[*Mechanical Translation*, vol.4, nos.1 and 2, November 1957; pp. 14-27]

Multiple Correspondence†

Roderick Gould, Computation Laboratory, Harvard University, Cambridge, Massachusetts*

It has been shown by Oettinger that the usefulness of rough Russian-English translations produced by an automatic dictionary is limited primarily by the large number of English equivalents which must be provided for many Russian words. The design of an additional machine stage for reducing the number of equivalents requires that the words be somehow classified; this classification might be according to meaning, grammatical role in the sentence, or both. Detailed examination of a model automatic-dictionary output revealed that the multiple-correspondence problem arose primarily from nouns, prepositions, and verbs, in that order. However, the extremely small number of distinct prepositions involved suggests that they should be given special individual treatment. It is proposed that the "meaning words" (nouns, verbs, etc.) of Russian and English be classified according to meaning and the "function words" (prepositions, conjunctions, etc.) be omitted from consideration. Lists of meaning-class sequences appearing in large samplings of Russian text would be tabulated and stored in the translator; comparison with these tabulated sequences would then allow the number of different classes of English words corresponding to any given Russian word to be reduced.

AN AUTOMATIC dictionary, as proposed by Oettinger, [1] is a machine for making rough translations of technical literature from one language into another. The machine contains a glossary of words in the input language and appropriate equivalents in the output language. When each successive word of a text in the input language is introduced into the machine, the corresponding equivalents in the output language are printed out. The original word order is unchanged. Almost no grammatical information, such as that given by tense or case endings, is preserved. Punctuation and mathematical symbols are passed through the machine unaltered.

When Oettinger prepared a text translation simulating the output of an automatic Russian-English dictionary and submitted it to a number of English-speaking subjects, he found that "The most frequent criticism was levelled at the excessive number of alternatives given for a single Russian word in some instances. " He concluded that "The absence of grammatical detail and the retention of the Russian word order seem to be of secondary importance only," and "... the proper selection of English correspondents is by far the major problem facing a reader..."

It is the purpose of the present paper to investigate some possibilities for refining the output of a Russian-English automatic dictionary by reducing the number of English alternatives for each word in the original text. Two approaches to the problem present themselves. The first is the reduction of the number of English equivalents provided in the glossary. The second involves an additional machine stage between the glossary and the output; in this stage a refining process would select the best equivalents for each word on the basis of the context.

It is certainly desirable to provide only a small number of English correspondents for each Russian word in the glossary, for conservation of storage space as well as for clarity of

---

† This paper has been adapted from Progress Report No. AF-45, The Computation Laboratory, Harvard University, Cambridge, Massachusetts.

* Now at Centre d'Etude et d'Exploitation des Calculateurs Electroniques, Brussels, Belgium.

1. Oettinger, A. G., "A Study for the Design of an Automatic Dictionary, " Doctoral Thesis, Harvard University, April 1954.

output. However, it is also essential that no important senses of the word be lost, or the text may become unintelligible to the reader. Since very few words in one language have one and only one correspondent in another, the great majority of dictionary entries will represent a compromise between these two goals.

The task of compiling the glossary will be simplified by a restriction to some specific scientific field. In this case, those word meanings having particular relevance to the field can be stressed, and specialized meanings unrelated to the field can be eliminated. The progress currently being achieved in the design of permanent storage media for electronic computers would seem to make this idea practical. For example, in such a photographic storage system as the "flying spot store" described by Ryan, [2] a number of specialized vocabularies could be stored, each on its own set of glass plates. The proper glossary to suit a given foreign text could then be inserted manually into the automatic dictionary.

It is hard to see how an optimum choice of word equivalents for even a specialized Russian-English glossary can be made without the aid of large-scale experiments on reader comprehension of machine output text. However, it is possible to establish some intuitive principles for minimization of the number of correspondents for a given Russian word:

(1) Try to select an English word, or words, covering the same range of meanings as the Russian word. Conversely, try to avoid English words having important senses which do not correspond to the Russian word.

(2) Include equivalents for all common senses of the Russian word; but be willing to omit the less common senses, particularly if they are at all suggested by the English words already selected. Sacrifice fine shadings of meaning.

(3) Preserve alternative grammatical roles which the Russian word may assume in English translation.

The problem of designing an additional operation in the machine is a much more complicated one than reducing the length of the entries in the glossary itself. The choice of alternative words on the basis of context as it is done by human beings[3] does not seem to be a process which can be mechanized. Since each of several consecutive foreign words may be provided with multiple English equivalents by the glossary, a refining device must be given some basis for choosing permissible sequences of alternatives from the myriad possible sequences. These facts seem to suggest a classification scheme which would distinguish between some, if not all, of the English alternatives for each Russian word.

The idea of an English word-classification scheme involving several hundred word classes has been proposed by Yngve. [4,5] He suggests that extremely large samples of English text be analyzed, each word be assigned to a class primarily on a grammatical basis, and all possible word class sequences of "phrase length" be listed. Sequences of phrases would then be tabulated, and so on up to sentence length. The method of approach to the problem of word classes to be adopted here is rather different from Yngve's, although his work will be alluded to occasionally.

Consideration will now be given in some detail to the question of distinguishing between English alternatives obtained from the output of an automatic dictionary. It will be useful to work with a sample output text. The one chosen is the model automatic-dictionary output mentioned above, constructed and used by Oettinger. It was derived from a Russian article whose title reads, in English: "The Application of Boolean Matrix Algebra to the Analysis and Synthesis of Relay-Contact Networks." The full text in Russian, a complete English translation, and a model dictionary output may be found in Reference 1.

---

2. Ryan, R.D., "A Permanent High Speed Store for Use with Digital Computers, " Transactions of the IRE. Vol. EC-3, No. 3, September 1954.

3. Kaplan, A., "An Experimental Study of Ambiguity and Context, " Technical Report P-187, The Rand Corporation, Santa Monica, California, November 30, 1950. Reprinted in Mechanical Translation. Vol.2, No. 2, November 1955.

4. Yngve, V.H., "Syntax and the Problem of Multiple Meaning," Machine Translation of Languages (W. N. Locke and A. D. Booth, editors). The Technology Press of M.I.T. and John Wiley and Sons, Inc., New York, 1955.

5. Yngve, V.H., "Sentence-for-Sentence Translation, " Mechanical Translation, Vol. 2, No. 2, November 1955.

Since the multiple-alternative problem is essentially one of multiple meaning, it is natural to consider word classification on the basis of meaning alone. One such classification scheme has already been set up, and has been in use for over a hundred years: Roget's Thesaurus. This work contains a large number of English nouns, verbs, adjectives, adverbs, and phrases, listed under slightly more than 1000 categories according to meaning or concept. These categories were set up with reference to general writing and are not well adapted for specialized scientific text. Still, some insight into the present problem is afforded by the classification of a small part of the model output text according to Roget's scheme. The Thesaurus used was the Authorized Edition, Revised 1941.

In Table 1 the first sentence of the Russian paper is given as it might appear in the output of an automatic dictionary. When a Russian word is provided by the dictionary with several English correspondents, these are enclosed in parentheses. The symbol "N" within the parentheses indicates that the word can sometimes be eliminated completely. One addition to the model output has been made by the present writer. In each case of multiple choice, the English word considered by an expert in the field of the article to be the best alternative is shown underlined. Thus the words outside parentheses, together with those underlined, constitute a nearly optimum word-for-word translation. In freer translation, the sentence reads: "In recent times Boolean algebra has been successfully employed in the analysis of relay networks of the series-parallel type."

In Table 2 the words of the model output are listed in columnar form. Next to each word, one or more appropriate categories from Roget, identified both by number and name, are given. The choice of categories was done not on the basis of the English words themselves but according to their usage as equivalents of the original Russian word. For example, the second English word shown, "at, " is listed in Webster's Collegiate Dictionary (Fifth Edition) as having six distinct meanings. However, "at" is important here only as a possible translation of the Russian word "v." The listing of the latter in the Russian-English dictionary used for reference, A. I. Smirnitskij's Russko-Anglijskij Slovar', appears to use "at" in only three of its six senses. Therefore, only these three were sought in Roget. Only one could be located. Where one or more pertinent senses of a word could not be located in Roget, an asterisk appears.

It should be noted that Roget categories seldom have a one-to-one correspondence with senses listed in a dictionary. A single category may include a number of concepts distinguished by Webster's.

As may be seen from the tables, most of the words could be located satisfactorily in the Thesaurus. Of those words having senses which could not be located, seven are prepositions. The Thesaurus contains no prepositions, and its categories are not well adapted to them. The remaining unplaced words include four words of a technical nature and two other words, "time" and "tense." The latter is a specialized grammatical term which probably should not have been included in the original glossary.

The Roget classification was quite successful in distinguishing between the various correspondents to a single Russian word. In no case do more than two correspondents fall in the same category, although two do so fairly frequently.

A listing of permissible sequences of word-meaning classes for use with an automatic dictionary can be obtained only through the analysis of very large samples of written material. The output of an automatic dictionary is arranged in Russian word order and according to Russian grammatical principles, e.g. there are no articles ("the," "a"). Therefore, word class sequences obtained from English text are of little or no value. It would appear that what is required is a tabulation of sequences of word meanings found in Russian language text. From this point of view, the categories shown in Table 2 are to be regarded as designations of the various senses which the original Russian word can assume. For example, consider the word "posledovatel'nyj," which is translated in Table 1 as "(series, successive, consecutive, consistent)." Inspection of a large sample of Russian scientific writing might show that a word used to indicate "Continuity "(i. e. unbroken sequence) sometimes occurs following a word indicating "Parallelism" and preceding a word denoting "Junction" or "Combination," but that words used to indicate "Sequence, " "Uniformity, " or "Agreement" never occur in

Table 1

(In, at, into, to, for, on, N) (last, latter, new, latest, lowest, worst) (time, tense) for analysis (and, N) synthesis relay-contact electrical (circuit, diagram, scheme) parallel - (series, successive, consecutive, consistent) (connection, junction, combination) (with, from) (success, luck) (to be utilize, to be take advantage of) apparatus Boolean algebra.

Table 2

(In	221 Interiority, *
at	199 Contiguity, *
into	294 Ingress, 300 Insertion
to	278 Direction
for	*
on)	*
(last	67 End
latter	63 Sequence, 122 Preterition
new	123 Newness
latest	118 The Present Time
lowest	649 Badness, 851 Vulgarity
worst)	649 Badness
(time	106 Time, *
tense)	*
for	*
analysis	49 Decomposition, 461 Inquiry
(and)	88 Accompaniment
synthesis	48 Combination, 54 Composition
relay-	*
contact	199 Contiguity
electrical	157 Power, *
(circuit	*
diagram	554 Representation
scheme)	626 Plan
parallel-	216 Parallelism
(series	69 Continuity
successive	63 Sequence
consecutive	69 Continuity
consistent)	16 Uniformity, 23 Agreement
(connection	43 Junction
junction	43 Junction
combination)	48 Combination
(with	88 Accompaniment, *
from)	*
(success	731 Success
luck)	156 Chance
(to be utilize	677 Use
to be take advantage of)	677 Use
apparatus	633 Instrument, 692 Conduct
Boolean	*
algebra	85 Numeration

this position. It would then be established that "posledovatel'nyj, " in the sentence translated in Table 1, could be given by the English words "series" or "consecutive" but not by "successive" or "consistent." The number of English alternate equivalents is thus halved. This principle could easily be extended so that Russian words requiring no English correspondent (i.e. the "N" alternative) would be eliminated altogether.

It must be recognized, however, that listing all word-meaning class sequences for the very large sample of Russian text that would be required represents a tremendous task. Each part of the sample would have to be read by a person well acquainted with the Russian language, who would assign to each word a meaning class designation (e.g. a Roget category number) according to its sense in that particular sentence. Alternatively, this might be done by an English-speaking person with the aid of an "unrefined " automatic dictionary. Once these class designations were assigned, tabulation of the sequences could be done comparatively easily on a digital computer.

A further problem is that the number of categories would have to be very large. If Roget's scheme were extended to cover technical material and perhaps to include more preposition-concepts, it would have to include perhaps 1200 categories at the very least. This figure yields 1.7×10^9 possible sequences of only three-word length. If the word class sequence method is to be effective, it is desirable that a large proportion of the possible sequences be ruled inadmissible. This is also a necessity from the point of view of storage of the admissible sequences. What proportion of the possible sequences might actually occur in written material is difficult to gauge. It would, of course, be essential to obtain a valid estimate before embarking upon such an ambitious project.

When a word is classified solely on the basis of the concept which it expresses, a certain amount of grammatical information is thrown away. In all Indo-European languages, words can be classified roughly into conventional groups called "parts of speech:" nouns, verbs, adjectives, and so on. These parts of speech assume fairly clear-cut roles in the construction of sentences. A noun meaning "a walk" and a verb meaning "to walk" belong to the same meaning category as far as Roget is concerned, but there is no reason to assume that the two words will occur in the same word—meaning class sequences. It is quite probable that they will not. If this is true, there may be reason for differentiating between the two words in the assignment of word classes.

The part of speech concept is of interest in another regard also. Since these basic distinctions between words do exist, it is pertinent to ask whether the multiple-meaning problem is more serious for some parts of speech than for others. Furthermore, these part of speech distinctions are not invariant in a translation between two languages; a word which is one part of speech in one language may sometimes translate into some other part of speech in another language. Also there exist homographs, pairs of foreign words which have identical spelling but quite different meanings, whose English correspondents must be lumped together in an automatic dictionary. One may wish to ask how often a Russian form will have English correspondents which belong to two or more part of speech groups. In order to shed light on such questions as these, Oettinger's model automatic-dictionary output was examined in some detail.

The Russian article contains 236 different word stems. In making up an English glossary for these stems, Oettinger strove to keep his entries general rather than slanted toward the text at hand. For each Russian word he listed English correspondents for all the important general senses and also for any technical meanings relevant to the electronic literature. The complete glossary and more detailed information about its construction are contained in Reference 1.

The division of words into part of speech classes as done by orthodox grammarians is not based on consistent definitions. Another scheme, which will be used here, is that devised by Fries. [6] His plan, illustrated in Table 3, is one of functional definition by means of contexts or "test frames" into which other words are substituted. Groupings of words are formed according to whether the words will fit into certain arbitrarily chosen contexts. The groupings are designated as Classes 1-4 and Groups A-O. However, since there is no functional distinction between a Class and a Group, both will be referred to here as classes. Since the groupings were formed on the basis

6. Fries, C.C., The Structure of English, Harcourt, Brace and Company, New York, 1952.

Table 3

FRIES' WORD CLASSES

(Adapted from Reference 6)

Name	Frames	Examples
Class 1	(The) _ was /were good The __remembered the __ The __went there	concert, difference, reports clerk, husband, tax, food team, husband, woman
Class 2	(The) 1 ____ good (The) 1 ___ (the) 1 (The) 1 ___ there	is, was, seem, become remembered, saw, signed went, started, lived, met
Class 3	(The) ___1. was/were __*	good, large, foreign, lower
Class 4	(The) 3 1 was/were 3 __ (The) 1 remembered (the) 1 __ (The) 1 went __	there, always, suddenly clearly, especially, soon out, upstairs, eagerly
Group A	__ 1 was/were 3 4	the, no, your, many, two
Group B	A 1 __ be/been 3 4	may, could, has, has to
	The 1 __moved/moving/move	had, was, got, kept, had to
Group C	The concert may ___ be good	not #
Group D	A 1 B 2 __ 3 (e.g. The concert may be ___ good/better)	very, any, too, still
	A 1 2 __ 4 (e. g. The men went __ down)	(a) way, very, much
Group E	The concerts ___ the lectures are ___ were interesting ___ profitable now___ earlier	and, or, not, nor, but, rather than #
Group F	A 1 __ A 1 2 ____ A 1 (e.g. The Concerts __ the school are __ the top)	at, by, of, across
Group G	__ the boy/boys 2 their work promptly	do/does/did #
Group H	__ is a man at the door	there #
Group I	_ did the student call	when, why, where, how
Group J	The orchestra was good ____the new director came	until, when, so, and, since
Group K	_ that's more helpful**	well, oh, now, why #
Group L	_ we're on our way now**	yes, no #
Group M	__ I just got another letter**	say, listen, look #
Group N	__ take these two letters**	please #
Group O	__ do them right away	lets [sic] #

* Word must fit both positions.
** Additional constraints, based on meaning, are used here.
\# All members of word class are listed.

of a large sampling of spoken English, many of them have little relevance for written text. Fries makes a point of giving no explicit definitions for his word classes. Particularly for this reason, nearly all comments made here about this classification system are the responsibility of the present writer.

Some general relations exist between Fries' plan and the conventional scheme. Class 1 words correspond in a general way to nouns and pronouns, class 2 to verbs other than auxiliaries, class 3 to most descriptive adjectives, and class 4 to adverbs which modify verbs. Class A words are "determiners," certain adjectives and other words which appear immediately before nouns. Class B consists of auxiliary verbs. Class D contains adverbs which modify adjectives. Conjunctions which join words and incomplete clauses are found in class E; conjunctions and other words which join complete clauses are in class J.* Class F contains the prepositions and class I the interrogatives. The present writer has included participles in class 3, and has added a new class P for abbreviations ("i.e. ") and certain phrases. For the purposes of this study, classes 2 and B and classes E and J have been combined.

The model automatic-dictionary translation was surveyed and each correspondent of each word in the original Russian was assigned to a word class, according to its usage in English as a translation of the Russian word. Smirnitskij's dictionary was the main reference for establishing this usage. In several cases the English correspondents were made up of two or more words rather than one. These phrases were treated as though they were single English words where possible. For example, the English correspondent for "naprimer" is the phrase "for example;" this was regarded as a member of class 4, rather than as a class F word followed by a class 1 word. Phrases like "one can, " which did not fit any Fries grouping, were assigned to class P.

In the majority of cases, the correspondents of a single stem were members of a single word class. Whenever the alternative "N" occurred, it was assigned to the same word class as the other correspondents. When there was a single English correspondent which fitted more than one word class, it was assigned to the one most appropriate class. The occurrences of the stems having correspondents of a single class have been tabulated in Table 4 according to the number of English correspondents and their class. Each of twenty Russian stems in the paper had English correspondents which fell into more than one word class. These stems will be treated separately later.

It is evident from Table 4 that nearly all of the multiple correspondence problems involve word classes 1, 2/B, 3, E/J, and F. The number of occurrences q of Russian words having their correspondents in each of these classes is plotted, in Fig. 1, against the number of English alternatives n. In Fig. 1, the class 1 curve stands well above the others in number of occurrences. The remaining curves lie fairly close together, except for the class F curve's large peak at n = 7.

The "Multiplicity Index" given in Table 4 is arrived at by summing the products of the number of correspondents n and number of word occurrences q within each word class for n > 1, or

$$M.I. = \sum_{n=2}^{\infty} nq_n .$$

This gives a first approximation to a linear measure of the multiple choice problem presented by each word class. The weighting by n is convenient but arbitrary, since it is not clear per se that, for example, a Russian word having four English correspondents presents exactly twice the problem of a word having only two.

Class 1 has the largest Multiplicity Index, 279. Class F follows closely with 233. The class 2/B Index is about half of that, and the Indices of classes 3 and E/J are still smaller The other Multiplicity Indices are negligible.

---

* Some difficulties appear in connection with class J. Consider the three sentences:

I wonder which he stopped.
I wonder which stopped him.
I wonder between which he went.

The first "which" is obviously a class J word, but the disposition of the others is not so clear. All such words have been assigned to class J. Pairs such as "if.. .then, " not mentioned by Fries, have also been included in class J.

Table 4

RUSSIAN STEM OCCURRENCES IN TEXT

by Number and Class of Correspondents

Word Class	No. of Correspondents "n"									Total	Multiplicity Index	Relative Multiplicity
	1	2	3	4	5	6	7	8	9			
1	141	34	34	12	7	3		1		232	279	1.20
2/B	34	24	11	6	2					77	115	1.49
3	60	23	1	1		1				86	59	.69
4	12	3								15	6	.40
A	18	1								19	2	.11
C	2									2		
D		1								1	2	
E/J	12	28		2	5					47	89	1.89
F	13	28	5	8			16		2	72	233	3.24
P	8									8		
Total	300	142	51	29	14	4	16	1	2	559	785	1.40

Table 5

DISTINCT RUSSIAN STEMS

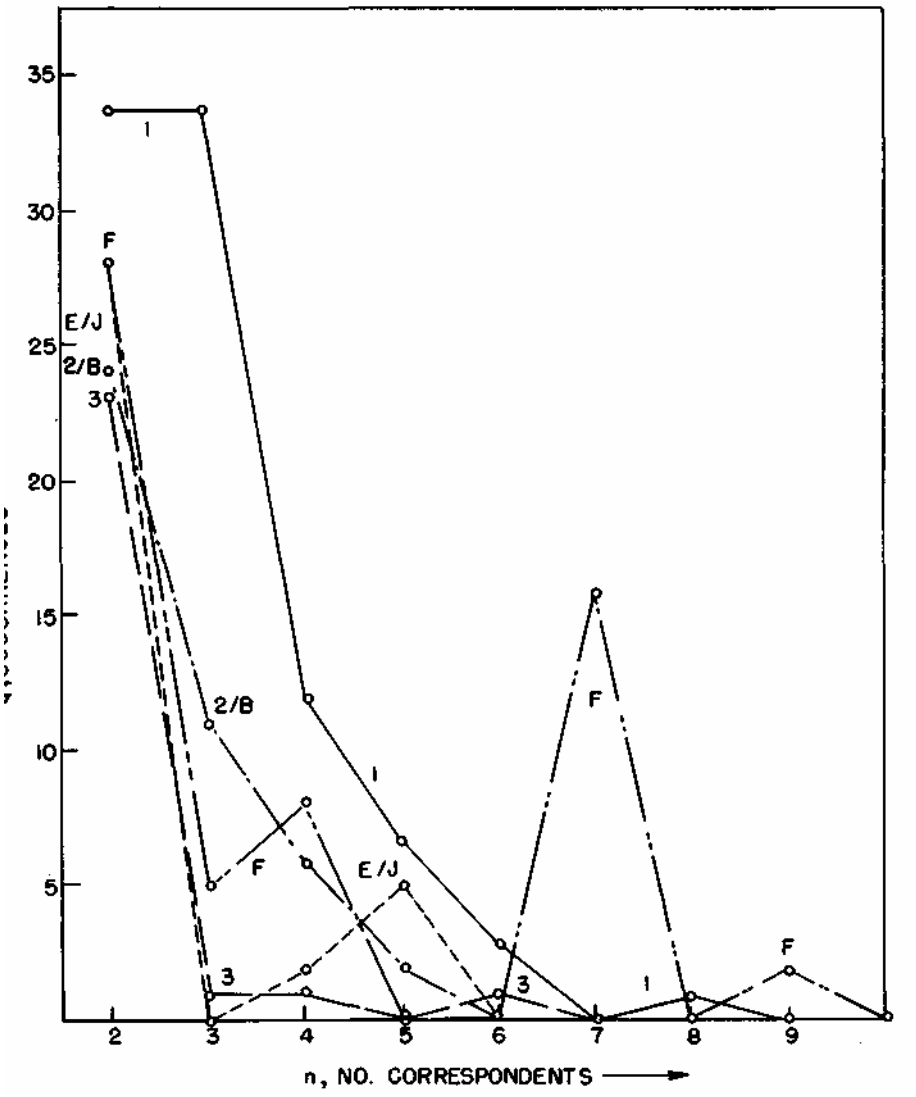
by Number and Class of Correspondents

Word Class	No. of Correspondents "n"									Total	Average Occurrences per Stem	Av. Correspondents per Stem
	1	2	3	4	5	6	7	8	9			
1	32	19	8	6	4	2		1		72	3.2	2.19
2/B	16	13	6	3	2					40	1.9	2.05
3	31	18	1	1		1				52	1.7	1.54
4	7	3								10	1.5	1.30
A	4	1								5	3.8	1.20
C	1									1	2.0	1.00
D		1								1	1.0	2.00
E/J	5	4		1	1					11	4.3	2.00
F	2	4	3	1			1		1	12	6.0	3.25
P	2									2		1.00
Total	100	63	18	12	7	3	1	1	1	206	2.7	1.97

The "Relative Multiplicity" is defined as the Multiplicity Index divided by the total occurrences for a word class:

$$R.M. = \frac{\sum_{n=2}^{\infty} nq_n}{\sum_{n=2}^{\infty} q_n}$$

Class F achieves its high Multiplicity Index in spite of the relatively small number of occurrences (72) of class F words in the sample. This fact is reflected by a Relative Multiplicity much larger than that of any other word class. The numbers of distinct Russian word stems producing the occurrences shown in Table 4 are tabulated in Table 5. Thus, for example, the 232 occurrences of class 1 words are produced by repeated occurrences of 72 distinct stems, so that each stem appears 3.2 times on the average; while the 72 occurrences of class F words are produced from 12 distinct stems, an average of 6.0 appearances per stem. It is particularly interesting to note that the 16 appearances of. class F words having 7 alternative correspondents, shown in



Occurrences of Russian Stems with Multiple Correspondents

Fig. 1

Table 6

COMPARISON OF MEANING AND FUNCTION WORDS

Word Classes	Total Occurrences	Multiplicity Index	Relative Multiplicity	Total Distinct Stems	Average Occurrences per Stem	Average Correspondents per Stem
1, 2/B, 3, 4	410	459	1.12	174	2.4	1.91
A, C - P	149	326	2.19	32	4.6	2.25

Table 4, are produced by repetition of a single Russian word. If this one stem were eliminated from the sample, the Multiplicity Index of class F would be reduced from 233 to 121.

The final column of Table 5 gives the average number of English correspondents for distinct Russian stems of each word class. This quantity is as small as 1.00 for certain word classes and ranges to 2.19 for class 1 and 3. 25 for class F.

It has been remarked by a number of observers that English words can be divided into two large classifications: the "meaning" words and the "function" words. Yngve[4] describes the latter as "... mostly grammatical words — articles, prepositions, conjunctions, auxiliary verbs, pronouns, and so on— the words that have so aptly been called the cement words. These are the words that provide the grammatical structure in which the nouns, verbs, adjectives, adverbs are held."

Fries[6] makes a similar distinction between his Classes 1-4 and Groups A-O. "In the four large Classes, the lexical meaning of the separate words are rather clearly separable from the structural meanings of the arrangements in which these words appear. In the words of our fifteen Groups it is usually difficult if not impossible to indicate a lexical meaning apart from the structural meaning which these words signal." * Fries found that each of Classes 1-4 had hundreds of members, but that in his entire language sampling the members of Groups A-O numbered only 154.

Although the number of distinct function words is small, these words make up a large proportion of the total word occurrences in English. Fries found them to be about 1/3 of the total in his verbal materials. According to the Eldridge word count, the 55 most frequent English words make up about half of ordinary newspaper text. Most of these are function words.

Table 6 shows the results of grouping the information of Tables 4 and 5 concerning occurrences of Russian stems into Fries' Classes and Groups. It should be remembered that not all of the stems in the sample are included, but only those whose English correspondents were all of one word class. However, the several correspondents of the twenty omitted stems are distributed fairly evenly between meaning and function words. The inclusion of Group B with Classes 1-4 probably has not affected the values appreciably, since the use of auxiliary verbs is not common in Russian.

Words of Groups A - P make up more than a fourth of the total occurrences. One would expect this proportion to be much less than the 1/3 quoted by Fries, for two reasons. First, Fries was dealing with conversational material, which in English at least is likely to contain a particularly high proportion of words of little meaning content; these fall into Groups A-P. Second, in Russian, word-endings fulfill many grammatical functions which in English require the use of function words. The figure of 1/4 is therefore higher than might have been expected.

---

* The prepositions, Group F, might seem to present an exception. But Fries points out that for the words "at," "by," "for," "from," "in," "of," "on," "to," "with," the average number of separate meanings given in the Oxford English Dictionary is 36 1/2! The lexical meaning apparently is at best an extremely vague one here.

Table 7

TWENTY RUSSIAN STEMS

with English Correspondents and their Word Classes

Russian Stem and No. of Occurrences		Correspondents	Word Classes										Combined Classes				
			1	2/B	3	4	A	C	D	E/J	F	I	P	1	α	β	2/B	γ
Homographic																		
l-	1	only, as soon as				x				x					x			x
mozh-	1	to be able, power	x	x										x			x	
prost-	1	simple, common, prime, stoppage	x		x									x	x			
uzh-	1	already, narrower			x	x									x	x		
Non-Homo.																		
vsiak-	2	any, every, anyone, all sorts of, anything	x				x							x	x			
vtor-	1	second, the latter	x		x									x	x			
dovol'n-	1	content, pleased, rather, enough			x	x			x						x	x		
drug-	5	other, different, the rest	x		x		x							x	x			
es-	5	to be, to eat, O.K.		x									x				x	x
eshch-	2	still, yet, as far as, only, some more	x			x	x							x	x	x		
in-	1	different, other, some			x		x								x			
kazhd-	1	each, every, everyone	x				x							x	x			
kak-	7	how, what, as, like, when, N								x	x	x						x
neskol'k-	2	several, some, somewhat, slightly	x			x	x		x					x	x	x		
odn-	5	one, the same, alone, nothing but, a certain	x		x		x							x	x			
ostal'n-	2	the rest of, the others	x				x							x	x			
pus-	3	let, though		x						x							x	x
sam-	1	the very, the same, most	x				x		x					x	x	x		
tak-	5	such, so, a sort of					x		x						x	x		
cht-	5	what, which, that, why								x		x						x

The Multiplicity Indices indicate that, despite their small number of occurrences, the function words contribute on the order of 2/5 of the alternate-choice difficulties. The average number of English correspondents is quite similar for the two word groups. This is perhaps accounted for by the fact that the prepositions have a great range of meaning, while the other function words have little range.

The Average Occurrences column of Table 6 shows that the meaning words are repeated, on the average, over half as often as the function words — seemingly a high figure. It is probable that meaning words receive much more repetition in scientific text than they would in more general writing.

Of the twenty Russian stems in the sample text whose English equivalents fell into more than one word class, four involved simple homographs. In each of these cases, two Russian words with identical stems had their English correspondents grouped together in the model glossary. The correspondents of each homograph fell into a single word class. The four homographic stems are listed at the top of Table 7. As for the remaining stems, given in the lower part of Table 7, the correspondents drawn from each listing in Smirnitskij fell into two or more word classes.

Table 7 shows the English correspondents and their word classes for each of the twenty stems, as well as the number of occurrences of each stem. It is difficult to see much pattern or regularity in the word class memberships. At the right of the Table similar information is given with certain of the word classes consolidated. Classes 3 and A are combined to form a general adjective grouping α, and classes 4 and D are combined into a general adverb grouping β. Classes 1 (nouns and pronouns) and 2/B (verbs) are left distinct, while the remaining classes are lumped together in γ. In terms of the new groupings, more regularity is evident. This is partly a reflection of the fact that Russian adverbs, like English, often modify either verbs or adjectives, so classes 4 and D are related. There is a similar close relation between adjectives and "determiners."

Eleven of the sixteen non-homographic stems have correspondents in grouping α and also in class 1, or grouping β, or both. Another stem has its correspondents in grouping α only. The remaining four stems involve grouping γ alone or with class 2/B.

The large number of stems which translate both as nouns and adjectives is traceable to the fact that Russian adjectives are often used as nouns, much as is done in English. The other word-class combinations are due either to vagaries of Russian usage or to peculiarities arising in translation. An example of the latter may be illustrative.

"Eshche" is a Russian adverb signifying continuity, as in "It is still raining." Here, the English equivalent is also an adverb. "Eshche" is also used in such a connection as "He gave me some more money." Here, though in Russian it modifies the verb, "eshche" must be translated into English as an adjectival phrase modifying "money." If there had been no object ("money") in the original Russian, the resulting translation "He gave me some more " would utilize "more" as a noun. Thus "eshche" may have an adverb, adjective, or noun correspondent in English. Here the languages differ in philosophy; does the "moreness" appertain to the action or to the thing given?

It appears that there is little to be gained from a more detailed study of the stems listed in Table 7. Each represents a highly individual multiple correspondence problem shedding little light on the general picture.

For the sake of completeness, the occurrences of mathematical symbols in the sample text were tabulated, as shown in Table 8. The symbols are of interest primarily because they sometimes enter into the sentence structure as subjects, predicates, etc. Symbols acting as sentence elements appeared most often as members of class 1: 49 times independently and 32 times in apposition with class 1 words. (The class 1 symbols were sometimes single symbols as listed in Table 8, sometimes groups of these such as " a + b," "x = y.") Symbols also appeared in sentences eight times as members of class 2, twice as members of class 3, and eight times as members of class A.

Table 8

SYMBOL OCCURRENCES

Type of Symbol	Examples	No. of Occurrences
Numbers	3, 21	7
Algebraic entities	x, A	107
Compound symbols	a^2, x_n, f(x)	53
Operations	+, ×	9
Relations	=, ∼	31

In general there seems to be some basis for doubt concerning the suitability of the word class scheme of Fries for the present application. Some rearrangements of the classes have been made for reasons of convenience during the course of this work. These rearrangements have resulted in a set of categories very similar to that of the conventional grammarian, whose example Fries strove to avoid. This suggests that Fries' scheme may not be appropriate for all types of linguistic analysis.

The data gathered in Tables 4-6 afford an opportunity for some tentative conclusions about the relevance of part-of-speech distinctions to the multiple meaning problem.

The Relative Multiplicities (Table 4) indicate that, word for word, the prepositions (class F) create more of a multiplicity problem than any other word class. Most of the trouble is caused by a very few words which have a large number of correspondents and which occur frequently. This certainly suggests that concentrated attention be devoted to these few words in an effort to reduce the confusion.

As has been pointed out above, prepositions seem to carry surprisingly little lexical meaning. In most Indo-European languages, prepositions are used in the expression of a large number of different concepts, and the combination of concepts embodied in a single preposition differs greatly from one language to another. Conversely, a single general concept is often expressed by a variety of prepositions, the appropriate choice of which must be considered idiomatic.

Can a machine reduce the number of alternatives through reference to the immediate context? Consider two uses of the preposition "v," translated in the machine glossary as "(in, at, into, to, for, on, N)." Reference to Smirnitskij reveals that when followed by the name of a place or object, "v" may be translated as "in," "into," "at," "to," or "for." In expressions of time it may appear as "in," "at," "on," or "N," as in the phrases "in three days," "at three o'clock," "on Thursday," or simply "Thursday." Evidently, knowledge of the preposition's object reduces the number of possible correspondents somewhat. Some rules can be invented for a further selection: reserve "into," "to," "for" for use with verbs of motion; use "at" with "o'clock;" and so forth. However, the method of context-reference which involves storing meaning class sequences of only three-word length is of little use in implementing these rules. The three-word context will not even include the object of a preposition if an adjective intervenes. On the whole, context-reference methods of the scale envisioned in this paper do not seem to hold much promise for reducing the multiplicity of prepositions.

A possible expedient might be to adopt some special convention for dealing with prepositions, e.g. transliterate directly the few extremely troublesome ones and then supply supplementary information concerning their usage along with each output text. However, such devices as this may add more difficulty than they remove.

The Multiplicity Indices of Table 4 show that class 1 words make the largest total contribution to the multiplicity problem. Class 1 supplies 36% of the total multiplicity, or 51% if the prepositions are omitted from the reckoning. The large contribution of class 1 words is due primarily to their frequent occurrence. Although a general study of class 1 words might prove rewarding, it would seem that the multiple correspondence problem is probably very similar for all meaning words.

The method of tabulating word meaning class sequences is useful primarily for the meaning words, Classes 1-4; it does not appear to be suitable for function words. This may not constitute a disadvantage of the method. Let the prepositions be disregarded for the present, inasmuch as they have been shown to present a very special sort of problem. Then it is evident from Table 4 that by far the largest proportion, at least 83%*, of the multiplicity trouble stems from Class 1-4 words. In view of this fact, it may be best not to try to assign the function words to meaning classes, but only to identify each with a special designation corresponding to its part-of-speech Group. This simplification would save much effort in making assignments to meaning classes, and would also reduce the number of distinct class sequences which must be stored in the translator.

---

* This figure is probably low. The only function word class other than class F (prepositions) having a significant Multiplicity Index is class E/J, with an Index of 89. Of this value, 44 is contributed by 22 occurrences of a Russian conjunction having the English equivalents "(and, N)." The null possibility occurs infrequently, and it is the present writer's feeling that "N" might well be omitted from the glossary.

Perhaps still better would be the complete omission of the function words from the class sequence scheme. Consider the English sentence: "Neither the positive nor the negative terminal was copper. " A context of three or even five words surrounding the word "positive" contributes no clue to its meaning. If the function words and also the verb "to be" are disregarded, the words "positive, negative, terminal, copper" are left. Some information about the proper choice of any word in this sequence could probably be gained by a knowledge of the meaning classes of its neighbors. If this technique were applied to a mechanical translation process, the number of correspondents for a given meaning word would be reduced by reference to the nearest other meaning words, with no attention being given to the intervening function words.

It is worth noting that Yngve, whose work concerning word class sequences was mentioned earlier, has come to a conclusion opposite to that proposed here. Yngve believes that "... a solution of grammatical and syntactical problems in translation.. . would also be a solution for considerably more than half of all the multiple-meaning problems, " and ".. .the multiple-meaning problem is less severe for the... [meaning] words."[4] By contrast, the evidence presented here seems to indicate that the multiplicity problem is best attacked by concentration on the meaning words, as long as some provision is made to handle a few troublesome prepositions.

From the ideas which have been discussed in this paper, a method of attack on the multiple-meaning problem can be formulated. First of all, the entries in the machine glossary must be made as short as seems advisable. Design of glossaries for special fields of knowledge will aid in this.

Next, let a scheme somewhat similar to Roget's be set up for classifying words on the basis of their meaning. Only the meaning words, comprising most of the nouns, verbs, adjectives, and adverbs, would be classified within this scheme. It seems doubtful that differentiation among these parts of speech would be advisable, since grammatical structure is otherwise ignored in the present method.

In a large sampling of Russian text, each meaning word would be classified according to the sense in which it is used at any particular occurrence. The class designations would be recorded in the order in which the corresponding words occur, with any intervening function words ignored. Then all of the distinct sequences of some convenient length would be sifted out. (Presumably, only sequences occurring within a single sentence would be used.) Three would seem to be an appropriate length for the sequences, although two is a definite possibility if storage space is limited. Use of longer sequences would multiply storage requirements tremendously.

The list of sequences would then be stored within the automatic translator. This list should be ordered, so as to reduce search time. With three-class sequences, ordering would be done on the second class of the three, so that an input meaning word whose translation was in doubt could be related to the meaning words preceding and following. If only two-class sequences could be stored, it would definitely be worth while to store the complete list twice, ordered on both first and last class. Then, to obtain information on a certain input word, separate comparisons with the list could be made using the preceding word and following word.

The programming of the context-comparison process within the translator is by no means straightforward. If several consecutive input meaning words each have a number of correspondents, the choice of alternatives for one word will depend upon the choices made for the others. For a simple example, suppose that two consecutive Russian words A and B have the multiple English correspondents a_1, a_2 and b_1, b_2 respectively. Consideration of A, taking into account the preceding word as well as B, shows that a_1 could occur if followed by b_1. and that a_2 could occur if followed by b_2. a_1 and a_2 are therefore left in the translator output as possible alternatives. Consideration of B, taking into account the word following, then shows that b_2 cannot occur. Is the machine to turn back and reexamine A? In a sentence containing many multiple correspondences, a reexamination process could become extremely complicated.

Furthermore, it is not certain that the meaning-class sequence method outlined here is basically sound. The amount of text to be analyzed as the source of the list of permissible sequences obviously must be extremely large if it is to provide all of the sequences possible in the Russian language. Such a list may be an impossibility, since there is no way

Continued on page 43

Gould from page 27

of being sure that a given machine input text does not contain sequences which have never been used elsewhere. The probability of this situation can be minimized by making the number of meaning categories small; but this also limits the usefulness of the method.

The proposals discussed here do nothing to improve the structure of the translating machine output as regards grammar, word order, etc. This appears to be a somewhat separate problem, and a complex one. On the basis of Oettinger's results discussed at the beginning of this paper, the multiple-meaning problem would seem to take precedence.

The writer is grateful to Prof. Anthony G. Oettinger for his valuable advice on the preparation of this paper.