Linguistics as a Basis for Analyzing Chemical Structure Diagrams*

Kirk Rankin** and Stephen J. Tauber†
National Bureau of Standards
Washington, D. C. 20234

Received June 7, 1971

Chemical structure diagrams constitute a language in the same sense that English and Fortran are languages. A language consists of combinations of members of a vocabulary. For chemical structure diagrams the vocabulary includes atom symbols, bond symbols, and numerical subscripts. In contrast to languages such as English and Fortran, wherein utterances are linear strings, the chemical diagram language uses arrangements of vocabulary members in two dimensions. The linguist seeks to construct a grammar which reflects the chemist's ability to distinguish between valid chemical structure diagrams and other patterns built from the same vocabulary. Such grammar should also account for "natural groups" in valid diagrams.

This note presents and motivates the claim that chemical structure diagrams are sufficiently language-like to warrant serious linguistic study. To our knowledge linguistic techniques for studying chemical structure information have been used only for nomenclature.[1,2] Since this journal is directed largely to chemists rather than to linguists, we begin by introducing some of the concepts and describing some of the activities in linguistic analysis, most importantly the basic concept of language itself.

LANGUAGE

A language is a set of strings constructed from some given vocabulary of symbols.[3,4] The strings are constructed by the operation of concatenation, which has the effect of combining the vocabulary symbols in a left-to-right order.

Consider English as an example of a language. For purposes of this rather simplified discussion, we assume that the vocabulary of symbols which combine into strings consists of *words* and *punctuation marks*. From the tens of thousands of English words and the many punctuation marks we will choose a much abbreviated vocabulary, V_1:

V_1 = {A, HOTEL, IS, LARGE,
NEAR, PARK, SMALL, THE, .}

The strings which constitute the language English are *sentences* of English. As your English teacher told you, not all strings of words are sentences. Some of the sentences which can be constructed by combining the symbols from V_1 are:

THE HOTEL IS NEAR A PARK.
THE HOTEL IS SMALL.
A LARGE PARK IS NEAR A SMALL HOTEL.

Another language that fits our definition is elementary algebraic notation. In this case, the vocabulary consists of such items as *variable* symbols (e.g., a, b, c) and *operator* symbols (+, −, ÷, =, etc.). We present a simplified vocabulary, V_2:

$$V_2 = \{a, b, c, +, -, \div, \times, =\}$$

The strings of symbols which constitute this language are the *equations* of elementary algebraic notation. Some examples are:

$$a + b = c$$
$$c - a = -b$$
$$c = c$$

Another language in our sense is FORTRAN (or any of the standard programming languages). The vocabulary consists of FORTRAN *terms*, such as the ones in the following highly abbreviated vocabulary:

V_3 = {.AND., GOTO, .GT., IF, .LE., N, .OR., 10, 17, (,), =}

The strings of terms which constitute the FORTRAN language are the FORTRAN *statements*, such as:

```
IF (N .LE. 10) GOTO 17
GOTO 10
N = 17
```

A final language which we will discuss is the set of chemical line formulas. Here the vocabulary of symbols consists of *atom* symbols, *bond* symbols, and *numerical subscripts.* An abbreviated vocabulary is:

$$V_4 = \{C, H, N, O, P, S, _2, _3, _4, _5, =, \equiv\}$$

Strings which are combinations of such symbols are the *line formulas,* such as:

$CH_3CH_2OCH_2CH_3$
H_2O
$CH_3CH_2OCH_2OCH_2CH_3$
$CH_3CH_2CH{=}NOH$

*Presented at the 6th Middle Atlantic Regional Meeting, ACS, Baltimore, February 5, 1971.

**To whom correspondence should be addressed.

†Present address: Informatics, Inc., 6000 Executive Blvd., Rockville, Md. 20852.

Contribution from the National Bureau of Standards. Not subject to copyright.

LINGUISTIC ANALYSIS

There are two central activities of linguistic analysis: isolation of grammatical strings and grammar construction.

In analyzing a language, the first task of the linguist is to study the full set of strings which in principle can be constructed from the vocabulary and to partition this set into two subsets: those which are grammatical and those which are ungrammatical. By definition, "grammatical" refers to those strings which are members of the language and "ungrammatical" refers to those combinations of elements of precisely the same vocabulary which are not members of the language. Taking our English subvocabulary, for example, the following strings can all be constructed from V_1, but only those labeled grammatical are English sentences:

THE A.
A HOTEL. NEAR
NEAR NEAR PARK A THE
grammatical: THE HOTEL IS NEAR A PARK.
grammatical: THE HOTEL IS SMALL.

Analogous situations occur in the study of the other three languages.

Algebraic Notation

$++$
$= a +$
$b\ c\ a \div$
grammatical: $a + b = c$
grammatical: $c = -c$

FORTRAN

.AND. .OR. GOTO
) .LE. (
IF .OR. .AND.
grammatical: N = 17
grammatical: GOTO 10

Chemical Notation

$C_3HC_2HOC_2HC_3H$
$CH_3{=}{\equiv}OH$
${}_{32}OHO$
O_5H
grammatical: H_2O
grammatical: $CH_3CH_2OCH_2CH_3$

We have been using the term grammatical in an intuitive sense. In the technical, linguistic sense, a grammatical string is one which conforms to certain underlying regularities.

For example,

THE HOTEL IS SMALL.

is grammatical. Equally grammatical are:

THE LARGE HOTEL IS SMALL.
THE LARGE HOTEL IS A SMALL PARK.

even though they are bizarre and contradictory. They are grammatical because they conform to the same regularities that underlie the grammatical (and nonbizarre) sentence above. These regularities can be expressed in terms of such notions as subject, predicate, agreement of verb with subject, etc.

In the case of chemical notation, the following line formula is grammatical:

$$CH_3CH_2OCH_2OCH_2CH_3$$

Equally grammatical is

$$CH_3CH_2OOOOCH_2OOOOOCH_2CH_3$$

even though it is bizarre because it does not describe a plausible compound. It is grammatical because it conforms to the regularities underlying more plausible grammatical line formulae; these regularities deal with subscripts having to follow element symbols and with acceptable valences. The type of regularities that a grammar can be expected to deal with are those concerned with the placement of symbols in the line formulas but not with characteristics of the molecules being represented such as bond lengths or conformation, which are not presented in the line formulas.

The second task for the linguist analyzing a language (after having made the partition into grammatical and ungrammatical subsets) is that of grammar construction. The linguist analyzing a language works through the users of the language and their intuitions concerning what is or is not grammatical. A formal grammar (or, simply "grammar") is a device which gives explicit account of the users' intuitions concerning the grammatical-ungrammatical partition. A grammar is basically a set of rules which put the vocabulary symbols together in the "right" combinations and never in the "wrong" combinations.

A grammar in fact is more. Not only does it put the symbols together into grammatical strings, it also makes certain claims about the structure of the grammatical strings. Exactly what aspects of structure a grammar should account for is a matter of linguistic controversy. For our present purpose, we assume that a grammar should account for the notion of "natural group" of vocabulary symbols. Those sequences of vocabulary symbols within grammatical strings which "go together" are natural groups. (Natural groups often have names: e.g., "subject", "predicate" in English; "carbonyl group", "methyl group" in chemical notation). For example the underlined word sequences in the sentence below are among the natural groups:

$\underline{\textit{THE HOTEL}}$ IS NEAR $\underline{\textit{A PARK}}$.

The underlined word sequences below are not natural groups:

THE $\underline{\textit{HOTEL IS}}$ $\underline{\textit{NEAR A}}$ PARK.

In chemical notation, we would expect a grammar to isolate the following sequences as some of the natural groups:

$$\underline{CH_3}\mathrm{CH_2}\underline{OCH_2CH_3}$$

but not the following sequences:

$$\mathrm{CH_3}\,C\underline{H_2O}CH\underline{{}_2\mathrm{C}}\mathrm{H_3}$$

Analogous comments could be made about natural groups in algebraic notation and in FORTRAN.

STRUCTURE DIAGRAMS

All the preceding discussion was based on the string (i.e., one-dimensional) nature of language. Structure dia-

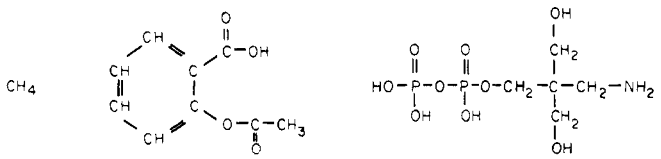

Figure 1. Grammatical structure diagrams

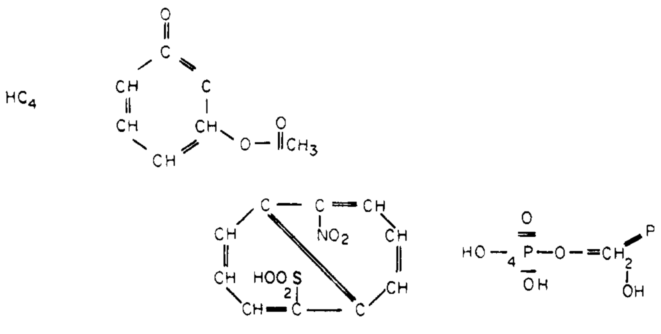

Figure 2. Ungrammatical juxtapositions of members of the structure diagram vocabulary

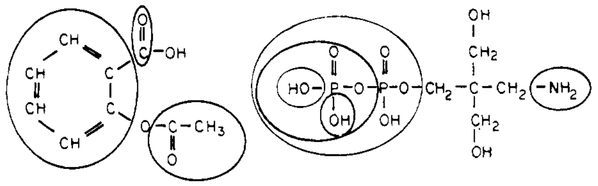

Figure 3. Natural groups in grammatical structure diagrams

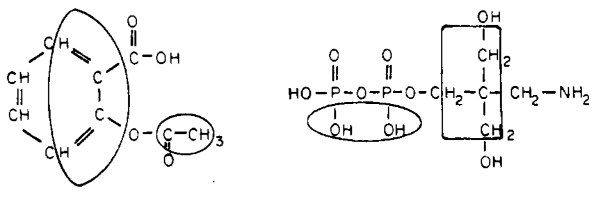

Figure 4. Happenstance groups within grammatical structure diagrams

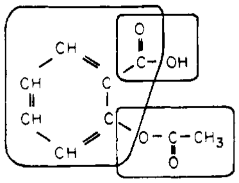

Figure 5. Overlapping natural groups in an *O*-acetylsalicylic acid structure diagram

grams—as they are normally used by chemists—are not in string form. The following discussion is based on modifying the definition of language given earlier. Instead of allowing only concatenation of symbols, we include sets of patterns in two dimensions. We will show that the concepts of linguistic analysis carry over.

Let a language be a set of patterns constructed from some vocabulary of symbols by juxtaposition.

In the case of chemical structure diagrams, the vocabulary consists of atom symbols, bond symbols, and numerical subscripts, as for the language of line formulas. However, the bond symbols are now distinguishable by orientation. V_5 is an abbreviated vocabulary of this type:

$$V_5 = \{C, H, N, O, P, S, {}_2, {}_3, {}_4, -, =, |, ||, /, //, \backslash\backslash\}$$

Some examples of grammatical patterns—i.e., valid structure diagrams are shown in Figure 1.

In analyzing the language of chemical structure diagrams, we would consider the full set of patterns that could be constructed by juxtaposing vocabulary symbols in various ways, and we would attempt as before to partition this set into a grammatical subset and an ungrammatical subset. The diagrams displayed in Figure 1 would be in the grammatical subset and those displayed in Figure 2 would be in the ungrammatical subset. These patterns are ungrammatical on various counts: atoms in the "wrong" valency condition, two bond symbols juxtaposed to each other, subscript and bond improperly juxtaposed, bond symbols of the wrong orientation juxtaposed with atom symbols, or else a graph violently distorted beyond what a chemist would write.

We would construct a grammar to embody chemists' intuitions concerning the grammatical-ungrammatical partition; as before, the grammar would make certain claims about natural groups. Thus it would claim that the encircled sub-patterns in Figure 3 are indeed natural groups. The grammar would deny status as natural groups to the sub-patterns encircled in Figure 4.

Preliminary results show that accounting for natural groupings will be much more difficult than in the non-chemical string languages. In chemical structure language, natural groups can overlap although neither group contains the other, as in the example of Figure 5.

We leave for presentation elsewhere examples of specific grammars for classes of chemical structure diagrams. For now, we satisfy ourselves by stating that these rules must deal with juxtaposition of atom symbols with one another, with subscripts, and with bond symbols, orientation of bond symbols, and valency, and that they should identify the natural groups.

LITERATURE CITED

(1) Garfield, E., "An Algorithm for Translating Chemical Names to Molecular Formulas," Institute for Scientific Information, Philadelphia, Pa., 1961.

(2) Elliott, P. M., and J. E. Rush, "Translation of Chemical Nomenclature by Syntax-Controlled Techniques," presented at 6th Middle Atlantic Regional Meeting, ACS, Baltimore, Feb. 5, 1971.

(3) Chomsky, N., "Syntactic Structures," Mouton and Co., the Hague, 1957.

(4) Hopcroft, J., and J. Ullman, "Formal Languages and their Relation to Automata," Addison Wesley, Reading, Mass., 1969.