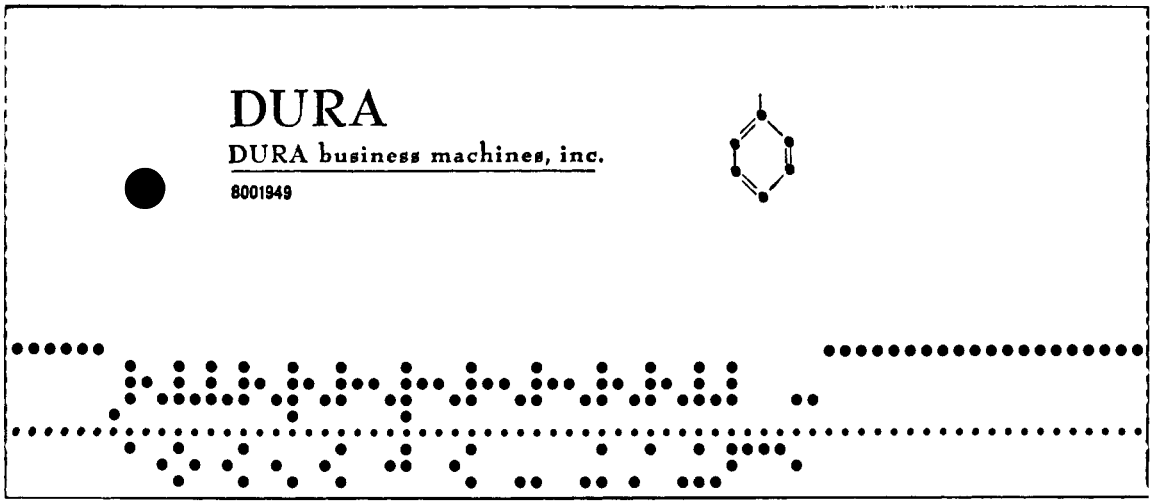


Figure 6. Edge-punched card for pendant benzene ring

representation of metallic complexes is not so clearly established.

ACKNOWLEDGMENT

A special acknowledgment is made to F. G. Stockton of Shell Development Co. who first perceived the benefit of the BACK INDEX function. Technical contributions were also made by Lloyd B. Campbell of Camwil, Inc., Los Angeles; Paul Enz, Dura Business Machines, San Francisco; and M. C. Lowman, R. W. Stevenson, and P. R. Wallace, all of Shell Development Co.

APPENDIX

Cost. This typewriter is competitive in price with earlier paper tape machines; its basic price is $2600 for an 11-inch carriage without accessories, $2750 for the 15-inch carriage shown. The alteration according to the most recent design costs about $1000, fully overheaded. Accessories raise the total price of the typewriter as shown to $4500 plus applicable taxes; it is probable that a commercial version to be offered by the manufacturer will be priced near this figure.

Additional Comments on Chemical Symbols. Most typewriters contain 88 characters and a blank space. In a system to be mechanized there can be no *ad hoc* symbols. The 88 characters suggested here were selected to suit the current internal needs of Shell Development Co. They are sufficient to type most if not all internal chemical reports in a form which can be accepted by the computer and are also acceptable to humans. Where a choice was necessary between conflict with a chemical convention or with a language convention, preference was most often given to abridging the latter. Thus, a semicolon requires: colon, back space, comma; no question mark is provided; etc. When using this character set it will be necessary to spell out or abbreviate symbols represented by @, #, $, %, ¢, &, fractions, etc. Use of the comma half a space high makes the usual computer substitution of a dash for an apostrophe unnecessary. Some compromises were made on the chemical side. A triple bond requires: double bond, back space, single bond (or in reverse order). The arrows were placed on the diagonals deliberately, since horizontal arrows, as used in typing reactions, can be made by two or more single bonds with "greater than" or "less than;" and vertical downward arrows, by a vertical bond and a lower case "v."

A sincere plea for deliberate early efforts at compatibility is made to avoid repetition of the controversies which have at least occasionally attended efforts in chemical documentation. Faced with his own problems not everyone would select the same character set. Chemists or chemical documentalists are less likely to criticize the nine basic symbols than the absence of a triple bond and semicolon or the presence of the arrow bonds, the elongated diagonals or the circle R. It may be that some or all of the latter five should be considered as optional positions to be replaced in companion systems by alternate symbols of greater use in, let us say, coordination chemistry. Here dotted diagonals may be strong candidates. We recognize that we have, ourselves, introduced one point of incompatibility. In the United Kingdom the fat dot represents hydrogen attached to carbon rising above the plane of the diagram. As an alternative we propose the open dot for hydrogen or the highest ranking substituent rising above the plane (head of an arrow approaching) and the asterisk for hydrogen receding below the plane of the diagram (tail of arrow receding). This convention is sufficient to distinguish between the rotational isomers of chloro-bromo-fluoro-methane (Figure 7).

Conversion from Typewriter Codes to a Connection Table. Consider the simplified example of propylene typed as shown:

$$C{=}C\diagdown_{C}$$

The tape codes are successively C, =, C, INDEX, \, INDEX, C. The conversion program instructs the computer to assign assumed coordinates to the first typed symbol and to calculate the coordinates of subsequent symbols. If the starting position is assumed as space 20

Figure 7. Rotational isomers.

(positive to the right) = X and line 20 (positive downard) = Y, the coordinates will be as follows:

Code No.	Symbol		X, Y
1	First	C	20,20
2		=	21, 20
3	Second	C	22, 20
4		\	23, 21
5	Third	C	24, 22

A further instruction causes the computer to recognize that Codes 2 and 4 are bonds and 1, 3, and 5 are atoms. The nature of each bond code is next examined and its terminals are identified: The = is a double bond and horizontal. The necessary condition for terminals of the horizontal bond is that its coordinates relative to the horizontal bond be (X − 1, Y) and (X + 1, Y). This is, of course, satisfied by the first and second C atoms respectively, and it is unnecessary to examine coordinates of the third C. The \ is a single bond. The requirement for terminals of a diagonal bond downward to the right is that its coordinates be (X − 1, Y − 1) and (X + 1, Y + 1) with respect to coordinates of the bond. Neither condition is satisfied by the first C but both are by the second and third, respectively. So-called housekeeping instructions are also executed to be sure that each atom is bonded at least once but not more than is permitted by its valence, that all atoms and bonds are accounted for, that each bond connects two atoms, or that exceptions are provided for, etc. Finally, the connection information is entered into a preliminary table such as:

Bonds / Atoms	First	Second
	=	—
First C	X	
Second C	X	X
Third C		X

from which it can be converted into the form considered most desirable in a particular system.

LITERATURE CITED

(1) Feldman, A., Holland, D. B., Jacobus, D. P., J. CHEM. Doc. **3**, 187 (1963); "Survey of Chemical Notation Systems," NAS-NRC Publication 1150 p. 424, Washington, D. C., 1964.
(2) *Chem. Eng. News* **30**, 2622 (1952).
(3) Luhn, H. P., personal communication.
(4) Gordon, M., personal communication.
(5) Chemical symbol typing element, Part No. 67M, Camwil, Inc., 835 Keeaumoku St., Honolulu, Hawaii 96814
(6) *Sci. Information Notes* **7**, No. 3, 11 (June–July 1965).
(7) Ofer, K. D., Rice, C. N., Bourne, R. B., Logan, S. W., "A Pilot Study for the Input to a Chemical-Structure Retrieval System," 151st National Meeting of the American Chemical Society, Pittsburgh, Pa., March 1966.
(8) Waldo, W. H., DeBacker, M., "Printing Chemical Structures Electronically: Encoded Compounds Searched Generically With IBM-702," Proceedings, International Conference on Scientific Information, NAS-NRC, Washington, D. C., 1958, Vol. I, p. 711.
(9) Crane, E. M., Horowitz, P., "Hecsagon: A System for Computer Storage and Retrieval of Chemical Structure," 142nd National Meeting of the American Chemical Society, Atlantic City, N. J., Sept. 1962.
(10) DURA MACH 10, with edge card punch, edge card reader, first line finder wiring, machine stand with tape supply reel, left hand shelf, right hand shelf, electric tape winder, tape unwinder, tape tenna, Mach 10 coded. DURA BUSINESS MACHINES, INC., 32200 Stephenson Highway, Madison Heights, Mich. 48071.
(11) Jacobus, D. P., personal communication.
(12) *Ibid.*, Electronic symbol typing element, Part 220M.
(13) Gilmore, J. T., Savell, R. E., "The Lincoln Writer," Group Report 51-S, Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., Oct. 6, 1959 (AD 235-247).
(14) Hiller, L. A., Jr., *IRE Student Quart.* 8, 36 (Sept. 1961).
(15) Hiller, L. A., Jr., Baker, R. A., *J. Music Theory* **9**, No. 1, 128 (Spring 1965).