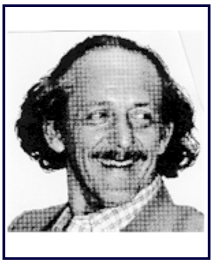
Reprinted in Essays of an Information Scientist, Vol:6, p.489-491, 1983



(Reprinted from Nature, Vol. 192, *No.* 4798, *p.* 192 *only, October* 14, 1961)

Chemico-Linguistics: Computer Translation of Chemical Nomenclature

I WISH to report what is believed to be the first successful mechanical translation of chemical names into chemical formulæ. Tsukermann and Terentiev[1] have indicated their intention to work on this problem.

Programming the initial experiments was simplified by imposing several restrictions. Simulated tests by manual methods have been successfully performed without these restrictions: (*a*) longest carbon chains allowed are ten carbon atoms in length; (*b*) maximum length of chemical names is 59 characters; (*c*) elements other than carbon, hydrogen, oxygen, nitrogen, sulphur, bromine, chlorine, fluorine and iodine were eliminated.

The most significant aspect of this research is the linguistic study of chemical nomenclature which preceded, for several years, the development of the computer programme[2].

In effect, a recognition grammar for organic chemistry has been written which can be used not only for mechanical indexing but also for teaching chemical nomenclature. Using Harris's methods[3,4] for grammatical analyses both the morphology and syntax of chemical nomenclature were analysed. Morphemes, allomorphs, homonyms, etc., were identified as well as their roles in the generation of molecular formulæ.

The difficult problem of calculating hydrogen values was overcome by application of the following formula, which is a modification of Soffer's[5] formula for determining double bonds and cyclic structures:

$$H = 2(\Sigma M_C - \Sigma M_{ab} + 1) + \Sigma M_N - \Sigma M_X$$

where M_i is an expression indicating the summation of the computational values represented by morpheme class i. For example, M_C is the number of carbon atoms implied by all morphemes in the chemical names which contain carbon. This completely resolved the tedious, though solvable, problem of constructing more complex rules for hydrogen calculation which are required when one attempts to follow the 'natural' inclination to record methyl as CH_3, amino as NH_2, etc.

It is the ultimate objective of this work to generate accurate molecular formulæ for all chemical names and also to display and print ideographs, that is, structural diagrams. Once the initial linguistic problems have been completely resolved, this can be done. It has already been shown that computers can be instructed to draw such diagrams[6,7]. However, neither Opler nor Waldo began with uncoded chemical names. It is one thing to programme a computer to draw a diagram for a known specific chemical. In this case, a human coder works from a diagram drawn by someone and calculates a cipher that is used by the computer to draw the same diagram. It is another matter to translate, without human intervention, from chemical names to structural diagrams.

The most difficult problem one encounters when trying to include all methods of naming chemicals, rather than just systematic International Union of Pure and Applied Chemistry names, is the ambiguity of different ring-numbering systems. While this does not affect the ability to calculate molecular formulæ, it does affect the ease of creating entirely accurate diagrams. In such cases the computer will render an invaluable service in reminding the chemist that several alternatives must be considered. It may also remind authors to stipulate whether they are following International Union of Pure and Applied Chemistry or some other system.

From a linguistic point of view, it is an interesting observation that the basic language of all naming systems in organic chemistry is essentially the same. While two chemists will name the same compound differently, both will be able to draw the same structural diagram. In this sense, the use of different systems corresponds to the problem of handling dialects rather than treatment of separate distinct languages. In addition, the complete understanding of the mechanisms for translating the language of organic chemistry provides useful experimental data for the more general problem of mechanical translation. If we cannot solve the problem of translating chemical nomenclature, there would seem to be little hope for translating natural languages such as English with a machine.

The following chemical compounds (Table 1) illustrate the capabilities of the present computer programme. While a *Univac* I was used for this experiment, the programme can be modified without

Table 1

Chemical name	Molecular formula
1,4-*bis*[*bis*(3-Diethylaminopropyl)amino]butane	$C_{32}H_{72}N_6$
2-[(2-Aminopropyl)ethylamino]ethanol	$C_7H_{18}N_2O$
8-Hydroxy-6-octene-2,4-diynenitrile	C_8H_5NO
1,3,3,4,4-Pentachloro-2-methyl*cyclo*butene	$C_5H_3Cl_5$
2-Phenyl-2,4,6-*cyclo*heptatriene-1-one	$C_{13}H_{10}O$
7-(2,4,5-Trichlorophenoxy)heptanoic acid	$C_{13}H_{15}Cl_3O_3$
Ethyl 2-cyano-5-phenyl-2,4-pentadienoate	$C_{14}H_{13}NO_2$
2-Naphthyl-2-*cyclo*hexen-1-one	$C_{16}H_{14}O$
Diphenyl-3-butynol	$C_{16}H_{14}O$

too much difficulty for smaller computers such as the I.B.M. 650 (ref. 8).

I wish to thank Dr. J. O'Connor and Mr. C. Meunch for their help in writing this programme, Dr. C. Borkowski for fruitful discussions on problems of mechanical translation, and Prof. N. Rubin, who acted as my chemical 'informant'.

E. GARFIELD

Institute for Scientific Information,
33 South 17th St.,
Philadelphia 3.

[1] Tsukermann, A. M., and Terentiev, A. P., *Proc. Intern. Conf. on Standards for a Common Language for Machine Searching and Translation*, 1, 493 (Interscience Press, 1960).

[2] Garfield, E., doctoral dissertation, Univ. of Pennsylvania (1961).

[3] Harris, Z. S., *Methods in Structural Linguistics* (Univ. of Chicago Press, 1951).

[4] Harris, Z. S., *Annual Report of the Computing Centre* (Univ. of Pennsylvania, Philadelphia, 1960).

[5] Soffer, M. D., *Science*, 127, 880 (1958).

[6] Waldo, W. H., and DeBacker, M., *Proc. Intern. Conf. Sci. Information*, 711 (Washington, 1958).

[7] Opler, A., and Baird, N., 133rd Nat. Meeting of the Amer. Chem. Soc., April 1958; *Amer. Documentation*, 10, 59 (1958).

[8] Garfield, E., *An Algorithm for Translating Chemical Names to Molecular Formulas* (Inst. Scientific Information, 1961).

Printed in Great Britain by Fisher, Knight & Co., Ltd., St. Albans.