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.
Received February 8, 1962

The purpose of the present paper is to outline procedures which have been developed for machine translation into English of Russian names of organic chemical compounds. These procedures were devised for use in general-purpose digital computers, within the framework of the Georgetown University machine translation routines. The Georgetown translation programs are not limited to chemical matter; research on Russian–English machine translation at Georgetown has now been extended to a number of subject-matter fields. However, the initial research employed organic chemical texts, and the problem of handling the names of organic chemicals naturally arose.

In all machine translation programs, the input text is fed into the computer (via punched cards and magnetic tape at present, since a general-purpose commercial print reader is not yet available) and is then compared with a long list of forms which may occur in the source language. This long list is usually called the "dictionary." The purpose of the comparison, or matching, is to establish not only the identity, but also the properties (morphological, syntactic, and semantic) of an input text item. One distinction among machine translation procedures is the length of the string of input text which is regarded as a potential unit for such treatment. Most programs now operating take the original textword as the unit (a textword is, roughly, a string of characters occurring between two blanks, or a blank and a punctuation mark, in the original text), and break the textword into a word stem (which indicates identity) and morphological affixes (which indicate properties). In other types of programs, all possible morphological forms of a given stem may be entered in the dictionary (identity and properties then are assigned in a single match, at the expense of a much longer list), and items consisting of a number of textwords may also be entered in the dictionary. On the other hand, some procedures attempt a finer splitting of word stems themselves into component units (lexemes), and the dictionary then consists of lists of such smaller units.

In the Georgetown Automatic Translation (GAT) procedures for Russian–English translation, the first-mentioned tactics are employed. Individual textwords from the input text are treated, and are matched against a "main dictionary" which is a list of Russian word stems and, for some items, full-form words. The stem having been matched, the identity is assigned, and the remainder of the textword is then compared with a list of possible morphological affixes in order to assign properties, such as number, case, aspect, and others. (Here, and in other instances below, it will be understood that only the general logic is being described. The actual programs will often attain the objectives more indirectly, for tactical reasons.)

When all the textwords of the input text have been submitted to this procedure, there will remain a certain number of textwords not matched, because the appropriate entries were not present in the dictionary. The GAT dictionary has been built up, not by copying any existing printed dictionary, but by lexical abstraction from actual texts in the subject-matter field. Obviously, no dictionary can contain a very large fraction of all possible organic chemical names. The problem is not only that the number of such names is very large, but that it is unlimited. The situation is almost unique linguistically, and a special program to handle it was therefore required. But a point to be noted is that the GAT main dictionary, since it was compiled from organic chemical texts, does contain many names of organic chemicals, and the non-rational or trivial names of common organic substances are most likely to occur because they are most likely to have been encountered in the texts. For example, such items as UKSUSNAYA (acetic), YANTARNAYA (succinic), and GLYUKOZA (glucose) are present. Furthermore, the number of such items in the main dictionary will increase in the normal process of dictionary building, as more texts are translated.

The textwords which were not matched in the above-described main dictionary lookup may or may not be organic chemical names. If they are not, their identity will remain unknown (it is still possible that their properties can be defined, from the morphological affixes). But if they are chemical names the possibility exists that they can be analyzed into the component nomenclatural fragments by the computer, and thus translated. For this to be successful, the name must clearly be one actually formed from such fragments—that is, a Russian chemical name such as SHCHAVELEVAYA, if not present in the main dictionary (it is), could not be handled because it is a trivial name, but one such as BUTANDIOL should be susceptible to analysis.

Our objective, therefore, was to devise routines for recognizing whether a textword not matched in the main dictionary was a systematic chemical name, and for producing the English equivalent if it was. By "systematic" here is not meant "official," but rather "formed from recognizable fragments to which an English equivalent can be assigned." We are concerned only with producing an English equivalent of the name used by the Russian author, and not with establishing the "meaning" of the textword in any other sense. For our

*Presented before the Division of Chemical Literature, ACS National Meeting, Washington, D. C., March 22, 1962.

**Professor of Chemistry, University of North Dakota, Grand Forks, North Dakota, on leave 1960–1961 at Georgetown University.

purposes, organic chemical names are most usefully classified as rational, non-rational, or half-rational.[1] Rational names, such as BUTANDIOL (butanediol), or IZOOKTAN (isooctane), are those formed entirely from conventional chemical nomenclatural fragments. Non-rational ("trivial") names, such as SHCHAVELEVAYA (oxalic), are not so formed, but are simply Russian words like any other Russian word. Half-rational names, such as TRIMETILUKSUSNAYA (trimethylacetic), are formed in part from the conventional fragments, in combination with a non-rational part, here UKSUSNAYA (acetic).

From the preliminary study of Russian chemical names and of the practices of Russian authors, we concluded that such a recognition and translation program could be successful in a sufficiently large number of instances to make it a practical procedure. The background has been discussed previously,[2] with notes on the linguistic features and comment on the data-processing operations; the background discussion will not be repeated here. Two points, however, are of particular interest. First, Russian chemical nomenclature follows the international practices, using the same chemical nomenclatural fragments (in Cyrillic spelling). These fragments are based mostly on Latin, and are therefore Romance, not Slavic. The result is that if a long textword is matched completely with such Latin chemical fragments there is almost no possibility that it is a non-chemical Slavic word. This makes the matching process itself a recognition routine. Second, in half-rational Russian chemical names, as in English, the predominant situation is that the non-rational portion forms the *final* (right-hand) portion of the textword, and has the grammatical endings attached to it. This distribution pattern makes it possible for us to use a relatively short list (about 150 items) of possible combining fragments. Rational names, composed entirely of such fragments, would be matched completely in this list. Non-rational names, of course, never can be matched. But the half-rational names, which are very frequent in Russian as in English, can be handled in most instances by taking advantage of the position distribution described; the principle is to match as long as possible in the chemical fragments list and then, if there is a remainder, to return to the main dictionary to look up this remainder. If what remains is the non-rational (trivial) name of an organic chemical, it is a Russian word in its own right, and is likely to be in the main dictionary, as explained above.

Here again the logical operations just described may be accomplished by different tactics, depending on the programming. In fact, two different programs have been written for the chemical name translation routine, both for the IBM 7090. In one program the return is actually not to the main dictionary, but to a subsidiary dictionary abstracted from the main dictionary. In the other program, the main dictionary itself is used for this purpose, but it was found desirable to program so that every textword has a remainder, and all are handled alike. This is done by ceasing to match in the chemical fragments list if the remainder of the textword is less than six characters long. The advantage is that the right-hand end of the word, carrying the grammatical suffixes, is then always looked up in the main dictionary. The main dictionary lookup exists already as a morphological routine, which not only matches the textword stem but also analyzes the grammatical endings and assigns properties. Therefore if the matching in the chemical fragments list will never involve the right-hand end of the textword, this sub-routine involves a simple straight match, and all morphological (grammatical) analysis uses the main dictionary morphological program. The chemical fragments which may become right-hand ends of rational chemical names must then be entered in the main dictionary with their morphology, as "artificial" entries; this is a relatively simple thing to do.

Briefly, the procedure is as follows: Textwords not matched in the main dictionary go to this "chemical names" routine. Terminal and initial locant numerals and letters are removed and recorded. The Russian item is then matched, from the left, against a relatively short table of chemical fragments (Table I). That match is considered best which involves the longest matching table entry.[3] If a match is obtained, the English equivalent of the chemical fragment is recorded, the matched characters are removed from the textword, and the matching process is repeated. If before any iteration of this process the remaining portion of the textword is less than six characters long, or if in any iteration no match is obtained in the table of chemical fragments, this matching process ceases, and the remaining characters are recorded as a "remainder." Finally, all the remainders are looked up in the main dictionary, and if found are analyzed grammatically and assigned an identity and properties.

Independently, Wahlgren[4] has developed a very similar procedure for Russian–English machine translation of chemical terminology.

To the organic chemist, this problem may at first sight appear a rather simple one because the chemist knows rules for nomenclature, and assumes that he and other chemists follow these rules and that the rules are themselves logical and consistent. In fact, the operation offers more complications than the chemist might foresee, although there are certainly many more difficult problems of machine translation. Initially, we considered seriously the possibility that a simple set of re-spelling rules could be used. But then half-rational names cannot be handled (DIKHLORGEPTAN can be respelled immediately as *dichlorheptan*, and with a little trouble as *dichloroheptane*, but respelling KHLORYANTARNAYA as *chloryantarnaya* is not useful); in addition, the recognition problem becomes serious. For if letter-by-letter re-spelling is used, all Russian words which come to the program will obviously be treated. To insure that the textword is a proper candidate for such chemical respelling would require matching strings of its characters as a recognition procedure, and therefore strings of characters might as well be matched as a translation procedure. If the input were, not a text, but a list of Russian words known beforehand to be organic chemical names, the problem would be different.

After the general procedure was decided on, the chief problem was of course the construction of the list of chemical fragments. A list of chemical nomenclatural units believed to be sufficient to handle a large proportion of organic chemical names was constructed; the list is given in reference (2). This list represents the target. It cannot be, itself, the table of chemical fragments for use in the computer, because the computer knows nothing of chemistry, and can make no judgments or allowances

Table I. Chemical Fragments List for Georgetown Chemical Names Program

Notes: Entries marked * are not used initially, but are switched into the list only if some portion of the textword already has been matched. Where more than one English equivalent is given specific sub-routines make the choice in each case. Entries marked ** have in Russian the reversed E.

*AL	al	EĬKOS**	eicos	*LAKTAM	lactam	PENTA	penta
*AL'	al	EĬKOZ**	eicos	*LAKTON	lacton/e/	PENTAKONT	pentacont
*AL'DEGID	aldehyd/e/	*EN	en/e/	MET	meth	POLI	poly
*AL'DOKSIM	aldoxim/e/	EN**	hen	META	meta	POLU	semi
ALK	alk	ET**	eth	METAN	methan/e/	PROP	prop
ALKA	alka	FEN	phen/e/	*MID	mid/e/	PROPA	propa
*ALYUMINI	aluminum	FENIL	phenyl	*MIN	min/e/	PROPIO	propio
*AMID	amid/e/	FORM	form	MONO	mono	PROPIOL	propiol
AMIL	amyl	FTOR	fluor/o/	*N	n/e/	PSEVDO	pseudo
AMIN	amin/e/	GEKS	hex	NAFT	naphth	SIL	sil
*AN	an/e/	GEKSA	hexa	NAFTALIN	naphthalen/e/	*SPIRAN	spiran/e/
ANIL	anil	GEKZA	hexa	NEO	neo	SPIRO	spiro
ANILIN	anilin/e/	GEMI	hemi	NITRIL	nitrile/e/	SUL'F	sulf
ANTR	anthr	GEN	hen	NITRO	nitro	SUL'FIN	sulfin
ARIL	aryl	GEPT	hept	NITROZO	nitroso	SUL'FON	sulfon/e/
*AT	at/e/	GEPTA	hepta	NON	non/e/	TETRA	tetra
*ATSEN	acen/e/	GIDR	hydr	NONA	nona	TETRAKONT	tetracont
ATSET	acet	GIDRAZIN	hydrazin/e/	*O	o	TI	thi
AZ	az	GIDRIN	hydrin	*OFORM	oform	TIA	thia
AZA	aza	GIDROKSI	hydroxy	*OIL	oyl	TOLIL	tolyl
AZIN	azine	GIDROKSID	hydroxid/e/	*OIN	oin	TOLUIL	toluyl
AZOKSI	azoxy	GIDROKSIL	hydroxyl	OKSA	oxa	TOLUOL	toluen/e/
BENZ	benz	*ID	id/e/	OKSAZ	oxaz	TRANS	trans
BENZOL	benzen/e/	*IL	yl	OKSAZA	oxaza	TRI	tri
BI	bi	IMID	imid/e/	OKSI	hydroxy, oxy	TRIAKONT	triacont
BIS	bis	IMIN	imin/e/	OKSID	oxid/e/	TRIS	tris
BROM	brom/o/	*IN	yn/e/, in/e/	OKSIM	oxim/e/	TSIAN	cyan/o/
BUT	but	IOD	iod/o/	OKSO	oxo	TSIKL	cycl
BUTA	buta	ĬOD	iod/o/	OKT	oct	TSIS	cis
BUTIR	butyr	IZO	iso	OKTA	octa	TSISTEĬN	cistein/e/
DEK	dec	KARB	carb	*OL	ol	TSISTIN	cistin/e/
DEKA	deca	KARBIN	carbin	*OLEFIN	olefin	UNDEK	undec
DETS	dec	KARBOKSI	carboxy	*ON	on/e/	UNDEKA	undeca
DI	di	KARBOKSIL	carboxyl	*ONITRIL	onitril/e/	UNDETSIL	undecyl
DODEK	dodec	KET	ket	ORTO	ortho	*ZAN	zan/e/
DODEKA	dodeca	KHINON	quinone	PARA	para	*ZO	zo
DODETSIL	dodecyl	KHLOR	chlor/o/	PENT	pent	*ZOL	zole

for ambiguities, fortuitous coincidence of spellings, and so on. The list then was reworked and modified in order to deal with these tactical difficulties. As an example of the tactical considerations, the following may be considered: To handle BUTEN (buten), the fragments BUT (but) and EN (en(e)) would presumably be provided in the list. But matching of BUTADIEN against this list will now match BUT, and leave ADIEN. To handle this the (structurally meaningless) A could be included in the list. One-letter fragments are undesirable, because all Russian words beginning with the letter in question will obviously give a match. The fragment BUTA (buta) can be provided in the list. Then BUTADIEN will be matched BUTA, DI, and EN, and translated *buta di en(e)*. But now a textword BUTANOL will encounter BUTA in the list, and this will leave NOL. To deal with such instances, we eventually provided N n(e) as an artificial fragment, but programmed so that this N cannot be used in the first matching attempt (to prevent all Russian words beginning with N giving a match). This one instance is simple enough, but multiplication of such problems causes the construction of the chemical fragments table to require a considerable amount of labor. A chemical fragments table now in use is shown in Table I.

To show how the above-described program operates, a few examples may be cited. (Table I is to be used for reference.)

Example I: DIKHLORETANOM.—This is matched successively with DI, KHLOR, and ET. The chemical fragments table contains AN, but matching stops after ET because the remainder is then less than six characters. The remainder ANOM is then eventually looked up in the main dictionary, where the artificial Russian word AN has been supplied, coded as noun, masculine, hard endings. This lookup is a morphological analysis program, and the instrumental ending OM will be analyzed. The translation is *dichloroethane* coded noun masculine singular instrumental.

Example II: DIKHLORUKSUSNAYA.—This is matched with DI, then KHLOR. Matching now fails. The remainder UKSUSNAYA eventually goes to the final main dictionary lookup, where it is matched with the stem UKSUSN-(acetic), a good Russian word which is in the GAT dictionary. Translation, *dichloroacetic*, adjective feminine singular nominative.

Example III: BUTLEROV.—This is a Russian family name, which comes to this program because it was untranslated in the original lookup of the text in the Russian-English dictionary. It is matched with BUT;

matching then fails in the chemical fragments table, leaving the remainder LEROV. Eventual main dictionary lookup of this remainder finds no such entry in the dictionary. The whole operation is therefore cancelled, and the original textword is simply transliterated and printed out in the translation as *Butlerov*, which is in this instance the proper treatment.

Example IV: BROMKAMFORSUL'FONOVAYA.—BROM is matched, then matching fails, and the remainder is eventually looked up in the main dictionary. There KAMFORSUL'FONOV- is not found (although KAMFOR and SUL'FONOV- are present). The treatment is cancelled, and the original textword remains untranslated; in this example our program fails, although the cancellation rule prevents any erroneous translation from being printed out.

Example I is a rational name; Example II a half-rational name with the non-rational part UKSUSN- at the right-hand end. These are the types the program is expected to translate. Example IV is a half-rational name with the non-rational part KAMFOR not at the extreme right. This type is fortunately less frequent. It is obvious that, by extension of the program along the same lines, and by extension of the dictionary and chemical fragments table, such types might be handled.

Thus our program will translate a large fraction of Russian organic chemical names, and in case of failure does not produce false translations. We believe that, with enough labor on the control lists and with extension of the programs on the basis of the same principles, translation of organic chemical names could be accomplished, not only from Russian to English, but in general from any language into whatever other form of representation might be desired.

REFERENCES

(1) A. M. Tsukerman and A. P. Terent'ev, Proc. Int. Conf. for Standards on a Common Language for Machine Searching and Translation, Vol. I, Interscience Press, New York, 1961; others have used this classification.

(2) L. Summers, Proc. 1st Intl. Conf. on Machine Translation of Languages and Applied Language Analysis, Natl. Physical Lab., Teddington, England, 1961.

(3) *Cf.* E. Garfield, "An Algorithm for Translating Chemical Names to Molecular Formulas," Inst. for Sci. Information, Philadelphia, 1961, p. 35.

(4) J. H. Wahlgren, Proc. 1st Intl. Conf. on Machine Translation of Languages and Applied Language Analysis, Natl. Physical Lab., Teddington, England, 1961.

A New Method for the Search of Scientific Literature Through Abstracting Journals*

By A. O. CEZAIRLIYAN, P. S. LYKOUDIS, and Y. S. TOULOUKIAN

Thermophysical Properties Research Center, School of Mechanical Engineering, Purdue University, Lafayette, Indiana

Received September 13, 1961

A new method of literature search using abstracting journals is presented. A model is adopted and by mathematical analysis relations are obtained which yield quantitative results. Statistical experiments are conducted on the literature covering the area of thermophysical properties of all matter to verify the analytical results. The experimental results are found to be in agreement with the theoretical predictions. The costs of literature search by the proposed method and another method (consisting of going through the volumes of abstracting journals page by page) are compared. It is found that the proposed literature search method gives a considerable cost reduction. Some statistical data pertinent to the research literature on thermophysical properties are given.

I. INTRODUCTION

The object of this study is to present a method for literature search using abstracting journals. A model is used for this purpose and by mathematical analysis useful relations are obtained which are verified experimentally. The reason for the investigation of the new method of search is to reduce the time and search cost in comparison to the cover-to-cover search method.** The proposed model is the following: An investigator starts searching the abstracting journal beginning with the most recent issue and goes back number of years a. Then, he searches the bibliographies of the papers that were located in the a years interval for new references. In this second step, a period of $(b-a)$ years is covered. Next, he starts using the abstracting journal again for a direct search of another a years starting from the year b. This may be repeated in cycles until the desired period of years is covered. The details of the approach and the derivations of the resultant equations are not presented in this paper. The reader interested in the mathematical procedures employed in the derivations may refer to the two studies on this subject.[1,2]

* The study reported here was conducted at the "Thermophysical Properties Research Center," School of Mechanical Engineering, Purdue University, and was sponsored in part by the National Science Foundation—Science Information Service.

** The cover-to-cover search method consists of going through the volumes of the abstracting journals page by page. It can be noted that there is also a method of literature search by the use of indexes, but in this study it is not considered as a basis for comparison, since the literature search by the use of indexes does not give the most nearly complete yield on publications.