[*International Conference on Machine Translation of Languages and Applied Language Analysis*, National Physical Laboratory, Teddington, UK, 5-8 September 1961]

LINGUISTIC ANALYSIS OF RUSSIAN CHEMICAL TERMINOLOGY

by

JOHN H. WAHLGREN

(University of California, Berkeley)

INTRODUCTION

THIS paper is a discussion of a specialized phase of linguistic research being carried on in the Machine Translation Project at the University of California in Berkeley. The material presented here is intended to illustrate in some detail the application of linguistic analysis to a particular problem. The fundamental approach upon which this work, is based has been described in a paper by Sydney M. Lamb.[1]

The first part of the following discussion deals with theoretical considerations underlying linguistic research into scientific terminology, with special reference to chemical terminology. The second part of the paper provides material which is illustrative of a linguistic description of chemical nomenclature. Examples are drawn from a detailed grammatical analysis of chemical terminology which is being conducted. Ultimately the results of this analysis will be incorporated into the total grammatical description of Russian which is to be employed in the machine translation process.

Relatively little attention is devoted here to the machine translation process, inasmuch as the application of the results of linguistic analysis constitutes a separate operation in the California Project. Some general comment on this aspect of the problem, however, will be made where necessary.

I. THEORETICAL CONSIDERATIONS

1.0 The continuing publication of large numbers of dictionaries, glossaries and lists dealing with terminology of many fields of science and technology is a strong indication that there exists, at least in the practical world of scientific writing and translating, an imposing "terminological problem" which is worthy of some special study. The decision to undertake ***linguistic*** study of the area of chemical terminology was

made without any preconceived ideas about linguistic peculiarities or limitations which that terminology might display. Only external, practical reasons prompted the setting aside of chemical terminology in the first place. It might otherwise just as well have been included in the overall structural analysis of the language, as it must be eventually in the final results. What linguistic peculiarities there may be in chemical (or any other) terminology remain to be determined in the course of investigation. Some basic assumptions, however, may be made.

1.1 The notion ***terminology,*** or even ***terminology of Science X i***s a notion which is not definable in structural linguistic terms. Scientific workers in a given field will agree more or less that a certain body of expressions which they employ, or, at least, with which they are familiar, belong to a special category which they will call the terminology of their science. In some cases, individuals among these workers will themselves have created some of these expressions, or they will be aware of the time, circumstances, and purposes of their creation. Scientific workers, furthermore, are all capable of creating new expressions which may become a part of the terminology. Terminology is intimately bound in with the development and practice of a science. It is an essential concomitant of all scientific endeavor. As a matter of fact, it is just this intimate tie, this "commitment", as it were, of these expressions to the concepts, methods, procedures and objects of a given science that makes them easily recognizable as terminology. Terminology, then, is terminology by virtue of its content, not by virtue of peculiarities in the linguistic structure of its expression. For example, one may employ the expressions "sulfuric acid" and "orange juice" in many environments which are structurally similar. One may, furthermore, make such grammatical statements as,

Sulfuric acid is a chemical compound.
Orange juice is a chemical compound.

But here the distinction between these two expressions begins to be revealed: the statement about sulfuric acid is chemically true and, more important, it is terminologically uniform. The statement about orange juice, on the contrary, is not only false in respect to specific fact, but it also would be regarded by a chemist as an inconsistent combination of a non-term with a term. A scientific term has a certain degree of general consistency with the theories, concepts, methods, or procedures of a given science; it has this sort of bond with the science over and above the individual reference it may have. Thus, both expressions "orange juice" and "sulfuric acid" refer to individual, clearly recognizable substances. But the science of chemistry does not frame its theories in terms of orange juice as an individual substance, and accordingly its name has no status as a chemical term. Sulfuric

acid, on the other hand, is a substance which has been singled out along with many others as significant and useful in the making of chemical theory, and thus its name has terminological status.

1.2 The restricted example just discussed is intended to characterize, but hardly to define terminology. For present purposes, it need not be defined. With such a characterization in hand we may obtain a body of expressions which is essentially the terminology of a field, and proceed with its linguistic analysis. Some further observations concerning the nature of terminology, however, may be useful in approaching the linguistic analysis.

1.3 Although terminology exists as a special phenomenon by virtue of the special role its expressions have in science, and not by virtue of structural distinctiveness, still it is not impossible that there may be a degree of ***correlation*** between terminology and linguistic structure. There are, for example, many areas of science and technology which have systematic aspects.[2] We may predict systematic linguistic correlates in at least some of these areas. Examples of structural correlation as well as lack of correlation are easily found. For instance, chemical theory classifies the chemical elements into groups within each of which the elements display great similarity in physical properties, combinatorial properties, etc. A comparison of the English names in group IA of the elements with those in group VIIA reveals a regular suffixal similarity within the groups and a contrast between them. If, however, the names within several other groups are examined, similarity in some of them is found to be small and contrast with other groups is not maintained.

TABLE I

chemical groupings with linguistic correlation		a chemical grouping without linguistic correlation
IA	VIIA	IVA
lithium	fluorine	carbon
sodium	chlorine	silicon
potassium	bromine	germanium
rubidium	iodine	tin
cesium	astatine	lead
francium		

Thus, within the entire system of chemical elements there is only a ***partial correlation*** between the terminological system (reflecting a certain chemical theory) and the linguistic structure.

1.4 In performing the linguistic analysis, and approaching terminological expressions from the point of view of their form, we note instances of structural systematicalness and attempt to discover correlation with terminological systems. In doing this, we take special note of another extremely important property which may be present. This property is ***productivity***. It may be viewed as a property of a system within the terminology which makes that system capable of being extended indefinitely. Linguistically, systems may be partially or wholly productive. An outstanding example of a wholly productive system is to be found in the naming of a class of chemical compounds called aliphatic hydrocarbons. The compounds in this class constitute a series in which the molecule of each succeeding compound is larger by one carbon atom than the preceding, and there is no definite limit on the size of the largest possible molecule in the series. One type of compound within this series (straight chain compounds) has a name which has a general form in Russian, ***R***-АН, where ***R*** symbolizes a root of numerical content and -АН is the actual shape of a suffix. Thus, a compound of this type containing five carbon atoms per molecule is named пентан, one with ten, декан, with fourteen, тетрадекан, and so forth. There are many further productive extensions on this basic system which yield a potential for literally millions of compound names, even without going beyond the bounds of "reasonable" molecule size. This, by the way, is an example of a productive terminological system for which normative descriptions exist in the literature of the science itself. See, for example, the 1957 Rules of the International Union of Pure and Applied Chemistry.[3]

1.5 Terminologies are generally capable of subdivision into more specialized areas connected with the major activities of a scientific or technical endeavor. One of the most important specialized segments of terminology is that one which is called ***nomenclature***. *A* nomenclature is a more or less self-consistent array of names applied to objects in a particular area of a scientific activity, as, for instance, the objects which the science investigates, the equipment and instruments, etc. Sometimes the term nomenclature designates the means by which names are generated, but the term is being used here to refer to the set of names itself. Hereafter in this paper a single item of terminology will be called a ***term*** a single item of nomenclature will be called a ***name***. More specific designations might be, for example, "chemical terminology", "electronic term", "nomenclature of chemical laboratory equipment", "botanical name", etc.

The particular orientation of this paper, as already suggested by previous examples, is toward that portion of chemical terminology which might properly be called "the nomenclature of chemical substances", but by common agreement among chemists is called simply "chemical nomenclature". In chemical terminology as a whole, chemical nomenclature is by far the greatest problem. It was therefore undertaken first, with a view to incorporating other areas of the terminology when their analysis is completed at a later date. In general, many chemical terms outside the nomenclature of substances are derivatives of the nomenclature, e.g., a chemical process, ацетилирование 'acetylation', from the name ацетил; a piece of equipment, хлоркальциевая трубка 'calcium chloride tube' from хлоркальций 'calcium chloride'.

1.6 Several characteristic features of terminology have been outlined: ***systematicalness***, ***productivity***, and ***specialization***. The degree to which these features are present varies widely in the terminologies of various sciences and fields of technology. Furthermore, the linguistic correlates of these features may be present in varying degrees or absent altogether. However, whatever correlates there are must be made apparent in any useful description of a terminology. A proper linguistic analysis will result in a description which is to the highest possible degree in accord with whatever features there may be in a particular terminology. If, for example, one were to overlook, the immense productivity of chemical naming systems and resort to the listing of all names which could be found in all chemical literature, one would not have made any provision for the vast number of new names which were being created at the very moment the list was being finished. Some estimates place the number of chemical substances described in the literature at over one million at the present time; the number of ***names*** greatly exceeds this by reason of the multiplicity of names applicable to a single substance. Even a limited recognition of productivity will not suffice. The breakdown of chemical names into word length elements and "obvious" component parts will fall short of the necessary goal of description to whatever extent it falls short of isolating the actual minimal productive elements. Thus, a basic criterion used here in the analysis of terminology is as follows: ***Segmentation*** must be carried out down to the minimal linguistically productive elements of the particular terminological system. Two other criteria place limitations, though not severe ones, on this first one: If a form which would otherwise be segmented proves to be ***homographic*** with some other form in the language, segmentation is not carried out. For example, the Russian chemical word пропил can be segmented into productive elements, проп and -ил; but there also exists in Russian an unsegmentable noun пропил 'a kerf', and so the segmentation is not made. The other criterion concerns the ***English representation*** in the

translation process. A form is not segmented if the segmentation would yield an incorrect English form upon combination of the English representations of constituents (this situation might be present when the English form is of an unproductive type). For example, there is a productive system of acid names in Russian wherein the suffix -OB regularly corresponds to English ***-ic***. However, the name ангеликовая кислота in Russian must be translated to English 'angelic acid' (not ****angelicic)***. The decision may therefore be made not to segment in this case.[4]

These three criteria form the main basis of the description of chemical nomenclature which is considered in the following section of the paper.

II. RUSSIAN CHEMICAL NOMENCLATURE

2.0 Chemical nomenclature, like other kinds of linguistic material, may be structurally described by (1) an enumeration of the distribution classes of elements, with the membership of each, and (2) a statement of the constructions into which the classes enter. A construction has the general form A B/C, which is a statement that "there are ***constitutes*** of distribution class C which are composed of ***constituents*** of classes A and B (in that order)". The information from this form of grammar may be readily adapted to the grammatical coding of dictionary entries for a machine system.[5]

Several paragraphs of the following partial grammar contain a few words of comment concerning methodology and problems of analysis in the particular area being considered. Some comment on possible application to other fields of terminology is also made.

2.1 Nature of the Corpus. The body of chemical nomenclature upon which the following description is based is made up of those names which can be characterized as chemical (according to some such considerations as were illustrated in paragraph ***1.1*** of this paper) and which are judged to be in any way possible occurrences in biochemical research literature.[6] It is not possible to state in detail in a report such as this exactly what areas are included or excluded, except to mention by way of example that certain nomenclature of industrial, metallurgical and mineralogical chemistry has been excluded. One area which has been included, but which poses particularly difficult problems in the determination of systematicalness and productivity, is the area of pharmaceutical abbreviated and trade names. Considerable study of them has been made and their structure is in general clear, but the requirement of maximal segmentation has not been satisfied with them.

It must be noted in connection with this preliminary step of the analysis (i.e. the delimiting of a corpus) that to some extent the limits of a terminology are indeterminate. There are a number of "borderline cases" in any field. If the decision is made that such cases are not terms of the science at hand, they will still in all likelihood be included in the dictionary either as general vocabulary items or as terms from some other field which are likely to occur, even though the dictionary is in general being designed for one major field.

2.2 ***Graphemics.*** The graphemic representation of Russian chemical names in this analysis follows the system of graphemic coding employed for Russian at the University of California MT Project.

Russian chemical names include in their graphemic form Cyrillic letters (transliterated, though not in the present paper), Latin letters, Greek letters (also transliterated), Arabic numerals, junctures (hyphen), punctuation and diacritics (comma, colon, parentheses, brackets, etc.). Letters occur both capitalized and small; letters and numerals sometimes occur as superscripts.

Chemical nomenclature is an example of a terminology having graphemic features which identify some, but not all, expressions as terms. For example пентан, гексоза, etc. are not graphemically distinct from Russian non-terms, but пентанон-3 or 2-метилпентен-3-овая-5 кислота are. Terminologies differ considerably in this respect. An extreme case is the taxonomic nomenclature of plants in botany which is purely Latin both graphemically and morphemically.

2.3 ***Lexemics.*** Segmentation with accompanying consideration of homography and English representation results in the isolation of lexical items or ***lexes*** which will be the dictionary entries of a translation system. Lexes are graphemic representations of ***lexemes,*** which are morphemic items. The description could be stated in terms of lexemes, hut is here maintained on the level of lexes in order that it may be more directly available for the form of dictionary entries.

The following paragraphs list the major distribution classes of lexes in Russian chemical nomenclature.

2.*31* ***Noun and Adjective bases.*** Certain chemical names are unsegmentable bases of nouns or adjectives. Those which are noun bases have a class symbol ***N-*** with a second position symbol ***m*** (masculine), ***n*** (neuter), or ***f*** (feminine), and, if undeclined a third position symbol ***u i***s employed. E.g. (only one or two examples of lexes are given for each class).

Nm - цези, кислород
Nn - олов (neuter nouns are quite rare among chemical names)
Nf - сурьм, кислот
Nnu - индиго

Rarely, an unsegmented adjective base may occur as a chemical name. Such adjectives are always masculine when having this function, e.g.

красн (a dye referred to simply as "red")

Ultimately all chemical names in Russian are, with respect to the grammar of the language as a whole, nouns or adjectives which function as nouns. Thus, the constructions into which chemical lexemes enter will lead eventually to constitutes which are also of one of the classes above, or a phrase of comparable class. The lexes shown above are simply monolexemic representatives of the class of all Russian chemical names. (Note that this step of the analysis establishes the status of terms in the grammar of the language as a whole).

2.32 ***Chemical Roots.*** Most bases may be segmented into elements which are not bases by the removal of prefixes and/or suffixes (q.v. below). These elements are lexes which are termed here ***chemical roots*** and given the general class symbol ***R.*** A subclassification indicated by second and third position symbols has been carried out, but is not given here except by way of example.

The series of aliphatic hydrocarbons mentioned previously (paragraph *l.4)* illustrates one subclass of roots, ***Rs.*** They are historically the bases of Greek numerals (except for a small subclass labeled ***Rss*** which contains such roots as эт in этан 'ethane'). These roots are characterized by their immediate co-occurrence with certain suffixes, e.g.

TABLE II

Russian Suffix	English Representation	Example
-ан	-an(e)	гексан 'hexane'
-ен	-en(e)	октен 'octene'
-ин	-yn(e)	пентин 'pentyne'
-ил	-yl	гептил 'heptyl'
-ат	-at(e)	does not occur

*(English graphs in parenthesis may be present or absent depending on certain morphographemic conditions

Another root subclass, ***Ra***, behaves differently in respect to these suffixes; for example, the root стеар :

-ан, -ен - do not occur

стеар + -ин - yields стеарин 'stearin'

-ил, -ат - yield стеарил 'stearyi', стеарат

Thus, root subclasses are determined primarily on the basis of their occurrence with certain suffixes and the English representation of such combinations (as in the examples пентин 'pentyne' vs. стеарин 'stearin').

The numerical roots of hydrocarbons occur in various other terminologies; a broader analysis of scientific terminology might still retain them in a single distribution class covering several fields. Combined with a suffix -од '-ode', for instance, they name a series of electron tubes, пентод, гексод , etc.

2.33 ***Chemical "Rootlike" Noun Bases.*** There are many forms which occur both as noun bases and as chemical roots. Such bases are a subclass of ***N,*** and have been given symbols ***Nr*** for masculine bases (which the vast majority of them are), and ***Na*** for feminines. Most of the monolexemic representatives of the class are either bases referring to chemical elements or organic bases unsegmented because of homography. Examples of such lexes are,

Russian	English	Examples with chem. suffixes
висмут	bismuth	вксмутил 'bismuthyl' висмутат 'bismuthate'
бор	bor/boron	борил 'boryl' борат 'borate' боран 'borane'
пропил	propyl	пропилиден 'propylidene'
камфор	camphor	камфорил 'camphoryl'

These classes have a number of subclasses reflecting co-occurrence with various sets of suffixes. They are chiefly important in the grammar as constitute classes (see para. 2.4 — Constructions).

2.34 ***Suffixes.*** Suffixes are of two main types — exclusively chemical (class symbol ***Z***), and those which also occur outside of chemical nomen-

clature (class symbol ***s***). Certain of the chemical suffixes ***(z)*** are homographic with "ordinary" suffixes, but may be segmented when occurring with unambiguous chemical roots or bases. The suffixes of chemical nomenclature are listed in ***table 3.***

The second position symbol for suffixes corresponds in general to the second position symbol of a constitute containing the suffix (e.g. пентоз is a constitute of class ***Na,*** -ОВ is a suffix of class ***za***). ***zo*** and ***so***, however, are special "connective" suffixes.

TABLE III

Class z			Class s		
Subclass:	Russ.	Eng.	Subclass:	Russ.	Eng.
zf	-аз	-as(e)	sq	-ов	-ic/-oic
za	-оз	-os(e)			-ate/-oate
zr	-ал(ь)	-al		-н	-ic/-ate
	-ол	-ol/-ole		-ист	-ous/-ide/-ite
	-он	-on(e)	so	-о/е	-о/ -Ø
	-ан	-an(e)			
	-ен	-en(e)			
	-ин	-yn(e)/-in/			
		-in(e)/-Ø/			
	-ил	-yl			
	-ид	-id(e)			
	-ит	-it(e)			
	-ат	-at(e)			
	(and a few others)				
zo	-а	-a			
	-и	-i			

2.35 ***Prefixes.*** For the present discussion, the major prefix classes[7] may be identified as ***uh*** (numerical — based on Russian numerical steins) ***um*** (numerical — based on Latin/Greek steins) ***uu*** (undifferentiated), for example,

uh — одно- , дву(х), трех-, etc.

um — мон(о)-, ди- , три- ,

uu — изо-, псевдо- , нор-, эпи- , пиро-, and others including various letter and Arabic numeral combinations with hyphen and other symbols.

2.4 ***Constructions***. The constructions are hardly amenable to illustration by way of example, inasmuch as they must form a consistent system in their totality. No attempt is made here to give a systematic account of constructions. The following examples are intended simply to illustrate the form they take and the way in which they account for combinations of lexes. The sets of constructions have been considerably simplified, both in regard to their form and the preciseness of subclassification of participating lexes. (Note, however, that some new subclasses have been introduced which have not been shown previously. Their characteristics are revealed to some extent in the examples. Subclass ***Nri***, for example, is one occurring initially with respect ***to Nrt*** /t — "terminal"/.)

(1) Constructions:[8] Examples:

Nm sq / Q водород -ист/ водородист

Q + Nm / Nm водородист натри/ водородист натри

Nm sns / NN = водородистый натрий

'sodium hydride'

Constructions of this form, of course, must be established for non-terminological Russian phrases as well, but in chemical nomenclature they will require translational rules to provide for necessary difference of order in the English representation.

The following sets exemplify constructions which are generally peculiar to chemical names.

(2) Rs zo / um тетр -а/ тетра

Rs zri / Nri эт -ил/ этил

um Nri / Nri тетра этил/ тетраэтил

(Nri :) Nm / Nm тетраэтил свинец/ тетраэтилсвинец

etc. =тетраэтилсвинец 'tetraethyl lead'

The notation (A :) means "A occurs zero or more times." Observe the following example.

(3) Rs zrt / Nrt пент -ан/ пентан
Nrt zrt / Nrt пентан -ол/ пентанол
(Nri :) Nrt / Nrt пропил пентан/ пропилпентан
этил пропил пентан/ этилпропилпентан
диэтил пропил пентанол/ диэтилпропилпентанол
etc.

(4) Rs zrt / Nrt пент -ен/ пентен
Nrt zri / Nri пентен -ил/ пентенил

(5) uh Nf / Nf пяти- окис/ пятиокис
Nf + Nm sgs / Nf пятиокис фосфор -а/ пятиокис фосфора
'phosphorus pentoxide'

III CONCLUSION

In this paper great stress has been laid upon the necessity of segmentation down to the minimal productive elements in a terminology. When this requirement is viewed in light of the problem of homography, it becomes clear that to be fully effective, similar segmentation analysis must be carried out wherever there are productive systems in the language. All this implies, to be sure, a great deal of work. In linguistic research, but holds promise in the long run for the machine handling of complex terminologies in an effective and accurate way.

The illustrative material in the foregoing discussion dealt specifically with chemical nomenclature. The examples, limited and simplified as they are, can do little more than give a general Impression of the shape which a detailed analysis has. The main point, however, should be clear, that linguistic analysis, just as in the language as a whole, reduces the large number of complex entitles in chemical nomenclature to a relatively small number of fundamental elements (lexes) which may be grouped into classes on the basis of their distribution. The possible combinations of these lexes are, in turn, accounted for by a series of constructions. A complete analysis of this kind should economically reveal the constituency of any chemical name, and, with the proper adaptation to the translation process, make

possible a correct translation utilizing a minimum number of dictionary entries.

NOTES

1. LAMB, S. M., "MT Research at the University of California," to appear in *Proceedings of the National Symposium on Machine Translation.*
2. This matter of linguistic correlation with non-linguistic systems has been discussed by Prof. M.B. EMENEAU in his article, "Language and Non-Linguistic Patterns" (Presidential Address, 24th annual meeting of LSA, 1949), *Language* 26.199ff, 1950. His article contains a chemical example along with a discussion of other types of organization — kinship, numerals, social structure — and their relationship to linguistic structure.
3. International Union of Pure and Applied Chemistry, Nomenclature of Organic Chemistry (1957 Rules), London, 1958.
4. The criterion of the English representation is of course not one which has anything to do with the pure description of the Russian terminology. What is being done here, in effect, is to "match" the analysis of Russian (considering its own productive elements, etc.) with a comparable analysis of English for translation purposes. As it happens, especially in the area of chemical nomenclature, there is a large degree of agreement between the two languages in areas where segmentation into productive elements is possible. This circumstance is obviously not accidental; the chemical terminologies of the languages have not evolved independently, but have grown up on the basis of much international exchange of chemical information and theory.
5. Individual dictionary entries are provided with a grammar code which indicates the distribution class of the lexical item or ***lex.***
6. Some of the sources of Russian chemical names in actual use as well as codifications of nomenclatural systems which were consulted are:
   a) AN SSSR, *Biokhimiya,* 1957, **22.**
   b) TERENT'EV, A. P., et al., *Nomenklatura Organicheskikh Soedinenij,* AN SSSR, Moscow, 1955.
   c) PEREL'MAN, V. I., *Kratkij Spravochnik Khimika,* Moscow, 1954.
   d) MOLOTKOV, I. G., Ed., *Khimicheskie Tovary (Spravochnik),* Moscow-Leningrad, 1954.
   e) BREJTBURG, A. M., *Biologicheskaya Khimiya,* Moscow, 1959.
   f) PAVLOV, B. A., and TERENT'EV, A. P., *Kurs Organichesko: Khimii,* Moscow, 1960.
   g) OREKHOV, A. P., *Khimiya Alkaloidov,* AN SSSR, Moscow, 1955.
   h) NAZAROV, I. N., and BERGEL'SON, L.D., *Khimiya Steroidnykh Gormonov,* AN SSSR, Moscow, 1955.

7. We use the symbol ***u*** in this project to designate the class of ***prefixes.***
8. Additional symbols not previously explained have the following meanings:

Q	--	Adjective base
sns	--	Nominative singular suffix
NN	--	Nominative Expression
sgs	--	Genitive singular suffix