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

MACHINE TRANSLATION OF RUSSIAN ORGANIC CHEMICAL NAMES INTO ENGLISH BY ANALYSIS AND RESYNTHESIS OF THE COMPONENT FRAGMENTS

by

LAWRENCE SUMMERS

(Machine Translation Research, Institute of Languages and Linguistics, Georgetown University, Washington, D.C.
Professor of Chemistry, University of North Dakota, Grand Forks, North Dakota, U.S.A.)

INTRODUCTION

THE Machine Translation Laboratory of the Georgetown University Institute of Languages and Linguistics has been working for some years on Russian-English machine translation. Different approaches to the problem, varying from essentially empirical code-matching techniques to techniques involving complete linguistic analysis, have been investigated and evaluated here. The procedure which was finally adopted and programmed for computers (and by which 100,000 words of Russian scientific prose have been translated) is the so-called GAT procedure, in which the text is submitted to a complete linguistic analysis within the computer. In the GAT procedure, dictionary look-up is a process of matching input items with an alphabetical list ("dictionary") containing Russian word stems or full-form words. When a test item is matched, it is assigned computer codes indicating its morphology, plus certain other codes for later use. Dictionary lookup of this general type is employed by most researchers now doing machine translation, and by many others now developing procedures which may eventually be applied to machine translation.

The vocabulary of organic chemistry cannot be handled satisfactorily in this way, however, because of the potentially infinite size of this vocabulary. The number of organic chemical compounds known is over one million; to enter the names of all these in a machine dictionary would be a very long task, and would also require a very large memory store, but computer scientists are developing large-capacity storage devices. However, the problem is not only that the vocabulary is very large (although it is), but that it is unlimited. Names of chemical substances are coined as needed. Many research papers in synthetic organic chemistry describe the preparation

of entirely new substances, whose names have never before appeared in the literature. (It is also possible, of course, that the name may be coined before the substance is made.) As a result, a machine dictionary containing the names of all known organic chemicals might still be unable to match any large fraction of the organic chemical names appearing in a given issue of the Journal of General Chemistry, for instance.

SUMMARY

We have developed procedures for machine translation of Russian organic chemical names into English by decomposition of the name into constituent Russian fragments, dictionary lookup of the fragments, and resynthesis of the English translation of the fragments. The routine has been programmed for computers (IBM 709-7090) and tested. Systematic or "rational" chemical names can be handled without great difficulty. In addition, our routines will handle the half-systematic or "half-rational" names perhaps more commonly used by chemists.

GENERAL DISCUSSION

A large amount of research has been devoted to problems of handling chemical formulas or names for purposes of indexing and data processing. These problems include machine conversion of names to chemical formulas, methods of indexing or encoding formulas, sorting and searching procedures based on chemical features, etc. (see the writings of Dyson, Wiswesser, Perry, Vleduts, Terent'ev, and many others). Our problem is different; we wish only to convert the Russian chemical name to an acceptable English equivalent. Thus from хлормасляной we do not wish to produce a chemical structural formula; we wish to produce chlorobutyric plus a morphological code, without having to enter the Russian word as such in the machine dictionary. The initial reaction of the chemist to this problem may be that it is not much more complicated than transliteration, because the chemist knows rules for systematic organic chemical nomenclature, and believes that chemists follow these rules. Actually, like any other set of normative rules for language, the rules of chemical nomenclature are not a description of the usages. (Even with names formed according to rule, problems arise—for a simple example from English, buta in butadiene is not to be equated to buta in butane, but is equivalent to but in butene. And Russian амино in аминобутан is to be brought into English as amino, but амино in бутиламином is not.)

Therefore, in this as in other areas of language, the language itself must be studied. For our purpose, the most useful classification of organic chemical names is that given by Terent'ev, as follows:

(I) ***Non-rational names.*** These are names which are ordinary words in the language, or which have been invented from some other point of view than that of chemical structure. They exist in a language on the same basis as other words, and must be handled like any other dictionary entry. Examples: глюкоза glucose; коронен coronene; янтарная кислота succinic acid.

(II) ***Rational names.*** These are systematic names, constructed in their entirety according to some accepted set of rules, so as to indicate the chemical structure and composition. Examples: 1,2-дибромэтан 1,2-dibromoethane; 2-бутиндиол-1,4 2-butyne-1,4-diol.

(III) ***Half-rational names.*** These are names constructed by a combination of the forms used in (I) and (II). They are therefore not systematic in their entirety, but only partly so. Examples: бромянтарная кислота bromosuccinic acid; диметилацетилен dimethylacetylene.

Our routines are intended to deal with Russian organic chemical names of types (II) and (III). Below there is given an outline of the procedures; this outline is followed by notes describing the "chemical linguistic" phenomena on which these procedures are based.

RESULTS

The procedure is as follows (see notes below):

(1) Submit the text to dictionary lookup; allow only items unmatched in the general dictionary to go to the routine for chemical names.

(2) Match the Russian text item, from the left-hand end, against tables of "chemical fragments".

(3) If a left-end string of characters in the text item is found to match a table entry, record the match, return to the head of the table, and repeat the matching process, beginning now with the leftmost unmatched character of the text item.

(4) If before any iteration of the matching process (including the initial one) the remainder of the text item is less than six characters in length, record this remainder and perform no further matching with this item.

(5) If the result of an iteration of the procedure is "no match", record the unmatched remainder of the item as in step (4).

(6) Submit the unmatched remainders to a second dictionary lookup in the main dictionary. (The matched portions may also be translated in this way, or their translations may be assigned from the table.)

(7) If a Russian item has not been completely matched as a result of steps (2) and (6), abandon that item (that is, send it on to succeeding routines in its original form, but do not attempt to furnish a "partial" translation).

(8) Reassemble the English translation, with its morphological codes (obtained in step (6); see also notes below).

The features considered essential to the satisfactory operation of this routine are the preceding main dictionary lookup (step 1), the matching procedure itself, the second dictionary lookup of unmatched remainders in the main dictionary (step 6), and the rule invalidating partial matching (step 7).

The computer program will depend on the particular translation routines in use. For example, our GAT programs are such that lookup of matched portions in the main dictionary will assign morphological codes, although actually only the right-hand fragment should carry such codes. For another example, it is possible to program this routine to get the effect of the second dictionary lookup without actually going twice to the main dictionary in the computer. Such programming considerations are not discussed here.

Notes:

(a) *Table 1* shows a list of about one hundred of the "building blocks" used in forming rational organic chemical names in Russian. This list was selected on the basis of our chemical experience, and that experience suggests that these building blocks will synthesize a very large proportion of rational organic chemical names. This list is not the table of "chemical fragments" used in the computer. *Table 1* defines the target; we wish to be able to deal with, at least, these pieces. How the computer control tables for accomplishing this are formed depends on tactical and programming considerations; we have altered our computer tables from time to time to meet programming requirements. The particular form of our computer tables is therefore of little interest here, and these tables are not shown.

(b) Soviet organic chemical writers follow the usages of international chemical nomenclature, and employ, in Cyrillic spelling, the international chemical "building blocks". These were originally derived from Latin (a few from Greek) word stems, and are therefore not Slavic. This fact contributes greatly to the feasibility of our routine. It is clear that the procedure described here represents a form of "lexemic analysis". A satisfactory lexemic analysis of any language is not yet available, although interesting work is being done. The problems of ambiguity, overlapping, partial agreement, "X-factors", etc., are numerous. In our limited area of study, however, overlapping of the chemical fragments ("lexemes"?) with syllables of words of the general Russian vocabulary is very much minimized.

(c) However, although the frequency of such overlapping is low, it is not zero. Study of the Russian vocabulary in general, and of chemical texts processed through the computer in particular, has produced a number of examples of such conflict—e.g., полу (a chemical fragment translating semi) with получать. This is the reason for the requirement (step 1) that the regular main dictionary lookup must precede this routine, and that only words not matched in the main dictionary be allowed to go on. This safeguards against possible mutilation and loss of text items present in the dictionary.

(d) *As* for text items not present in the main dictionary, but not actually chemical names, these will go into this routine. Since they were not in the dictionary, they could not have been translated anyway; but they may match in part with some of the chemical fragments. If partial translation is permitted, the form of the original text will be lost. This is the reason for the requirement (step 7) that there must eventually be a complete match. For example, the Russian proper name Полушкин would match with полу, and would give rise to something like semishkin, without this safeguard. If a complete match is required, the item must then be in fact a chemical name or a complete homograph of such name, and such homographs are extremely rare, for the reason mentioned under (b) above. We have as yet been able to find only one example, namely декан (the chemical decane, or dean of a faculty), and this is not a Slavic word even in the non-chemical meaning.

(e) The combining fragments listed in *Table 1* show some useful patterns of position distribution in the constructed name. For instance, изо occurs very frequently at the beginning of a chemical name, and never finally; ол occurs very frequently finally, and never initially (apart from олово or олеиновая and derived forms). Some use has been made of these position frequencies in our programming. For many of the fragments, however, the

position frequency distributions are rather flat, and not very useful. It is in the half-rational names of type (III) that there is a position-frequency situation which is quite useful (see below).

(f) The entries listed in ***Table 1*** will of course produce only rational chemical names. We would like to be able to deal also with the half-rational names of type (III), but the non-rational portions of such names are so numerous that it does not seem practical to include them in the matching tables. Apart from the increase in size of the tables, this would introduce a large number of conflicts within the tables, and also with the general vocabulary since these non-rational parts are very often Russian rather than Latin.

The half-rational names are formed in various ways. There is one very frequent pattern, however, in which the rational fragments of ***Table 1*** are employed as prefixes for a non-rational name of type (I). An example is метилкротоновая кислота, where метил (composed of systematic fragments) is prefixed to кротоновая кислота. Here кротоновая кислота is itself the common name of a chemical substance, so that кротоновая is an ordinary main dictionary entry. With high frequency in such cases, that part of the name which is an ordinary dictionary entry will form the final portion of the name. The procedure is therefore devised so that, as a last step, any unmatched right-hand remainder is recorded and submitted to a second dictionary lookup (step 6). Then the possibilities of handling names of this type increase with the natural growth of the dictionary, with no necessity for alteration of the tables of chemical fragments.

In a sample of 100 long organic chemical names picked at random from different pages of an issue of the Journal of General Chemistry, 42 were found to be rational (these would be handled by direct matching from our tables), and 33 were half-rational with the non-rational part in the final position (these would be handled as described here, if the non-rational part was in the dictionary). The rest of the sample consisted of half-rational names with the non-rational part not final (for these our procedures would fail) and of 5 examples of names not classified in these three groups. These 5 cases were also amenable to our routines. A procedure intended to deal only with rational names would therefore have failed with more than half this sample, but by providing for non-rational right-hand components it is possible to handle at least 80% of it. This is too small a sample for statistical conclusions; our statements about usages in chemical nomenclature are based on experience in chemistry, and the sample is cited mainly as illustration.

(g) The Russian text items involved are nouns or adjectives. The fragments out of which the rational chemical names are formed were defined originally in the Latin alphabet, and those that can occur finally end in consonants. As a result, if the item can be matched completely from our tables alone, it is a noun, masculine, inanimate, hard endings, first declension. (The only exception is аль which is commonly spelled with soft ending.) If the item is an adjective it will have full-form inflectional endings, which may indicate any gender, number, or case. Those of the noun forms which can become adjectival do so by addition of a derivational suffix such as -н- or -ов-, to which the inflectional ending is then suffixed (e.g., амин becomes аминный).

(h) Very frequent in the final position are two-character fragments such as ол or ен. These are noun forms, declined as indicated under (g) above, and the inflectional ending may therefore be as many as three characters, to make a total of five. The length test (step 4 above) for more than five remaining characters avoids mutilation of these final fragments and their inflectional endings. Application of this test in the initial iteration also is convenient, since if the item is less than six characters long it is surely not a compound chemical name.

(i) As a result of the procedure described above, the text item (unmatched in the main dictionary) is either matched from the tables, or not. If not, it is rejected. If there is a match, the remainder in the text item may be zero or not. If it is zero, English translations can be assigned as a result of the matching, and the item is a noun, masculine nominative-accusative singular (see part g above). If the remainder is greater than zero, it is either an inflectional suffix, a final fragment plus inflectional ending (see part h), or a non-rational portion which is itself a Russian word. For any of these remainders the procedure is the same, namely lookup in the main dictionary.

If the remainder is a non-rational portion which is a Russian word in the machine dictionary, the dictionary lookup deals with it as usual, no special dictionary entry being necessary. The regular dictionary updating procedures will increase the number of such remainders that can be handled.

The other possibilities require special dictionary entries, artificial in the sense that they are not words of the Russian language, but otherwise like any other dictionary entry. The possible inflectional suffixes (which are few in number—see part g) are not matched in the table, because the table has been so devised that they cannot be. They will therefore be remainders, and are entered in the dictionary like full-form words with

fixed morphological code, but with zero translation. The adjective-forming derivational endings and the remainders that can result from the length requirement are entered like word stems, with morphology to be assigned by the computer routines, and with an appropriate translation. The number of these "artificial" dictionary entries is fixed, and once the required set is entered no updating is necessary (unless, of course, the table itself is changed).

(j) Two problems not treated here are resynthesis of the English chemical name from the translations of the fragments, and the problem posed by prefixed, suffixed, or infixed numerical or alphabetical indicators, as in бутин-1,4-диол or бенз [а] антрацен. These are problems of tactics and programming (but their solution may present complexities).

The following will serve as examples of the above procedure:

Дибромбензол will be matched completely and translated as dibromobenzene, noun masculine nominative-accusative singular (because there is no remainder).

Дибромбензолы will be matched with remainder ы , translation dibromobenzene from the matching; in the dictionary lookup of remainders the "artificial" entry ы will be found with zero translation and the fixed morphological codes for noun nominative-accusative plural (there is no ambiguity with the feminine genitive singular because ы is marked in the computer as being the result of this routine).

Дибромуксусной will be matched through дибром, translation dibromo; in the dictionary lookup уксусной will then be looked up as the stem уксуск- which represents a Russian word, and the translation acetic and the morphological codes will be assigned.

Хлорэтаном will be matched through хлорэт translation chlorœth; in the dictionary lookup аном (which is a remainder because it is shorter than six characters) will be looked up as the artificial stem ан- which has been entered, and assigned the translation ane and the morphological codes.

CONCLUDING REMARKS

We have described here an analysis of the usages in Russian organic chemical nomenclature, and a practical procedure, based on that analysis, for machine translation of Russian organic chemical names into English. This procedure was devised for incorporation into our GAT system. These routines involve dictionary entries in the form of word stems or full-form words; the computer analysis of the text is carried out on morphological, syntagmatic, and syntactic levels. In machine translation programs based on a different type of analysis, or different dictionary procedures, a "chemical names" routine would be programmed differently. The same facts concerning the usages would have to be dealt with, however, and it seems very likely that similar tactical problems would be encountered.

Machine translation research at Georgetown has not been carried out in an attempt to furnish support for any particular ***a priori*** theory of the nature of language or of translation; instead, the objective has been to develop theories and procedures suitable for practical machine translation of actual texts not previously abstracted or analyzed. Certain working principles have been adopted, however; among these is the concept (Dostert) of initial investigations carried out on texts from one subject matter field at a time. It would follow that, after development by linguists and programmers of procedures for translation, specialists familiar with the particular subject-matter field should participate in the work. The initial subject-matter field was organic chemistry; it is for this reason that a chemist familiar with the special problems of chemical literature has been included in the Georgetown staff. It is anticipated that, as the work is extended to other subject-matter fields, specialists in those fields will also become associated with the work.

The tactics and programming methods developed for solution of this problem may find application in other cases, since there are other situations in language where the problems are similar to those encountered here. For instance, these procedures may prove to be useful for machine treatment of languages such as Arabic, in which inflectional affixes may be either prefixed or suffixed to the stem. Here there is the same problem of removing left-hand fragments whose translational meaning is fairly well defined, and of recognizing and translating the remaining word stem, which may itself carry derivational or inflectional suffixes. German compound words are another example; this problem has been studied by others (Reifler). From another point of view entirely, it seems clear that text-centered studies of scientific nomenclature and scientific writing should be interesting and useful to scientists themselves—for example, in the future design of

nomenclature systems. As we noted above, research along such lines is being carried out by a number of workers. To these different facets of the investigation we have as yet given little attention, however; our work, has been devoted essentially to the solution of the particular machine translation problem indicated by the title.

TABLE 1

Components for Rational Russian Organic Chemical Names

аконт	acont	геми	hemi	кис	kis
аз	az	ген	hen	лактам	lactam
азо	azo	гепт	hept	лактон	lacton/e/
азин	azine	гидразин	hydrazine	мет	meth
азксои	azoxy	гидразино	hydrazino	мета	meta
ал	al	гидр	hydr	моно	mono
аль	al	гидро	hydro	нафт	naphth
альдегид	aldehyde	гидрокси	hydroxyl	нафталин	naphthalene
альдоксим	aldoxime	гидроксид	hydroxide	нео	neo
алк	alk	гидроксил	hydroxyl	нитрозо	nitroso
амид	amid/e/	дек	dec	нитрил	nitrile
амил	amyl	децил	decyl	нитро	nitro
амин	amin/e/	ди	di	нон	non
амино	amino	додек	dodec	оил	oyl
ан	an/e/	додецил	dodecyl	оин	oin
анил	anil	ен	en/e/	окса	oxa
анилин	aniline	ид	id/e/	окси	oxy,hydroxy
антра	anthrax	ил	yl	оксид	oxide
арил	aryl	изо	iso	оксим	oxime
ат	at/e/	имид	imid/e/	оксо	oxo,keto
ацен	acene	имин	imin/e/	окт	oct
ацет	acet	имино	imino	ол	ol
бенз	benz	ин	yn/e/	олефин	olefin
бензол	benzene	иод	iod/o/	он	on/e/
би	bi	йод	iod/o/	орто	ortho
бис	bis	карб	carb	пара	para
бром	brom/o/	карбинол	carbinol	парафин	paraffin
бут	but	карбокси	carboxy	пент	pent
бутиро	butyro	карбоксил	carboxyl	поли	poly
гекз	hex	карбонов	carboxylic	полу	semi
гекс	hex	кето	keto	проп	prop

TABLE 1

Concluded

пропио	propio	тиа	thia	форм	form
пропионо	propiono	тио	thio	фтор	fluor/o/
псевдо	pseudo	толил	tolyl	хинон	quinone
сил	sil	толуил	toluyl	хлор	chlor/o/
спиро	spiro	толуол	toluene	циан	cyan/o/
спиран	spiran/e/	транс	trans	цикл	cycl
сульф	sulf	три	tri	цикло	cyclo
сулъфин	sulfin/e/	трис	tris	цис	cis
сульфо	sulfo	ундек	undec	эикоз	eicos
сульфон	sulfon/e/	ундецил	undecyl	эикос	eicos
тетр	tetr	фен	phen/e/	эн	hen
ти	thi	фенил	phenyl	эт	eth
				ят	at/e/