MECHANICAL TRANSLATION

DEVOTED TO THE TRANSLATION OF LANGUAGES *WITH THE AID OF* MACHINES

VOLUME FOUR, NUMBERS ONE/TWO NOVEMBER, NINETEEN FIFTY SEVEN

News

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Problems of German syntax in the context of MT will be the area of investigation and discussion for a group of persons who will meet and work together during July and August, 1958, at the Massachusetts Institute of Technology. The principal agendum will be exploration of German syntax in detail; the objective, to work toward that kind of full analysis of German syntactic structures which is prerequisite to MT with German as either source or target.
All persons who may be interested in attending the meeting should write to V.H. Yngve, 20B-101D, M.I.T., Cambridge, Massachusetts.

HARVARD UNIVERSITY

The compilation of an automatic Russian-English dictionary suitable for operation on a Univac computer is continuing at the Harvard Computation Laboratory. This dictionary will contain words of general currency plus specialized terms for electronics. Ordinary dictionaries are being used as the initial source of words. Up-dating procedures are planned to be semi-automatic: unused words will be periodically purged on the basis of zero look-up frequency, and new words occurring in texts being processed will be added. Paradigms are represented by stems, of which the number ranges from one per paradigm for most nouns to five or six per paradigm for most verbs. Dictionary entries are being provided with coded symbols for eventual use in syntactic analysis.

A. G. Oettinger

A GLOSSARY OF RUSSIAN PHYSICS ON PUNCHED CARDS

A glossary of 6, 000 Russian forms has been prepared in a cooperative effort of the University of Michigan and The RAND Corporation.

The entire text of Volume 28, Number 1 (January, 1955), of the Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki was punched onto IBM cards; this text consists of about 30, 000 running words, punched one word to a card. These cards were alphabetized, and tentative English equivalents were assigned.
The tentative English equivalents were then listed in textual order. Editors compared this list with the text of Soviet Physics (a translation of the Zhurnal) and selected a preferred equivalent in each context. These editors also assigned a code to each item in the glossary to indicate the syntactic function (or functions) of the form as inflected, and grouped inflected forms into words.
The final glossary consists of about 6, 000 inflected forms of about 2, 300 words. Each form card contains: the complete Russian form, a form identification number, the identification number of the word of which this is a form, a grammar code, and one or more English equivalents .
A brief list of idioms has also been prepared; references are made to the idiom list, as appropriate, in the glossary. The frequency of preference of each English equivalent of each Russian word is also available.
Duplicate sets of glossary cards, and listings of them, are available at cost to research workers.

K.E. Harper
D.G.Hays
A.Koutsoudas

A Model for Mechanical Translation

John P. Cleave, Birkbeck College Research Laboratory, University of London*

A mathematical model for a translating machine is proposed in which the translation of each word is conditioned by the preceding text. The machine contains a number of dictionaries where each dictionary represents one of the states of a multistate machine.

IN MECHANICAL TRANSLATION the foreign language (input text) words are operated on by a computer, which is programmed to effect certain formal rules to produce a series of target language (output text) words. Mechanical translation may therefore be regarded as a transformation of a series of data S_1 to a series S_2.

Suppose the series S_1 is composed of elements of the finite set $(a_1, a_2 \ldots a_n)$ and S_2 is composed of elements of $(b_1 \ldots b_m)$. These elements $a_1, a_2 \ldots a_n, b_1 \ldots b_m$ correspond to the words of the input text and output text respectively. Let $S_1(n)$ denote the n-th datum of the series S_1. Then the simplest type of transformation by which the output series S_2 is printed is expressed by the rules,

"rule r: If $S_1(n) = a_{\mu_r}$ print b_{ν_r}, add 1 to n and go to rule 1.

If $S_1(n) \neq a_{\mu_r}$ go to rule r + 1,"

where r = 1 .... n and where the set (a_{μ_r}) is identical with $(a_1, a_2 \ldots a_n)$. The transformation corresponds to a word-for-word translation and also to a simple coding expressed by the table

S_1 element.	S_2 element.
a_{μ_1}	b_{ν_1}
a_{μ_2}	b_{ν_2}
⋮	⋮
a_{μ_n}	b_{ν_n}

which may be regarded as a dictionary. If the input data S and the output data are punched tape on an automatic computer with unidirectional reading and printing devices, then the above transformation is effected by a single-state machine.

A word-for-word translation in which the equivalents selected for an input word depend upon the context of the preceding text is represented by a compound coding, effected by a multistate machine. This type of transformation, called "conditional" is effected by the rules:

"rule r: If $S_1(n) = a_{\mu_r}$ and if $S_1(n)$ is preceded in the S_1 series by elements $a_{\mu_{r,q}}, a_{\mu_{r,q-1}} \ldots a_{\mu_{r,1}}$ in that order (not necessarily juxtaposed) then print b_{ν_r}, add 1 to n and go to rule 1.

If $S_1(n) \neq a_{\mu_r}$ or if $S_1(n) = a_{\mu_r}$

---

* Now at Southampton University, Southampton, England.

and either

(i) $S_1(n)$ is not preceded by elements

$a_{\mu_{r,q}}$, $a_{\mu_{r,q-1}}$ ..... $a_{\mu_{r,1}}$, or

(ii) if $S_1(n)$ is preceded by these elements,

and they are not in the required order in S_1,

then go to rule r + 1,"

where r = 1, 2..... We suppose that the sequence of rules provides a course of action for each possibility. (The exact conditions on the number of rules will not be investigated here, but it should be noted that the rules are in a certain order.) If we let the sign ' »' denote 'precede in the message' then rule r can be abbreviated to

"rule r:

$$a_{\mu_{r,q}} \gg a_{\mu_{r,q-1}} \gg \ldots \gg a_{\mu_r} \longrightarrow b_{\nu_r}"$$

Thus the ordered list of rules is:

$$a_{\mu_{1,q}} \gg a_{\mu_{1,q-1}} \gg \ldots\ldots a_{\mu_{1,1}} \gg a_{\mu_1} \longrightarrow b_{\nu_1}$$

$$a_{\mu_{2,s}} \gg a_{\mu_{2,s-1}} \gg \ldots\ldots a_{\mu_{2,1}} \gg a_{\mu_2} \longrightarrow b_{\nu_2}$$

$$\vdots$$

$$a_{\mu_{p,t}} \gg a_{\mu_{p,t-1}} \gg \ldots\ldots a_{\mu_{p,1}} \gg a_{\mu_p} \longrightarrow b_{\nu_p}$$

$$a_{\mu_{p+1}} \longrightarrow b_{\nu_{p+1}}$$

$$a_{\mu_{p+2}} \longrightarrow b_{\nu_{p+2}}$$

$$\vdots$$

$$a_{\mu_{p+n}} \longrightarrow b_{\nu_{p+n}}$$

The last n rules cover those instances where a datum of S_1 is not preceded by its relevant context. These rules cannot be reduced to the simple dictionary with a finite number of entries as in the previous simple transformation.

Instead a connected series of dictionaries may be constructed by the following method, which is best illustrated by supposing one conditional rule only. Suppose the sequence of rules is

$$a_p \gg a_{p-1} \gg \ldots \gg a_2 \gg a_1 \longrightarrow b_q$$

$$a_1 \longrightarrow b_{\mu_1}$$

$$a_2 \longrightarrow b_{\mu_2}$$

$$\vdots$$

$$a_n \longrightarrow b_{\mu_n}$$

The sequence of dictionaries will contain some entries which will refer the operator to another dictionary. If we let, say

$$"a_s \longrightarrow b_{\mu_s}(t)"$$

denote an entry in dictionary u which prints b_{μ_s} when a_s occurs in S_1 and then changes the dictionary from u to t, and let

$$"a_s \longrightarrow b_{\mu_s}"$$

denote an entry which does not affect a change of dictionary, then the list of rules above may be replaced by the dictionaries

Dictionary (1)	Dictionary (2)	Dictionary (3)
$a_1 \longrightarrow b_{\mu_1}$	$a_1 \longrightarrow b_{\mu_1}$	$a_1 \longrightarrow b_{\mu_1}$
$a_2 \longrightarrow b_{\mu_2}$	$a_2 \longrightarrow b_{\mu_2}$	$a_2 \longrightarrow b_{\mu_2}$
$\vdots$	$\vdots$	$\vdots$
$a_p \longrightarrow b_{\mu_p}$ (2)	$a_{p-1} \longrightarrow b_{\mu_{p-1}}$ (3)	$a_{p-2} \longrightarrow b_{\mu_{p-2}}$ (4)
$a_{p+1} \longrightarrow b_{\mu_{p+1}}$	.	$a_{p-1} \longrightarrow b_{\mu_{p-1}}$ (1)
.	.	$a_p \longrightarrow b_{\mu_p}$ (1)
.	.	$a_{p+1} \longrightarrow b_{\mu_{p+1}}$
.	.	.
.	.	.
.	.	.
$a_n \longrightarrow b_n$	$a_n \longrightarrow b_n$	$a_n \longrightarrow b_n$

Finally

Dictionary
(P)

$a_1 \longrightarrow b_q$ (1)

$a_2 \longrightarrow b_{\mu_2}$ (1)

$a_3 \longrightarrow b_{\mu_3}$ (1)

$\vdots$

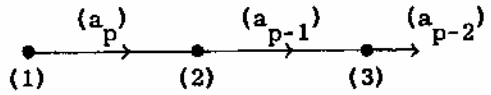
$a_p \longrightarrow b_{\mu_p}$ (1)

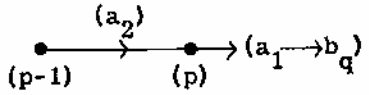
$a_{p+1} \longrightarrow b_{\mu_{p+1}}$

$\vdots$

$a_n \longrightarrow b_n$

With obvious convention the connection of the dictionaries may be represented by





For two conditional rules

$a_p \gg a_{p-1} \gg \ldots\ldots \gg a_1 \longrightarrow b_q$

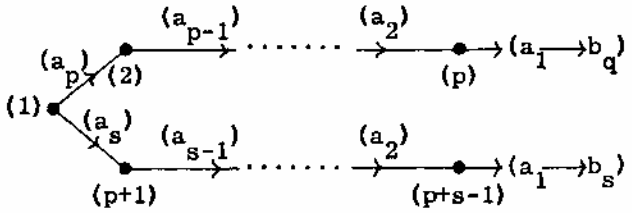
$a_s \gg a_{s-1} \gg \ldots\ldots \gg a_1 \longrightarrow b_t$

$a_1 \longrightarrow b_{\mu_1}$

$\vdots$

$a_n \longrightarrow b_{\mu_n}$

The connection of dictionaries is represented by



If the conditional rules are effected by a computing machine, each dictionary represents a state of the machine. A transformation which depends upon context therefore can be represented as a compound coding or a multistate machine.

Structural Grammars†

R. B. Lees, Research Laboratory of Electronics,
Massachusetts Institute of Technology, Cambridge, Massachusetts

We adopt the view that the grammar of a language is a predictive theory which isolates the grammatical sentences of that language by means of immediate constituent analyses, morphophonemic conversions, and grammatical transformations. A sample grammatical analysis is given for the development of the verb phrase in German independent clauses. Simple rules are given for converting the verb phrase as a sequence of personal affixes, various auxiliaries, and the main verb into passive, future, or conditional clauses, and then introducing word boundaries, choosing the proper auxiliaries, arranging the word-order, and finally mapping the resulting morpheme sequence into the correct sequence of words in the independent clause.

ANY reasonably general, mechanized program for translating texts from one language into another can avoid dealing with each and every sentence as a completely new and arbitrary sequence of dictionary items only if it provides each source-language sentence with a grammatical analysis.

Traditional notional or semantic-based grammatical descriptions are useless for this purpose, since an analysis using such a grammar can be carried out only if the meanings of all of the constituents of the sentence are given. These meanings cannot be assumed: one of the main purposes of a syntax program is to aid in determining them so that they can be transferred, i.e., translated, into the appropriate target-language equivalents. Furthermore, contemporary descriptive linguistic grammatical practice is also faulty, especially when it is to be employed in a machine program; for, while the descriptive linguist no longer designates sentence constituents by means of meaning-labels but refers exclusively to their perceptible shapes, the description is still largely ad hoc — each particular grammatical category is designated by an arbitrary stigma or mark of class membership and must be devised differently for each language. Moreover, descriptive sketches are deficient in their presentation of the syntax in that they are either fragmentary or else require very complicated, arbitrary, and often-repeated rules for specifying the constituent structure of even fairly Simple sentences. This is largely the result of assuming that all sentences of a natural language are describable in terms of an immediate-constituent analysis or branching-diagrams.

N. Chomsky [1] has described a theory of language which avoids these difficulties by relaxation of requirements on a grammar to the weaker position of satisfying some evaluation procedures (instead of requiring a discovery or decision procedure), introduction of natural canons of simplicity or elegance, statement in terms of a set of expansion rules for generating all grammatical utterances, and, above all, introduction of a level of grammatical transformations. These grammatical transformations convert the constituent-structures of a set of the most central sentences (i.e., basic, nonderived sentence types, such as affirmative assertions) into the derived structures of a more complex, less central, and infinitely extendible set of sentences.

† This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research); and in part by the National Science Foundation.

1. Chomsky, N., "The Logical Structure of Linguistic Theory", Preliminary Draft, M.I.T., 1956, 713 + li pp.

Following certain suggestions of Chomsky and Lukoff [2] a scheme has been constructed as an illustration of a small, isolated portion of such a grammar for German. The scheme is intended to generate all verb phrases of independent clauses, active and passive, subject to the following limitations:

a) The device generates several types of verb phrase which would occur only rarely in natural speech, not for any clearly grammatical reason, but simply because they are too long or clumsy. Three types generated are probably only semigrammatical, containing two past participles in ge-. In addition, several very long, but not obviously excluded, types will not be generated.

b) There is no provision for conforming the affixes of the finite verb to those of the accompanying noun phrases in the sentence, or for adjusting the selection between particular verb-phrase morphemes and other morphemes external to the verb phrase, such as subject, object, or adverbial, or between the verb and the separable prefix. (This last selection would devolve upon the lexicon.)

c) No provisions are made to generate impersonal constructions, zu- infinitives, nominalized verb phrases, dependent clauses, or other verbal constructions.

d) The rules for generating the proper allomorphic shapes of the stems and affixes are only suggested by reference to a few examples, since a complete listing of morpheme spellings would be as long as the lexicon.

2. Chomsky, N. and Lukoff, F., "Construction of the German Verb Phrase", Mechanical Translation Group Memo, Aug. 12, 1955, Research Laboratory of Electronics, M.I.T.

GLOSSARY OF SYMBOLS

Af	Any affix or connected sequence of affixes of the set Ps, sbj, pst, I, G
Af_1	Affix of subject nominal
Af_2	Affix of object nominal
Aux	Auxiliary verb stem, hab or sei
D	Any post-verbal objects, adverbials, predicate nominals or adjectivals
G	Affix of past participle
I	Affix of infinitive
M	Modal stem, könn, müss, etc.
M^{ob}_{Af}	Obligatory mapping which places the affixes after the appropriate stem
M^{ob}_{Aux}	Obligatory mapping which selects the proper auxiliary stem
M^{ob}_{DI}	Obligatory mapping which replaces a participle with an infinitive in the so-called "Double Infinitive" construction
M^{ob}_{ge}	Obligatory mapping which introduces the special participle of werden after another participle
M^{ob}_{ord}	Obligatory mapping which rearranges the word-order, placing non-finite verb forms at end in correct order
M^{ob}_{W}	Obligatory mapping which introduces word boundaries at proper places
N_1	Subject nominal
N_2	Object nominal
Ps	Any personal affix
pst	Past tense affix
Q	Any St + Af or sequence of these
S	Sentence
sbj	Subjunctive morpheme
sep	Any separable prefix
St	Any stem of the set V, M, W, or Aux
t_2	2nd person plural personal affix
t_3	3rd person singular personal affix
T^{op}_{P}	Optional transformation of kernel sentences producing passive sentences

T_W^{op}	Optional transformation of kernel sentences or passive sentences producing future and conditional sentences	Y	Any string
		Z	Any St + Af
V	Any verb stem = either V_h or V_s	+	Symbol of grammatical concatenation
V_h	Any verb stem which takes haben as auxiliary	/	Word boundary
		*	Sentence boundary
V_s	Any verb stem which takes sein as auxiliary	=	The grammatical rule "Rewrite the foregoing as:"
W	Stem of the verb werden	( )	Optionally present
X	Any St(+Af)	{ }	Alternatively present

DEVELOPMENT OF THE VERB PHRASE

1. PHRASE-STRUCTURE RULE to yield verb phrases of kernel sentences

$$S = N_1 + Af_1\,(+pst)(+sbj) + Ps\,(+ \left\{\begin{matrix}(M+I) + M + I\,(+Aux+G)\\(Aux+G)(+M+I)(+M+I)\\M+I+Aux+G+M+I\end{matrix}\right\}) + V(+D)(+sep+)*$$

(Abbrev.:

N_1	Subject nominal	N_2	Object nominal
Af_1	Subject nominal affix	Af_2	Object nominal affix
pst	Past tense morpheme	sep	Separable verb prefix
sbj	Subjunctive morpheme	S	Sentence
Ps	Personal affix	St	V, M, W, Aux, hab, or sei
M	Modal stem	Af	Ps, sbj, pst, I, G, or any connected sequence of these
I	Infinitive affix		
Aux	Auxiliary verb stem	D	Objects, adverbials, predicate nominals, adjectives, etc.)
G	Past participle affix		
V	Verb stem		

2. Optional GRAMMATICAL TRANSFORMATIONS to yield non-kernel sentences

a. Passive transformation:

$$T_P^{op}: \quad N_1 + Af_1\,(+Af) + Ps\,(+X) + V + N_2 + Af_2\,(+D)(+sep+)$$
$$= N_2 + Af_2(+Af) + Ps(+X) + W + G + V(+D)(+von+N_1+Af_1)(+sep+)$$

b. Werden transformation to yield future and conditional phrases:

$$T_W^{op}: \quad \left\{\begin{matrix}pst + sbj\\(+sbj)+Ps\end{matrix}\right\} + St = \left\{\begin{matrix}pst + sbj\\(+sbj)+Ps\end{matrix}\right\} + W + I + St$$

3. Obligatory MAPPINGS to yield proper word-order, word boundaries, and auxiliary selections

a. Word boundary:

$$M^{ob}_{W}: \quad \left\{ \begin{matrix} St \\ D \\ von \\ Af_1 \\ Af_2 \end{matrix} \right\} + = \left\{ \begin{matrix} St \\ D \\ von \\ Af_1 \\ Af_2 \end{matrix} \right\} \quad /$$

b. Affixation:

$$M^{ob}_{Af}: \quad Af + St = St + Af$$

c. Auxiliary selection:

$$M^{ob}_{Aux}: \begin{cases} Aux + Af \; / \left\{ \begin{matrix} V_h \\ M \end{matrix} \right\} = hab + Af \; / \left\{ \begin{matrix} V_h \\ M \end{matrix} \right\} \\ Aux + Af \; / \left\{ \begin{matrix} V_s \\ W \end{matrix} \right\} = sei + Af \; / \left\{ \begin{matrix} V_s \\ W \end{matrix} \right\} \end{cases}$$

d. Word order:

$$M^{ob}_{ord}: \begin{cases} X \; / \; (Y/) \; Z \; / \; V + Af \; / \; (Q/) \; * = X \; / \; (Y/) \; V + Af \; / \; (Q/) \; Z \; * \\ X \; / \; (Y/) \; Z \; / \left\{ \begin{matrix} D \; / \; (sep+) \\ sep+ \end{matrix} \right\} (Q/) \; * = X \; / \; (Y/) \left\{ \begin{matrix} D \; / \; (sep+) \\ sep+ \end{matrix} \right\} (Q/) \; Z \; * \end{cases}$$

(where X = St(+Af) + Ps; Y = any string; Z = St + Af;
Q = St + Af or any sequence of these)

e. Double Infinitive:

$$M^{ob}_{DI}: \quad M + G \; / \; * = M + I \; / \; *$$

f. Special participle:

$$M^{ob}_{ge}: \quad V + G \; / \; W + G = V + G \; / \; worden$$

MORPHOPHONEMIC RULES

a. Personal endings:

$$Ps = \left\{ \begin{matrix} e \\ st \\ t_3 \\ en \\ t_2 \end{matrix} \right\}$$

$$St + X + \left\{ \begin{matrix} e \\ t_3 \end{matrix} \right\} = St + X$$

pst = te

sbj = e

b. Stems:

M = könn, müss, woll, soll, dürf, mög

W = werd

V_h = sing, hör, mach, trag, geb, ...

V_s = geh, fahr, bleib, werd, ...

c. Separable prefixes:

sep + = an, auf, ein, hin, vor, ...

d. Infinitives and participles:

I = en (special rules for verbs like tadeln where I = n)

sing + G = gesungen
hör + G = gehört
studier + G = studiert
etc.

e. Past and subjunctive stems:

sing + te = sang
sang + e = sänge

f. General morphophonemic rules:

e + e = e

g. Finite verb affixes:

sing+e	= singe	fahr+e	= fahre	hör+e	= höre
sing+st	= singst	fahr+st	= fährst	hör+st	= hörst
sing+t_3	= singt	fahr+t_3	= fährt	hör+t_3	= hört
sing+en	= singen	fahr+en	= fahren	hör+en	= hören
sing+t_2	= singt	fahr+t_2	= fahrt	hör+t_2	= hört
sing+e	= singe	fahr+e	= fahre	hör+e	= höre
sing+e+st	= singest	fahr+e+st	= fahrest	hör+e+st	= hörest
sing+e	= singe	fahr+e	= fahre	hör+e	= höre
sing+en	= singen	fahr+en	= fahren	hör+en	= hören
sing+e+t_2	= singet	fahr+e+t_2	= fahret	hör+e+t_2	= höret
sang	= sang	fuhr	= fuhr	hör+te	= hörte
sang+st	= sangst	fuhr+st	= fuhrst	hör+te+st	= hörtest
sang	= sang	fuhr	= fuhr	hör+te	= hörte
sang+en	= sangen	fuhr+en	= fuhren	hör+ten	= hörten
sang+t_2	= sangt	fuhr+t_2	= fuhrt	hör+te+t_2	= hörtet
sänge	= sänge	führe	= führe		
sänge+st	= sängest	führe+st	= führest		
sänge	= sänge	führe	= führe		
sängen	= sängen	führen	= führen		
sänge+t_2	= sänget	führe+t_2	= führet		

A SAMPLE DERIVATION

VP

(1.) $N_1 + Af_1 + pst + sbj + Ps + Aux + G + V + N_2 + Af_2 + D + sep + *$

(2.a.) $N_2 + Af_2 + pst + sbj + Ps + Aux + G + W + G + V + D + von + N_1 + Af_1 + sep + *$

(2.b.) $sbj + Ps + W + I + Aux + G + W + G + V + D + von + N_1 + Af_1 + sep + *$

(3.a.) $N_2 + Af_2 \,/\, pst + sbj + Ps + W \,/\, I + Aux \,/\, G + W \,/\, G + V \,/\, D \,/\, von \,/\, N_1 + Af_1 \,/\, sep + *$

(3.b.) $W + pst + sbj + Ps \,/\, Aux + I \,/\, W + G \,/\, V + G \,/\, D$

(3.c.) $N_2 + Af_2 \,/\, \underbrace{W + pst + sbj + Ps}_{X} \,/\, \underbrace{sei + I \,/\, W + G}_{Y} \,/\, \underbrace{V + G}_{Z} \,/\, \underbrace{D \,/\, von \,/\, N_1 + Af_1}_{D} \,/\, sep + *$
$/\, sep + *$

(3.d.) $\underbrace{W + pst + sbj + Ps}_{X} \,/\, \underbrace{sei + I}_{Y} \,/\, \underbrace{W + G}_{Z} \,/\, \underbrace{D \,/\, von \,/\, N_1 + Af_1}_{D} \,/\, sep + \underbrace{V + G}_{Q} \,/\, *$
$/\, sep + Q \,/\, *$

(3.d.) $\underbrace{W + pst + sbj + Ps}_{X} \,/\, \underbrace{sei + I}_{Z} \,/\, \underbrace{D \,/\, von \,/\, N_1 + Af_1}_{D} \,/\, sep + \underbrace{V + G \,/\, W + G}_{Q} \,/\, *$
$/\, sep + Q \,/\, *$

(3.d.) $\underbrace{W + pst + sbj + Ps}_{X} \,/\, \underbrace{D \,/\, von \,/\, N_1 + Af_1}_{D} \,/\, sep + \underbrace{V + G \,/\, W + G \,/\, sei + I}_{Q} \,/\, *$

(3.f.) $/\, worden \,/$

(4.a.) $W + pst + sbj + t_3$

(4.a.) $W + pst + sbj$

(4.a.) $W + te + e$

(4.b.) $/\, werd + te + e \,/\, D \,/\, von \,/\, N_1 + Af_1 \,/\, sep + trag + G \,/$

(4.c.) $/\, ein + trag + G \,/$

(4.d.) $/\, ein + getragen \,/\, worden \,/\, sein \,/\, *$

(4.e.) $/\, würde + e \,/$

(4.f.) $N_2 + Af_2 \,/\, würde \,/\, D \,/\, von \,/\, N_1 + Af_1 \,/\, eingetragen \,/\, worden \,/\, sein \,/\, *$

Let $N_2 + Af_2$ = Das / Geld

D = früher / auf / meinem / Konto

$N_1 + Af_1$ = der / Gesellschaft

Das / Geld / würde / früher / auf / meinem / Konto / von / der / Gesellschaft / eingetragen / worden / sein / *

Semantic Frequency Counts

Paul Pimsleur, University of California, Los Angeles, California

The success of a mechanical translation should be measured in terms of the level of depth required by the situation. To determine whether a careful translation is desirable a rough scanning will suffice. The use of cover-words, high frequency words that may be substituted for low frequency words, in the output language is an essential part of this process. The preparation of trans-semantic frequency counts resulting in dictionaries of reduced size that require less computer storage capacity is recommended.

ACCORDING to Y. Bar-Hillel, "The central problem in mechanizing translation is the preparation of methods that permit a more restricted memory. Hitherto accepted methods require a rapid access mechanical memory with storage capacity greatly in excess of that of available electronic computers."[1]

Though work is now in progress on machines featuring large density storage units and rapid access time, [2] the development of such machines will not substantially change the problem. The goal is, and will remain, the creation of the most efficient dictionary for MT purposes, containing the smallest number of entries and featuring the most rapid search procedures.

The reduction of dictionary size is directly related to the matter of multiple -meaning. The ideal dictionary will be the smallest possible one which still suffices to meet the requirements of translation, within the limits of accuracy we have chosen to accept. However, such a dictionary presupposes considerable knowledge of the frequency with which words occur, in each of their several meanings. "In effect, what is needed are true ideoglossaries, based on actual, rather than potential, behavior."[3] Though some attempts have been made to attack this problem as it has arisen in particular research contexts, [4] no concentrated effort is being exerted toward the establishment of semantic frequency counts per se. It appears, however, that such counts are essential to the future development of MT. Some additional incentive may also be derived from the recent indications that Russian MT specialists have been working for some time on a "polysemantic dictionary" which is a central part of their MT procedure.[5]

A semantic frequency count is a listing of the words of a language, with the several meanings of each word, and the relative frequency of occurrence of each meaning in general and/or specialized contexts. Valuable as such a count might be to scholars and educators in various domains, it appears that a somewhat different count is needed for purposes of MT. The need is for TRANS-SEMANTIC FREQUENCY COUNTS. A trans-semantic frequency count is a listing of the words of the source language, together with the various possible renderings of each in the target language, and the frequency of occurrence of each of the latter. Such a listing would resemble a normal translation dictionary, with the addition of information, probably in the form of percentages, giving the

---

1. Y. Bar-Hillel, "Can Translation be Mechanized, " (abstract) MT, Vol.3, No. 2, p. 67.

2. G.W. King, "Stochastic Methods of Mechanical Translation, " MT, Vol. 3, No. 2, pp. 38-39.

3. K.E. Harper, "Contextual Analysis in Word-for-Word MT, " MT, Vol.3, No. 2, p. 40.

4. A. Koutsoudas and R. Korfhage, "Mechanical Translation and the Problem of Multiple Meaning," MT. Vol.3, No. 2, pp. 46-51, 61.

5. D. Panov, "On the Problem of Mechanical Translation, " MT, Vol.3, No. 2, pp. 42-43.

frequency of occurrence of each meaning in the target language. Alternate frequencies should also be given for various subject areas, scientific, military, etc.

As described here, such an undertaking would be enormous, even for any two languages. However, it may be argued that: 1) the need for such information is great for MT; 2) any partial listing would provide data that could immediately be useful in the preparation of MT dictionaries.

In connection with the problem of multiple-meaning, it may be useful to dwell briefly on another approach. Virtually all non-mechanical translators, and even some who are concerned with MT, think in terms of sure translation. By sure translation is meant a sort of one-to-one semantic mapping from the words of the source language to the best possible "mots justes " of the target language. The suggestion is offered that the issue be rephrased in terms of probabilities (a "stochastic approach"[6]), in which we aim at the degree of success in translation which the situation seems to demand. By success is meant a comprehensible, non-misleading rendering. The degree of success may well vary with the danger or inconvenience resulting from imperfect translation. In many instances, there may be quantities of material to be merely scanned for purposes of determining whether any use is to be made of any part of it. In such cases, a very rough translation has been shown to suffice,[7] with a consequent saving in cost and intricacy of machine operation. A minimum probability coefficient of .80 for each ambiguous word may be sufficient for such rough scanning. This sort of translation is probably attainable in the relatively near future, though anything like a "perfect" translation is still on the distant horizon.

Thus the concept of levels of depth becomes important. The first level of depth may be a translation in which the chances are 80 or more out of a hundred that each ambiguous word has been translated acceptably. The second level of depth might involve a minimum confidence of 90% per word; the third and most refined level (the one on the distant horizon) would provide confidence .95 or perhaps even .99 per multiple-meaning word. This concept may be symbolized as:

$$\Pr(X \text{ is acceptable}) \geq 1-\alpha$$

where Pr means "the probability that. .. ", X represents a given rendering of a source word in the target language, and a stands for the maximum tolerable error per word. In the levels of depth just discussed, the alphas would be .20, .10, and .05 or .01, respectively. Obviously, each successive level will require considerably more search-time, an improved and probably a larger dictionary, and more detailed programming.

An illustration may serve to clarify several concepts. In the German sentence

Die Aufgabe ist zu schwer.[8]

the word schwer presents a typical problem in multiple-meaning. A dictionary of modest dimensions [9] lists the following eight meanings, for each of which we have provided an English translation. (Several sub-meanings listed as colloquial have, perhaps unfairly, been omitted.)

1) 'weigh-s' (verb). Die Kiste ist drei Zentner schwer, 'the box weighs three hundredweight .'
2) 'heavy'; 'strong.' ein schwerer Stein, 'a heavy stone;' ein schwerer Wein, 'a strong (intoxicating) wine.'
3) 'laden.' Das Dach ist schwer von Schnee. 'the roof is laden with snow.'
4) 'difficult.' Das fällt mir schwer, 'I find 'that difficult.'
5) 'unfortunate'; 'hard.' Er hat ein schweres Schicksal, 'he has an unfortunate fate.' Sie nimmt es schwer, 'she takes it (the news) hard.'
6) 'very.' Der Mann ist schwer reich, 'the man is very rich.'
7) 'slow-ly.' Er ist schwer von Begriff, 'he is slow to catch on,' or 'he catches on slowly.'
8) 'pregnant.' Die Lage ist schwer an Entscheidungen, 'the situation is pregnant with decisions.'

---

6. G. W. King, "Stochastic Methods of Mechanical Translation," MT. Vol.3, No. 2, pp. 38-39.

7. J.W. Perry, "Translation of Russian Technical Literature by Machine, " MT. Vol. 2, No. 1, (discussion of results) p. 16.

8. T.M. Stout, "Computing Machines for Language Translation, " MT, Vol. 1, No. 3, p. 41.

9. Der Sprach-Brockhaus. Eberhard Brockhaus, Wiesbaden, 1954.

There are thus ten possible translations for the German word schwer, in this no doubt incomplete list. They are: 'heavy, strong, laden, difficult, unfortunate, hard, pregnant, slow-ly, very, weigh-s.' By introducing the concept of COVER-WORDS, the number of these translations can be substantially reduced.

A cover-word is a word of relatively high semantic frequency which can be used in place of words of lower semantic frequency, with little possibility of misinforming the reader.

Referring back to the list above, let us examine each of the meanings of schwer in turn.

1) 'weigh-s' (v.i.) requires the translation of a predicate adjective in German by a verb in English — though these grammatical concepts may be operationally meaningless in MT, they are retained here for convenience. The importance of the problem depends on the frequency of occurrence of this locution, which is unknown at present. A trans-semantic frequency count would help us to decide how situations of this sort are to be handled. In any event, the possibility should be considered of using the awkward translation, 'the box is three hundredweight heavy,' thereby using the cover-word 'heavy' for 'weighs.' The loss is primarily of elegance, not of correct understanding.

2) 'heavy' needs no comment; it is a primary, or high-frequency rendering. 'Strong' would seem to be infrequent enough to render it inconsequential, but this again must be confirmed empirically.

3) 'laden.' If we rendered 'the roof is laden with snow' by 'the roof is heavy with snow,' the cover-word is used and no misinterpretation can result.

4) 'difficult' is a high-frequency meaning and appears irreduceable. This again must be checked empirically, which presupposes a trans-semantic frequency count.

5) 'unfortunate' may be replaced by 'heavy' in the sentence 'he has a heavy fate,' with a loss of elegance but little semantic distortion. The meaning 'hard,' as in 'she takes it hard' is somewhat more troublesome. Whether it is worthwhile to program special instructions for dealing with this case will depend on the frequency with which it can be expected to occur. In scientific literature at least, the frequency may be negligible. Should special provision for this case be necessary, it might be best to treat it as a compound, etwas schwernehmen.

6) 'very.' Schwer reich should be translated as 'very rich,' while schwer verletzt means 'badly wounded,' and schwer enttäuscht may be either 'badly disappointed' or 'very disappointed. ' The solution seems to lie in translating schwer in this context as 'very,' thus forcing acceptance of 'he was very wounded' instead of 'he was badly wounded.' It appears necessary to allow 'very' as a third rendering of schwer, alongside 'heavy' and 'difficult.' However, its occurrence as 'very' may be limited to cases such as those cited above, where it is directly followed by one of a small number of adjectives and can thus be identified rather easily by the machine.

7) 'slow-ly.' Schwer von Begriff requires special treatment as an idiom.

8) 'pregnant' can be rendered by the cover-word 'heavy' without serious loss.

Thus the ten meanings of schwer have been reduced to three cover meanings, 'heavy, difficult and very,' of which only 'difficult' and 'heavy' may be expected to occur in many different settings which we cannot at present predict. No loss of comprehension has resulted from the use of cover-words, though stylistic violence has been done to a varying extent. This drawback is offset by a substantial gain in terms of machine time and storage space.

SUMMARY AND CONCLUSIONS

1. It has been suggested that work be undertaken with all possible speed toward the establishment of trans-semantic word counts, with the goal of attaching a probability coefficient to the occurrence of a given meaning of a given word in a given subject field. Without underestimating the enormousness of the task, it is submitted that it is indispensable to MT. The work should commence with the subject areas of most immediate concern, i.e. scientific, and with the words which occur with greatest frequency, as shown by existing word-counts of the major languages. New machine methods may lighten the task considerably.

2. The concept of levels of depth has been used to describe translations of differing (but predictable) degrees of accuracy.

3. The concept of cover-words has been used, as well as that of trans-semantic frequency counts, to assist in reducing the contents of a storage dictionary.

Multiple Correspondence†

Roderick Could, 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

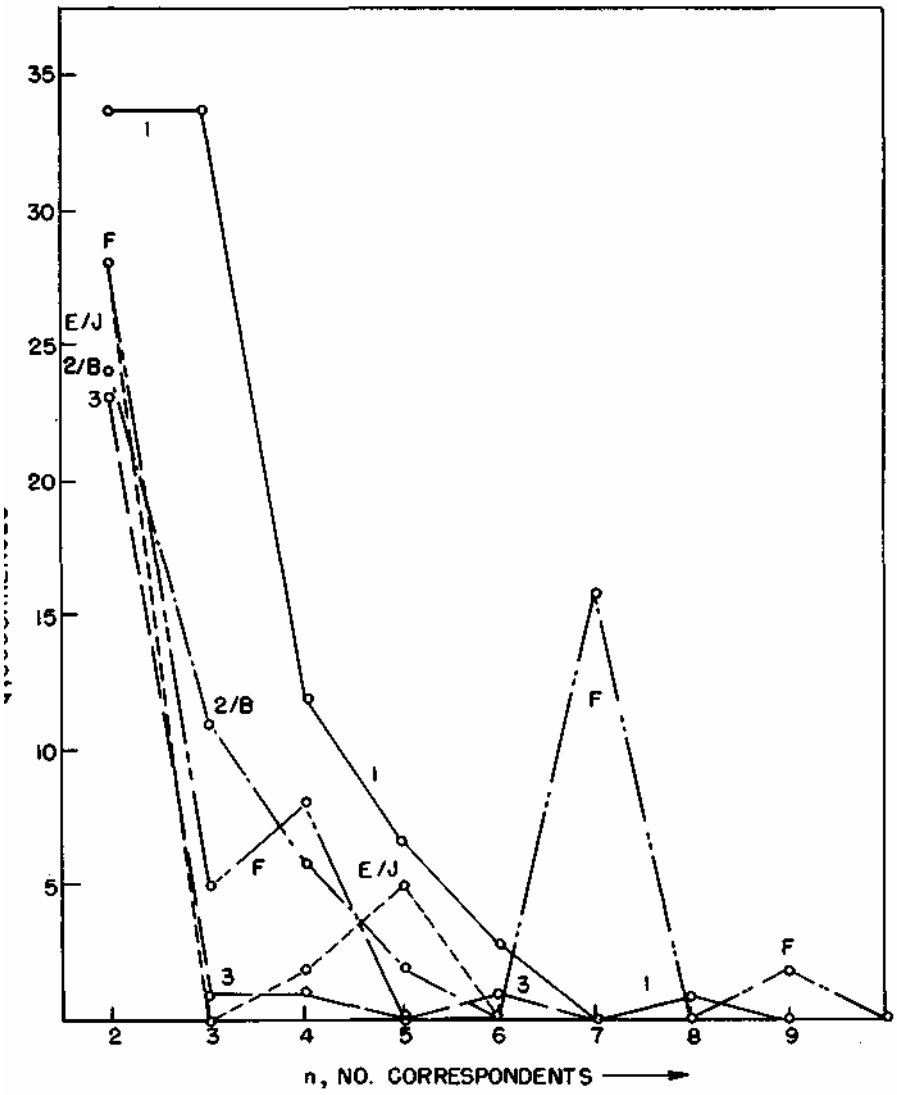
by Number and Class of Correspondents

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

Syntactical Variants†

Bjarne Ulvestad, Research Laboratory of Electronics,
Massachusetts Institute of Technology, Cambridge, Massachusetts*

Traditional grammar is normally eclectic and vaguely formulated, and it often tends to overgeneralize or fails to state the range of validity for its rules. Grammars for mechanical translation must be all-inclusive and rigorously explicit. While the input language grammar must register all the grammatical constructions possible, the existence of basically synonymous morphological and syntactical variants permits considerable inventorial reduction in the output grammar. These considerations are discussed with reference to English and German examples: verb phrases with 'remember'/ (sich) erinnern as the head; 'as if' / als ob clauses.

IT IS POSSIBLE to imagine a series of poor but successively 'better' machine-made translations, ranging from, say, 'very poor' to 'fair' or 'not so very poor,' which might be found to be substantially adequate for their various purposes. Thus even a lowest-grade or 'very poor' translation would conceivably have a demonstrable adequacy, provided its purpose were merely to acquaint its prospective readers with the subject matter of the original (input language) text.[1] Leading up from this kind of primitive, low-standard mechanical translation to one that would be regarded by the pundits as 'correct,' to the finest shades of idiomatic nuances, there is an almost discouragingly long, devious path, or rather a long series of shorter excursions each of which is more complex and laborious than its predecessor. If we, as we should, consider it imperative never to compromise with perfection where perfection is attainable, all the words and all the syntactical constructions of a given pair of languages, and especially of the one on the input side of the translation machine, will ultimately have been 'tagged' or assigned their specific memberships in a large number of groups and subgroups of linguistic entities, and the more exhaustive this intricate taxonomy, the more adequate, i.e., the less liable to produce ungrammatical and nonsensical sentence sequences, will be the corresponding translation mechanism.

The tantalizing question as to whether an absolutely foolproof apparatus for the mechanical transfer of information from one language to another can be constructed, if only in theory, need not bother us too much at this stage, for even if the answer to the question should in the end turn out to be negative, less-than-perfect mechanical translation will nevertheless be useful for scholars, whose main concern is naturally to obtain an adequate communication of scientific facts and ideas rather than stylistically impeccable texts, desirable though the latter may be.

Judging from reports on the highly significant work which is at present carried on at various universities, we have every reason to believe that most of the general technical problems of mechanical translation are approaching their solution. As an example of this kind of promising study, one may mention N. Chomsky's and V. Yngve's research into workable recognition devices for use in sentence-for-sentence translation, which is vastly preferable to word-for-word transfer. While the bulk of linguistic work in the field of mechanical translation has thus far admittedly been of a rather general

---

† This work was supported by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S.Navy (Office of Naval Research); and in part by the National Science Foundation.

* On leave from University of California, Berkeley, California; now at University of Bergen, Bergen, Norway.

1. Cf. J. W. Perry, "Translation of Russian technical literature by machine," MT, Vol. 2, No. 1, pp. 15-24 (1955).

and preliminary nature, researchers on both sides of the Atlantic are becoming more and more aware that the most pressing requirement for further progress is the composition of total-coverage grammars deliberately executed with mechanical translation in mind. We do not have such grammars for any language, except in rudimentary and fragmentary form, but even at this early date we can discuss some of their conspicuous features, as distinct from those of what we may term traditional grammars.

In this article a few problems in mechanical translation grammar will be presented and discussed, with some reference to their practical relevance to the input language and to the output language. English and German are the two languages chosen for this exposition. However, substantially similar problems will no doubt be found in any language.

We can state without reservation that in constructing grammars for the input language and for the output language, the input grammar must be subjected to the more piecemeal examination of particular problems. One of the most transparent reasons for this lies in the relatively large number of basically isosemantic morphological and syntactical variants that exist in every linguistic system. While all these variants will presumably have to be identified and registered in the input language grammar, considerable reduction in the number of corresponding variants will ordinarily be possible in the output grammar, as will be seen below. It must be emphasized that the chief difference between traditional grammar and what may be called mechanical translation (input language) grammar is that the former is eclectic and normally vaguely formulated, whereas the latter will be all-inclusive and rigorously explicit and formalized. Traditional grammars overgeneralize and rarely state the actual range of the validity of each rule; mechanical translation grammar must, ideally, explicate all the cases for which the given rule applies as well as those for which it does not. Furthermore, mechanical translation grammar must of necessity account for the total number of linguistic constructions that occur in a given language even if traditional grammars categorically state the nonoccurrence of certain members; [2] and misleading transformation rules must be recognized as such and correctly restated. [3] Whereas variant constructions of low statistical probabilities may on the whole be disregarded in the grammar of the output language, [4] they cannot, as a rule, be left out of the grammar of the input language without more or less serious consequences for the quality of the eventual translation. It is obvious from the remarks made above that the mechanical translation point of view will compel linguists to examine in detail problems that have hitherto been regarded as trivial or inconsequential. We can therefore expect that mechanical translation research will be of fundamental value to structural linguistics.

The important task of registering all syntactical variants, including those that are ordinarily overlooked in standard grammars, need not necessarily lead to a correspondingly greater complexity on the part of the eventual encoding program, although it may seem so at first glance. An example will perhaps help.

(1) Ich erinnere mich an ihn (den Mann)

(2) Ich erinnere mich auf ihn (den Mann)

(3) Ich erinnere mir ihn (den Mann)

(4) Ich erinnere mich ihn (den Mann)

(5) Ich erinnere ihn (den Mann)

(6) Ich erinnere mich seiner (des Mannes)

These German sentences are built around the weak verb (sich) erinnern 'remember' and corresponding to the English sentences 'I remember him' and 'I remember the man.'

---

2. Cf. B. Ulvestad, "Object clauses without dass dependent on negative governing clauses in modern German," Monatshefte, 47.329-38 (1955).

3. A typical instance is furnished by E. E. Cochran, A Practical German Review Grammar. 11th printing (New York, 1947), p. 241: "Note: zu after sagen is dropped in an indirect statement." The example illustrating this dropping of zu is: Er sagte zu mir: "Ich kann es mir nicht leisten," vs. Er sagte mir, er könnte es sich nicht leisten. That this rule is invalid in its present categorical formulation is seen from such sentences as: Er sagte zu Sabine, er werde sie . . . abholen (Brentano), Franz... sagte einmal zu mir, es gebe in jedem Dorf ein oder zwei schwere Taten (Wittich).

4. This consideration will be taken up for separate discussion in a later article.

Only (1) and (6) belong to the generally accepted standard language, and for that particular code the traditional formula, 'sich (acc.) erinnern is followed by a genitive construction or by the preposition an with an accusative construction,' is correctly stated, provided, of course, that one does not take 'followed by' literally. In normal modern German literary prose, however, one may encounter any one of the six types. Now, if we want to register every one of the sentence types with reflexive erinnern in the input code (this excludes 5), we need only add the verb erinnern not only to the class of reflexive verbs with the reflexive pronoun in the accusative case, but also to the class of verbs that may occur with the reflexive pronoun in the dative, and subsequently state, e.g., that the verb erinnern with accusative reflexive may 'govern' the accusative, the genitive, or a prepositional phrase with an or auf followed by an accusative noun phrase (NP). Since these entities will presumably have been registered and classified in some department of the grammar anyway, they do not have to be restated, but only referred to in terms of a defined code signal. This signal will indicate, for instance, that the verb (sich) erinnern belongs with denken in that it 'governs' an an-phrase with the accusative, and with sehen in that it takes an auf-phrase with the accusative.

If the purpose of the mechanical translation grammar and translation apparatus were restricted exclusively to the transfer of German scientific texts, sentence types (1) and (6) above would probably be the only ones that would need to be encoded. Even for translation of current novelistic prose we need only add (5), which occurs much more frequently than (2) and (3). In this kind of literary prose, the frequency continuum runs as follows, from very high to very low: (6)— (1)— (5) — (2) — (3)— (4).[5] If, on the other hand, a speaker of the Hamburg Umgangssprache were to be used as 'informant,' the first part of the frequency sequence would probably be (5) — (1); (6) can hardly be said to belong in this city language at all.[6]

Whatever the tasks for which the translation machine is designed, the encoding will not be made too difficult by the requirement of full coverage. It is the patient grammar writer whose difficulties are enhanced by new decisions to improve the translation.

It is interesting that if German were the output language, the situation in the examples above would be reversed and considerably less complex. As input, we would have English sentences with the verbs 'remember,' 'recall,' and possibly 'recollect,' all of which are closely related from the point of view of multiple-class memberships. With German as the output language, one of the six types above is sufficient for mechanical translation purposes since we are primarily interested in cognitive meaning transfer, not in the kind of additional information 'natural language' may furnish (age, sex, dialect, education, business background, etc.)

Naturally, the reduction of the number of variants in the output language to one is advisable only if the variants are absolutely free or if there is no possibility of making a meaningful selection out of two or more output variants on the basis of clues found in the input language. We snail explain this below with reference to a typical mechanical translation problem, using as examples German and English clauses which may be termed 'quasi clauses' (in English, 'as if'-clauses; in German, als ob-Sätze). Presentation of a grammar of these clauses for mechanical translation is the purpose of the remainder of this paper.

Variations on the following statement, with its examples, are current in textbooks of German: 'The secondary subjunctive (past subjunctive) is usual after als ob 'as if.' Er sprach, als ob er das Buch gefunden hätte. . . . ob may be omitted and inverted order used. ... Er sprach, als hätte er das Buch gefunden.'[7] It is not difficult to see that this 'quasi clause grammar' is far

5. The data for this were obtained from a corpus of 52 recent German novels; (3) and (4) occurred only five and three times, respectively, and there was a considerable frequency drop between (6), (1), and the rest.

6. Native informants refer to (6) as "stilted," "constructed," "archaic."

7. P.H. Curts, Basic German, revised ed. (New York, 1946), p. 71. It does not matter much whether one's description of als (ob, wenn) reads, (1) 'the ob, like the wenn, may be omitted,' or (2) 'the quasi conjunction is als, but ob or wenn may be added,' although logically (1) is preferable in a grammar of the spoken standard (Hochsprache popularly also called Schriftsprache). and (2) better corresponds to the usage actually found in the written (novelistic) language.

too fragmentary to be used except for introducing the 'rudiments of elementary German' to beginners; so we shall not take time to demonstrate its shortcomings. Rather, we shall attempt to write as complete a grammar of the German 'quasi clauses' as possible from the data available to us. Subsequently some practical problems with reference to the transfer processing will be discussed.

Let us consider the following six sentences.

(7) Ihm war, als habe er sie seufzen gehört (Waggerl)

(8) Es war, als ob noch einmal die Sonne, Wasser und Wind ... dem Oberleutnant in dieser Gestalt vor die Augen treten wollten (Tügel)

(9) Mister Wenner ging durch das Dorf, als wenn es gar keine Schwalbacher gäbe (Kirschweng)

(10) Und doch war es, wie wenn ein schieferblanker, tödlicher Ernst sich auf den ganzen Platz gelegt hätte (Goes)

(11) Wenn ich im Fahren lange hinaufsah, war es mir, der ganze Himmel käme auf mich zu (Bauer)

(12) Ich lief schnell, wie als gälte es, sich ein Landgut zu erobern auf diesem Gang (Goes)

Sentences (7) to (12) have different 'quasi' conjunctions (QC's), namely, als, als ob, als wenn, wie wenn, zero (Ø), and wie als. The internal relationships between these sentences will be seen from the following regrouping of (7) to (12) symbolized in terms of significant constituents (the symbol / is read 'or'):[8]

(7) --------, als + Vfin + NP + (Vinf / Vpp)

(12) --------, wie als----------

(8) --------, als ob + NP + (Vinf / Vpp) + Vfin

(9) --------, als wenn ----------

(10) --------, wie wenn ----------

(11) --------, Ø + NP + VP ----------

We symbolize the noun phrase and the potentially succeeding infinitive or past participle under one sign, Z [NP + (Vinf /Vpp) = Z]; and the relationship between (7), (12) on the one hand, and (8), (9), (10) on the other will be seen to be one of constituency permutation to the right of the QC. For further simplification of the structural statements, we may operate with three classes of QC's: QC_1 (als, wie als), QC_2 (als ob, als wenn, wie wenn), and QC_3 (zero).[9] Note that a comma always separates a clause from a succeeding dependent clause and accordingly stands in an immediate concatenation relationship with the conjunction. We can therefore (and this may be useful for mechanical translation encoding) subsume under the term 'conjunction,' for maximum mechanical translation signal power, the conjunction itself with the preceding comma, so that, for example, the symbol QC_1 shall be henceforth taken to mean 'comma followed by QC_1.' The six 'quasi' sentences can accordingly be written as follows:

I. (7), (12) ----------QC_1 + Vfin + Z

II. (8). (9), (10) ---------QC_2 + Z + Vfin

III. (11) ----------QC_3 + NP + VP

Further reduction, stating the transformation relationship between I and II in formal terms, is possible. For instance, one might state the rules: 'for transforming I into II. rewrite QC_1 as QC_2 reversing the order of Vfin + Z, and for transforming II into I, rewrite QC_2 as QC_1 reversing the order of Z and Vfin,' but further study would disclose that T I $\rightarrow$ II is correctly stated, and not the reverse T II$\rightarrow$ I. From er tat, als hätte er ihn nicht gesehen (I) we clearly obtain by this transformation: er tat, als ob er ihn nicht gesehen hätte (II), but there exist instances of so-called elliptic II-sentences that do not permit a direct transformation T II $\rightarrow$ I, for instance, er tat als ob er ihn nicht gesehen, in which the finite verb (here,

8. The mode of the finite verb in the ' quasi' clause is not considered at this point. Note that the term 'Vfin' in parentheses is used in a wide sense and includes so-called passive infinitives such as gehört werden, gehört worden sein, etc.

9. On a different level of analysis, one might make use of the structural relationships between (12) and a sentence such as es war mehr so, als hielte sich etwas an ihrem Bein fest (Nossack) and state that the adverb so in the governing clause can be shifted into the dependent clause and changing its status into that of a corresponding conjunction particle, thus: X + so, als + Y $\rightarrow$ X, wie als + Y. Note the positions of the comma in the two formulas.

hätte or habe) is dropped, or more correctly stated, does not occur. The ellipsis of the (readily predictable) finite verbs haben and sein after past participles is encountered occasionally in all subtypes of II, in (8) as well äs in (9) and (10), whereas the finite verb must always be made explicit in I. And the omission of haben / sein is not restricted to 'quasi' clauses. [Cf. the dependent clauses of sentences like er fragte, ob er ihn gesehen [habe / hätte] and als er nach Hause gekommen [war], fand er, dass. ....] This 'dropping' of haben / sein after past participles thus need not be specially explicated in the grammar of 'quasi' clauses; it will have been taken into account elsewhere. Another distinctive feature differentiating I and II may be adduced: The subjunctive mode of the finite verb, or rather the subjunctive ([er] höre, [er] ginge) or the nonovert, 'neutral, ambiguous' mode (indicative or subjunctive, such as [er] hörte, [er] suchte) is obligatory in the I-sentences, but not in the II-sentences; for instance, er tut, als höre / hörte er nichts, but er tut, als ob er nichts hört / höre / hörte, where hört is an overtly indicative weak verb. In a recent study of German 'quasi' sentences, based on twenty-four novels, no overt indicative finite verbs were found among 737 als-clause s (I), but fifteen were found among the 187 als ob- / als wenn-clauses (II) found in the corpus.[10] Consequently, the establishment of groups I, II, and III appears so far to be the simplest possible classification and if we include reference to the mode of the finite verb in the 'quasi' clause, the following three statements or formulas describe the grammar of the 'quasi' clauses in German:

I. QC_1 + Vfin subj + Z

II. QC_2 + Z + Vfin subj / ind

III. QC_3 + NP + VP subj /ind

Formulas I and II uniquely define German 'quasi' clauses. They can therefore be used directly, i.e., without additional specification, as clause identification formulas in standard written German. Thus X + I + Y or X + II + Y is normally sufficient information for establishing that one is concerned with sentences or sentence sequences that include 'quasi' clauses, e.g., er sagte, als hätte er nichts verstanden, dass er es morgen Versucher werde.[11] Here the 'quasi' clause is included in an indirect discourse sentence, and its special formula is simply X + QC_1 + Vfin subj + Z. Note that 'Vfin + Z' is an indispensable element in formula I, because of the nonunique function of als as a dependent clause conjunction (cf. als er nach Hause kam, etc.), whereas in formula II the element ' Z + Vfin' can be considered predictable, and the simplified formula X + QC_2 + Z would perhaps be an adequate statement for a sentence like am nächsten Tage lag er ganz still, als ob er tot wäre. The unique function of als ob as a conjunction makes this reduction possible.

Formula III is more recalcitrant in that its primitive form, (---------Ø + NP + VP) is also the statement of the structure of indirect discourse sentences with zero conjunction; e.g., er sagte, er sei krank. Actually, III formalizes a genuine overlapping or ambiguous sentence type. [Cf. such sentences as mir scheint, dass....., mir scheint, Ø..... , and mir scheint, als ob .....] Note that our token sentence (11) above can be translated either as '... it seemed to me as though..' or as '... it seemed to me (that)...,' with only trivial difference in cognitive meaning. There are two possible ways of solving the recognition problem in this case: (1) We can add specifications as to the context of the clause and state that zero is used as a 'quasi' conjunction after governing clauses such as mir ist, es scheint, or (2) we can drop III from our 'quasi' clause formulations altogether and consider it an indirect discourse formula only (the term 'indirect discourse' being used here in its traditional meaning). The second solution seems preferable for the following reasons: The zero

10. B. Ulvestad, "The Structure of the German Quasi Clauses," to be published in Germanic Review (1957).

11. This statement needs to be qualified to exclude some rarely occurring clauses that would seem to correspond to II in its present formulations. The following sequence was found in W.v.Niebelschütz, Verschneite Tiefen, (Berlin, 1940), p. 144: 'Doch wessen das Herz hier gierig ist, weiss niemand; nur ich. Vielleicht weiss es der Ritter auch? Mag sein. Mag es sein, es wäre leichter für mich, als wenn ich's ihm sagen müsste.' The clause starting with als wenn means: 'than if I had to tell it to him.' Such dependent clauses as this are found only after comparatives in the governing clauses, here, leichter.

Table I

	als		als ob		als wenn	
	c.pr.	c.pt.	c.pr.	c.pt.	c.pr.	c.pt.
1		63	1			
2	15	15				
3	63	15		2		
4	1	2		1		
5	3	35	1			
6	16	25	5	11		1
7		1	1			
8	9	5	1	2		
9		11		7		
10	1	8		5		
11	1	31		2		1
12	1	6	4	10		
13			1		11	12
14	1	38				
15	2	5		5		
16	38	7				
17	6	26		11		6
18	9	18				
19	4	24		7		
20	13	5	1	1		
21	3		1			
22	8	5	2	3		
23	10	14	1	4		
24	1	3	2	7		
Sums	205	362	21	78	11	20

Frequencies of chosen present subjunctive (c.pr.) and chosen past subjunctive (c.pt.) in three different 'quasi' clause types in novels by 24 authors.

conjunction occurs only after governing clauses like es scheint, mir ist, es kommt mir vor, and it is infrequently found. Only thirteen examples [such as mir schien, ich könnte sie aussprechen, jedoch fehlte das Wort (Zweig)] were found among 1168 'quasi' sentences taken from twenty-four works. This in conjunction with the basic similarities in meaning ('it seemed to me that / as though ....'), appears to furnish sufficient justification for operating with only two types of 'quasi' clauses, I and II, and our reduced grammar now simply reads:

I. QC_1 + Vfin subj + Z

II. QC_2 + Z + Vfin subj / ind

The tense-forms of the subjunctive in such clauses need not occupy us for long. In most traditional grammars, which are usually of the prescriptive type, statements indicating the obligatory nature of past subjunctive finite verbs are found. Table I amply demonstrates that these statements are untenable and unwarranted.

12. The term 'chosen present/past subjunctive' means that either tense form in a given case would represent the subjunctive mode unambiguously. In other words, we are interested in the ratios between the numbers of occurrence of such forms as, e.g., [er] sei, gehe, bringe (present subjunctive) and [er] wäre, ginge, brächte (past subjunctive). The names of the authors are of no importance in this context.

We would therefore be wrong in adding the word 'past' after 'subj' in formulas I and II; the correct statement is obviously one that does not specify tense-form. If German were the output language, (in which case we would be faced with a choice, see below) the grammar would read, at least for the literary style level:

I. QC_1 + Vfin subj past + Z

In this formula, QC1 would include only als, not wie als, and formula II would not occur in this grammar at all, unless compelling reasons for its inclusion were discovered.[13]

A similar problem emerges with regard to the translation of German into English: Should we register both 'as if' and 'as though' as correspondent conjunctions, and if not, which one would be preferable? Let us discuss this from the point of view of a particular transfer situation. The following German sentences are all grammatically correct:

Er tat, als ob er krank wäre
-------, als wenn----------
-------, wie wenn ----------
-------, als wäre er krank
------, wie als ----------

These sentences are, at least from the point of view of mechanical translation, isosemantic and can be translated as either 'he acted as if he were ill,' or 'he acted as though he were ill.' Therefore, NP + VP + 'as if' + NP + VP seems just as good a correspondence formula as NP + VP + 'as though' + NP + VP.[14] However, we would reasonably argue that the slightly 'elevated,' 'literary' connotation of 'as though' in contradistinction to the more 'colloquial' one of 'as if' corresponds to that of the German als (I) and als ob (II), respectively, in which case one may suggest as an adequate German-to-English transfer grammar of 'quasi' clauses:

I. QC_1 + Vfin subj + *Z*
→ 'as though' + NP + VP

II. QC_2 + Z + Vfin subj / ind
→ 'as if' + NP + VP

The concise 'quasi' clause grammar which we have worked out above could be further simplified within the context of a full-scale input grammar of German, because most, perhaps all, of the constituents would already have been described and classified. For instance, the two clauses in the sentence wenn er mich sähe, würde er grüssen belong in the same classes as some of the 'quasi' clause constructions after als in [er tat,] als wenn er mich sähe and [er tat,] als würde er grüssen, respectively.

The classification and coding of sentence elements and the subsequent elaboration of the simplest possible grammatical rules in terms of these classes are indispensable preliminaries to a successful construction of a workable translation machine. Every new grammatical statement will also represent a step forward in our scientific description of the language whose structure the grammar explicates and formalizes. The ultimate grammar will constitute the central prerequisite for a translation machine.

---

13. The reasons for preferring I (with als) to II (with als ob, als wenn) for the output grammar, if only one formula were to be employed, can be read out of the table.

14. A more complete discussion of the English correspondences would, of course, include such 'quasi' clauses as 'as though being ill.'

The Thesaurus in Syntax and Semantics†

M. M. Masterman, Cambridge Language Research Unit, Cambridge, England

The recent work of the Unit has been primarily concerned with the employment of thesauri in machine translation. Limited success has been achieved, in punched-card tests, in improving the idiomatic quality and so the intelligibility of an initially unsatisfactory translation, by word-for-word procedures, from Italian into English, by using a program which permitted selection of final equivalents from "heads" in Roget's Thesaurus, i.e. lists of synonyms, near-synonyms and associated words and phrases, instead of from previously determined lists of alternative translations. The Unit is investigating whether the syntactic properties of a word in a source language may be defined by a simple choice program, with reference to extra-linguistic criteria, which might be of universal or extensive interlingual application. It is hoped to combine or reconcile such a program with R.H. Richens's procedure for translating syntax by means of an interlingua, which has proved effective in a small-scale test. Studies have been made of the complementary distribution in literary English of words and phrases from "heads" in Roget, and of the construction of discourse from the contents of selected "heads." The possibility of producing a thesaurus better suited for machine translation purposes than Roget's, to be based on a more restricted lexis and a simpler categorization, is to be examined.

AT THE Second International Conference on Machine Translation, held at the Massachusetts Institute of Technology October 16-20, 1956, members of the Cambridge Language Research Group[1] presented four papers[2] which together opened up a new approach to certain linguistic problems of machine translation. As a result of discussions which followed, a Research Unit was formed at Cambridge, with the support of the National Science Foundation of the United States, to investigate these problems further.[3]

One of the great problems of machine translation is that of providing any device, programable on a machine, for translating idiomatic or metaphoric uses of word when these uses cannot be foreseen, since they may be occurring for the first time in the language which is being translated. To meet this problem, three of the Cambridge research workers, M.M.Masterman, A.F.Parker-Rhodes and M.A.K.Halliday, recommended that a mechanizable procedure for producing non-literal, "idiomatic" translations should be tried. This procedure required an

---

† This paper has been written with the support of the National Science Foundation, Washington, D.C.

1. The Group is a private, informal research society, most of whose members hold appointments in the University of Cambridge (see MT. Vol. 3, No. 1, p. 4). The Unit, concerned specifically with machine translation and library retrieval methods, was formed mainly from members of the Group, with some additional workers.

2. M.Masterman, "Potentialities of a Mechanical Thesaurus"; A.F. Parker-Rhodes, "An Algebraic Thesaurus"; R. H.Richens, "A General Program for Mechanical Translation between Any Two Languages via an Algebraic Interlingua" (reported MT, Vol.3, No.2); M.A.K. Halliday, "The Linguistic Basis of a Mechanical Thesaurus", now published MT, Vol. 3, No. 3.

3. See Annual Report of the National Science Foundation 1957 (in the press).

extra dictionary, compiled not on the principles of an alphabetic dictionary, but of a thesaurus,[4] to be inserted into the machine handling the target language. Thus, if the target language were English, the main part of the procedure would consist in retranslating an initially unsatisfactory translation, obtained by the word-for-word procedures long known to be feasible in machine translation, into idiomatic English. The actual translation procedure, moreover, did not consist, as had all mechanical translation procedures up to that time, of programing the machine to make a selection between the members of a finite set of antecedently given translations of a source language word. It consisted, on the contrary, of a procedure for mechanically producing from a thesaurus a finite set of extensive lists of synonyms of a particular word; that is, of a total dictionary in miniature; and of then choosing, by a two-stage procedure, firstly from among the lists, and secondly from among the synonyms. Thus, by looking up the word 'plant,' say in the cross-reference dictionary of a thesaurus, a set of numbers can be obtained, each standing for a list of synonyms, which might appear in one context, of the word 'plant:' "plant as place, 184: as insert, 300: as vegetable, 367: as agriculture, 371: as trick, 545: as tools, 633: as property, 780: – 'a battery,' 716: – 'oneself,' 184: – 'ation,' 184, 371, 780." This last represents an actual extract from the cross-reference dictionary of Roget's Thesaurus. Initially, the machine cannot know which of these lists of synonyms of 'plant' it should choose. But suppose that the word 'plant' were preceded, in the text, by the word 'flowering.' The cross-reference dictionary entry for flowering' is as follows: "flower as essence, 5: as produce, 161: as vegetable, 367: as prosper, 734: as beauty, 845: as ornament, 847: as repute, 873: – 'of age,' 131: – 'of flock,' 648: 'of life,' 127: – 'painting,' 556, 559." There is only one context in common between the context list of 'plant' and the context list of 'flowering,' namely, 367, 'Vegetable.' We therefore correctly assume that the synonym list under Vegetable is the synonym list required, if a synonym is in fact required for the basic word 'plant.'

The last stage in the procedure consists in comparing, in twos, the synonym lists which have been selected by the procedure given above in order to find which synonyms occur in common in these. Thus, if 'Woman' and 'Animal' are looked up in Roget's Thesaurus, and the synonym lists under each compared for common words, a single common word will be discovered, namely 'bitch.' These common words are then ordered, in descending order of frequency and the most frequent provide the retranslation output, certain restrictive rules having been brought into play which are designed to decide unambiguously which synonym shall replace each initially given pidgin English word. Sometimes, as in the case of 'plant,' in 'flowering plant,' the output is the same as the initially given word; this is taken as confirmation that the original translation was right. But sometimes, in the test cases presented at the Conference, the final output was significantly different from the original word. Thus, by using what came to be known as the "thesaurus procedure," it was shown that the Italian phrase alcune essenze forestali e fruttiferi. which had been translated, by a word-for-word translation procedure, 'forest and fruit-bearing essences,' could be retranslated 'forest and fruit-bearing examples [or specimens];' that the Italian phrase tale problema si presenta particolarmente interessante, which had been translated, by the word-for-word procedure, "such problems self-present particularly interesting,' could be retranslated 'such problems strike one as, [or prove] particularly interesting;' and that the Italian word germogli, which had been translated by the word-for-word procedure 'sprout,' could, though with difficulty, be retranslated 'shoot.' The papers made clear that the use of such a thesaurus procedure by no means always produced a correct translation. For instance, the phrase particolarmente interessante, which had been correctly translated by the word-for-word procedure 'particularly interesting,' was retranslated by the thesaurus procedure as 'What's the matter?' Nevertheless, the examples showed that a trans-

---

4. The only way of defining the notion of a thesaurus, in practice, is by reference to the famous work of Roget, Thesaurus of English Words and Phrases (Longmans, Green and Co.

5. Locke and Booth, Machine Translation of Languages (New York and London, 1955). See esp. Chapter II; Richens and Booth, Some Methods and Mechanized Translation.

6. I.S.Mukhin, An Experiment in the Machine Translation of Languages Carried out on the B.E.S.M. (Moscow. 1956); examples: 'category' (chart on p. 16); 'of' (chart on p. 17).

lation device which was programable on an electronic digital computer, but which made use of the intrinsic elasticity of words, could hope to deal, in a significant number of cases, with the hitherto unsolved problem of translating idiom, metaphor, and pun.

The fourth paper presented at the Conference, by R. H. Richens, made a different, though cognate, recommendation. In it the author recommended that a completely general interlingual notation, or set of symbols, should be used to produce syntactically correct translations between languages of different types, without any effort being made to translate directly between any given pair of languages. Richens showed, moreover, that by the use of such an interlingua, and by a mechanical procedure so simple that it could be effected not only by a digital computer, but by a punched card machine, a sentence could be translated with complete syntactical correctness from Japanese into the interlingua, and from the interlingua into English, German, Latin and Welsh. Thus the Japanese passage conventionally translated as: KETSU SAKU HO GO HEI ni ICHI SAKU to[2] ri SHU SHI RYU SU[2] ha KO HAI JI KI ni yo tsu te I ru was rendered into English as 'the percentage of matured capsules and the number of grains of seeds of one capsule are different according to the time of hybridizing;' into German as der Prozentsatz der gereiften Kapseln und die Zahl der Grane der Samen einer Kapseln sind gemäss der Zeit des Bastardierens verschieden; into Latin as ratio per centum capsulas maturandi et numerus granorum seminum capsulae unius secundum temporem hybridizandi diversa sunt; and into Welsh as y mae canran oeddfedu masglau a rhif gronynnau hadau un masgl yn wahanol yn ol amser croesi rhywiau. And Richens' claim, made in his paper, that his interlingua was algebraic has since been justified. When subjected to mathematical logical analysis, the Richens interlingual notation was shown to possess the characteristics of a weak mathematical system.

It might be thought that such revolutionary translation proposals as these, requiring as they do such an immense amount of computer storage, would be of merely academic interest to machine translators until computer research had developed to a point considerably in advance of that at which it now is. This is by no means the case, however. Information presented at the same conference, notably in a paper by Dr. Gilbert King, [7] made it clear that in the machine translation field, computer research is far in advance of language research; that, if the linguistic problems can be solved by any mechanizable procedure, computer engineers will find a way of programing the solution on to a machine. At a speech made at the Conference's final day, for instance, Dr. King said that procedures which had been brought forward at the Conference had convinced him that a machine could translate not merely as well as, but better than, an M.I.T. professor; since, having more storage space, it could produce a bigger vocabulary. Thus the papers presented by the Cambridge research workers at the Conference produced an atmosphere of technological hopefulness about the future prospects of mechanical translation, which did not, perhaps, take sufficient account of the fact that the basic linguistic problems, though tackled, were not yet solved.

After the Conference, it rapidly became clear to us that the generality of approach implied by the proposal to use a target language Thesaurus was cognate to, but not identical with, the generality implied by the proposal to use an algebraic syntactic interlingua. The more recent work of the members of the Unit has, therefore, been primarily directed towards making explicit the exact nature of the interrelations between these two proposals. For it is evident, on the one hand, that an interlingual claim is being made by the assertion that Language is such that, in it, metaphors and proverbs can, in some cases, be interchanged by means of a thesaurus. And, on the other hand, the analytic examination of Richens' interlingual algebra has established that it, itself, when interpreted, showed some, though not all the characteristics of a thesaurus. The question therefore arose: could the two methods be unified? Could an interlingual thesaurus somehow be conjoined to an interlingual syntactic notation to produce completely interlingual idiomatic mechanical translation from any language into any other? Conversely, could syntactical correctness as well as semantic elegance be introduced into the translation program at the stage of target-language retranslation by including a syntactic section within a thesaurus, so as to produce idiomatic multilingual mechanical translation from any source language into a single target language ?

---

7. King and Wieselman, Stochastic Methods of Machine Translation (International Telemeter Corporation, 1956).

Up to this point, the nature of the mechanical translation technique had required that the major part of the Cambridge Unit's analytic work should be performed by programmers and mathematical logicians, not by linguists; for the Unit's first need was to produce an analysis of the translation process which was both sufficiently general to justify the commercial production of a future mechanical translator, and also mathematically definite enough to be mechanizable. Now, however, it became clear that essential and fundamental considerations, regarding both the nature of comparative descriptive linguistics, and the nature of philosophic logic, were tied up in all this analytic work. For, to mention only one such consideration, the promoters of the thesaurus target-language procedure could, and on occasion did, claim that they were mathematicizing Plato; Richens, with an equal justice, could be said to be mathematicizing Aristotle. Thus, with sophistications on both sides, the age-old controversy in philosophy between nominalists and realists took, in the research conferences of the Cambridge Language Research Unit, a strange, fascinating, esoteric new turn.

Secondly, it became clear that if a well-grounded decision was to be made between the policy of interlingualizing the thesaurus, (that is, of assimilating semantics to syntax) and that of thesaurizing the syntax (that is, of including syntax within semantics) the linguists would have to be called in. In fact, for a time, they would have to be given charge. In the attempt to decide between these two alternatives, the Unit had developed two complementary lines of research. In the first, Richens designed an interlingual program complete with dictionary for translating syntax, beginning with translation from Italian into English, but subject to continual test by translation from other languages. In this test the object was to see how, with a very rough-and-ready method of translating metaphor and idiom, but with a very advanced and sophisticated method of translating syntax, intelligible translations of scientific texts could be made without using a thesaurus. In the second line of research, transformations were made from thesaurus-heads to texts and then back again within one language, without any procedure being used to translate from one language to another, or to translate syntax. The linguists were then invited to comment on and improve both of these lines, in order to see whether or not they tended to contrast or converge.

Halliday's sophistication of the Richens interlingual syntax translation program was of the following general form. For the general description of it I quote his own words:[8]

".. Translation.. is a form of comparative descriptive linguistics; but whereas translation between a given pair of languages requires only particular (one language) and comparative (in this case transfer, i.e. two languages) description, we envisage it as a requirement of mechanical translation that the program should be applicable to translation among all languages, and therefore we must face the necessity of universal (all languages) description ... Clearly if work was concentrated on a one-one translation field, where only a straight transfer description is required, results might be expected much more quickly. But the whole program might have to be remade for each pair of languages, and [so] it seems preferable to aim at a universal linguistic translation program applicable to translation between any pair of languages.

"This wider aim can only be achieved by a rigorous separation of the particular from the comparative universal range of validity (in MT terminology, of monolingual from interlingual features), and by their separate handling in the program ... The basic problem in the grammar is the setting up of relations among the particular grammatical structures of different languages ... It seems clear that considerable use can be made, both in the dictionary entry and in the operations, of the descriptive distinction between those chunks [separable segments of words[9]] which can be fully identified in the grammatical analysis (i.e. grammatical chunks or 'operators') and those only partially identified in the grammar and requiring further, lexical, information (i.e. lexical chunks or 'arguments'). This is of course an arbitrary distinction made for mechanical translation purposes; it reflects the different fields of application of the grammar and the dictionary in

---

8. From "The Linguistic Basis of Mechanical Translation" (Report for the Eighth International Congress of Linguists, University of Oslo, 1957; in the press).

9. See Richens and Halliday, "Word Decomposition for Machine Translation;" presented to the Georgetown University Eighth Round Table Meeting on Linguistics and Language Studies, April, 1957, and to appear in its Proceedings (in the press).

descriptive linguistics ... Comparative linguistics has the theoretical equipment [for establishing a universal description of syntax] by reference to categories of context grammar; and the systems of context-grammar categories set up for mechanical translation make up a grammatical interlingua such that any single language is capable of comparison with them. This grammatical interlingua .. is not a universal language, which would merely turn the number of languages we have to deal with from n to n + 1, but a set of systems of grammatical relations identified in context grammar, of the type that one sets up for the comparative identification of grammatical categories in descriptive linguistics .. The method [of setting these systems up] which seems at present likely to be most fruitful, and [which] is being tried out on a limited number of languages, (Italian, Chinese, English, Russian and Malay in the first instance), is [first] to establish a rigid operator/argument distinction, and [then] to identify the operators by their placing in a number (provisionally about 60) of two term systems each term being a yes-or-no function, . . The arguments are then classified by reference to grouping of these systems .."

Halliday's method, then, stripped to its essentials, is first to make a monolingual grammar of each language, and then, distinct from this, an interlingual analysis. The monolingual grammar is of the kind normally produced by descriptive linguists, except that it is only for the operators of each language; it is by reference to these operators that the arguments are, later, to be defined. This monolingual grammar can, at a later stage, be mathematically related to the interlingual analysis of these same operators, but is initially sharply to be contrasted with it, since it is to be based on extra-linguistic, not on intra-linguistic context.[10] The interlingual analysis, the making of which is the key to the whole problem, is achieved by the following method. With regard to each operator in question, the analyst asks himself a number of extremely simple questions, questions so simple, in fact, that he can unhesitatingly answer, with regard to them, "Yes," "No," "Both," "Neither" ("Neither" meaning "The question is inapplicable"). For instance, take the French operator la, the function of which, for mechanical translation purposes, is always very difficult to define, since, speaking vaguely, it can serve either as a feminine definite article or as a feminine accusative pronoun. We assume that la has already been monolingually placed within a set of monolingual grammatical systems, including a two-gender system, which apply to French only. We therefore feel free to ask, interlingually, not "Does la belong to any gender system?" because it is notorious that gender systems, as between languages, do not correspond, but, far more simply, "Can la, under any circumstances, tell us anything about sex?" Thus, by this change of question, we are exchanging a reference to the intra-linguistic context, (i.e., that of French) for the far more stable extra-linguistic context, i.e., that of the division of the human race into two sexes. English has no genders, French two, German three, Icelandic six; but Englishmen, Frenchmen, Germans and Icelanders alike all fall into communities consisting of two, and only two, sexes. Thus, with regard to the French operator la, when we ask, "Can it, ever, tell us anything about sex?" we can instantly and unhesitatingly answer, "Yes, it does." Proceeding to the next question, we ask, "Does la apply to animate/inanimate objects?" to which the answer is, "It applies to both." To the next question, "Does la apply to present/non-present time?" the answer is, "Neither; the question is inapplicable." "Does la refer to proximate/distant regions of space?" Answer, "Neither; the question is inapplicable. " (With regard to the French operator là this question could be answered; but not with regard to la), and so on. The heart of the whole method lies in the application of the precise and elegant methods used by contemporary descriptive linguistics to analyze monolingual context grammar (methods which amount in effect to analyzing the older compendium units "verb," "adjective," "noun" and the rest into weaker but more stably definable unitary components from which any required variant of the compendium units can be built up) to analysis of extra-linguistic context also (Halliday; June, 1957). In this latter case the extra-linguistic contexts can be universal ones, and the compendium units are the actual operators themselves. In other words, by taking seriously the analogy which has always been known to exist to some extent between intra-linguistic and extra-linguistic context, and by

---

10. M.A.K.Halliday, "Some Aspects of Systematic Description and Comparison in Grammatical Analysis" (Studies in Linguistic Analysis; Philological Society Special Volume, London, 1957).

treating the first as a straight extension of the second, Halliday has shown that he can achieve, for practical purposes, a non-contentious method of universal grammatical description. (By 'non-contentious' I here mean only, 'a method which will produce the same answers to the same questions when applied to the same operators by different analysts.') Moreover, the preliminary use of this method gives some provisional reason to think that the more complete and comprehensive the series of "Yes/No" questions which are asked (however large it is, the list will be objectively determinable and finite) the more closely the numbers of operators in each language come to approximate to one another. The result, if it is confirmed, will be very useful for mechanical translation, since it means that, with regard to any language, the operator category will be checked and redefined by the interlingual analytic process itself.

Thus Halliday's suggestion for sophisticating Richens' translation program is already of considerable research interest, since it shows that even so initially general and purely logical a research project such as that of Richens can be re-envisaged as arising out of a valid linguistic field. Halliday's suggestion is also hopeful in that preliminary research trials show that it does provide a paradigm, or model, for the rapid construction of operator dictionaries. Thus the Unit has plans to prepare such dictionaries in Italian, Standard Chinese, Cantonese, Malay, Hindi, Russian, Turkish, English, French, and German, these being the languages for which the dictionary makers are readily available. If the method justifies itself, other languages, without too much strain, can be added to these. The second consideration which can be derived from studying Halliday's schema is that he is, in effect, making a syntactical thesaurus. Several of the yes-no questions by which he establishes the components of his categories, for instance, "Does this operator apply to animate/inanimate objects?" "Does this operator assert a fact / give an implication?" "Does this operator indicate completion/non-completion?" "Does this operator indicate duration/non-duration?" could equally well be used as part of a schema for classifying synonyms under given thesaurus-heads. Thus a convergence between the interlingual and thesaurus approaches is detectable here.

What is not yet established, as must be made clear, is whether the additional complexity which Halliday desires to insert into the very simple and elegant translation program of Richens will really improve the quality of the translation produced by it. A test is being devised of the capacities of the original and amended versions to translate prepositional phrases. Meanwhile, another feature has emerged, in that Halliday's amendments to Richens' program have strengthened the case for coding this program to go through the computer by using the very general mathematical system known as lattice theory. (The use of lattice theory for the analysis of language will effect an analysis congruent to the ideas of those linguists who can, in any sustained way, imagine language as a net. On a first approximation, a lattice is an asymmetric net; a finite lattice is a fishing net or hammock, though an asymmetric one; that is, a net with a single top point and bottom point. Such nets are built up from a single asymmetric binary relation, which itself derives, though over some distance of time, from the asymmetric binary relation used by George Boole, and which was suggested to him by the linguistic adjective-noun relation.) Preliminary grounds for using this mathematical system to algorithmize the translation of syntax had already been given in earlier papers by the members of the Unit. [11] Moreover, the fact that the Richens interlingua had already been shown to constitute an algebraic system weaker than lattice theory, though not incongruent with it, increased the ground for re-mathematicizing it by trying on it a mathematical system of the same kind as itself, though of more algorithmic power. And Halliday's analysis, being as it is in terms of dichotomies, (and of systems which can be constructed by successions of dichotomies) straightforwardly uses lattice theory by its very nature. Either, therefore, it must be compressed and coded by initially using this system, or it cannot be compressed and coded at all. Some idea can be gathered, however, of the extent of the complication which Halliday's suggestion introduces into Richens' program from the fact that whereas an entry of 20 bits (20 binary digits) per chunk would have sufficed Richens to translate both meaning and syntax, Halliday's amendment will require an entry of at least 120 bits

---

11. See MT, Vol. 3, No. 1, pp. 2-28 (report on the Colloquium of the C. L. R. Group, August, 1955); and M. Masterman, "The Comparative Analysis of a Chinese Sentence, " (annex to the report, available from the Editor of MT).

per chunk for syntax translation alone. Fortunately, Dr. Gilbert King, who was mentioned earlier, and who now is a member of the Unit's Consultative Committee, considers it feasible, from the engineering point of view, to construct a mechanical translator which will perform lattice operations but not arithmetical ones, and which will allow of chunk entries 1, 000 bits long.[12] For existing computers, however, Halliday's schema would be too complex by far. This should not blind us to its intrinsic interest or to its many potential advantages; but it should be borne in mind by those linguists who are seriously interested in developing machine translation as a concrete reminder that, for every increase in linguistic analytic complexity, a heavy electronic price has to be paid.

Turning now from syntax without semantics to semantics without syntax, a word must be said about the Unit's second research project, namely that of examining the interrelations between texts and their constituent thesaurus-heads without the complicating intervention of a foreign language. Dr. E. W. Bastin, Karen Jones, M. M. Masterman, R.H.Needham, A.F. Parker-Rhodes, A.R.Penny, Dr. R.H.Thouless and W.F. Woolner-Bird have made the principal contributions.

The first provisional discovery made by the members of this research group was that paragraphs of lecture-style discourse could, without difficulty, be constructed by the intuitive use of a minimum number of thesaurus-heads. Thus a paragraph dilating pompously but not vacuously on the present peculiar scientific position of the study of parapsychology was constructed by Dr. Thouless and Margaret Masterman, for thesaurus demonstration purposes, using only four lists of thesaurus synonyms to supply all the argument words. These lists concerned the generic ideas of 'Wonder' (with a cross reference to 'Interest'), 'Science,' 'Parapsychology,' and of a very general topic within which 'Appearance in Thought' contrasted with 'Instantiation in Reality,' the two combined heads forming an antithetic pair. The method by which the paragraph was constructed was suggested by one of the Unit's programmers, Lady Hoskyns. If Interest be A1, Wonder A2, Instantiation in Fact B, Psychical Research C and Science D, then the paragraph constructed by Dr. Thouless can be thesaurized as follows:

" 'Interest' [A1] in 'psychical research' [C] is often 'motivated' [A1] by 'wonder' [A2] at 'phenomena [C] which 'appear to be' [B] 'marvellous' [A2]. The 'sitter' [C] is 'amazed'[A2] at the 'wonderful' [A2] 'results' [D and B] of 'card-guessing experiments' [C] which 'leave him in a state of' [B] 'bewilderment' [A2], 'seeming' [B], as they do, 'to savour of' [B] 'necromancy' [A2]. This 'attitude' [A1] of 'awe' [A2] (or of 'admiration' [A2], as it would earlier 'have been called' [B]) 'produces' [B] a 'fascination' [A2] with the 'subject' [C and D]. The 'new-comer's' [C] 'surprise' [A2] 'leads' [B] often to 'stupefaction' [A2], and the 'research' [D] is 'treated' [D] as a 'sensation' [A2] rather than as a 'serious' [A1] 'branch of science' [C and D]."

Other paragraphs, giving the obituary of an imaginary well-known biologist, an advertisement for a film star, and a denunciation of the British Conservative Party, were similarly constructed. The introduction of a randomizing procedure, with the object of mechanizing the selection of synonyms, caused a paragraph of esoteric theology, and also one denouncing philosophic scepticism, to be a little more irrational than they would otherwise have been, but not very much. Attempts rapidly followed to use this method to construct parody (Thouless and Parker-Rhodes); to simulate essay writing (Woolner-Bird); and to employ it to analyze chapters instead of paragraphs (Needham and Jones). Several facts of considerable interest emerged. One was that, in any kind of writing which builds up into an argument, thesaurus-heads tend to be introduced in powers of two, each topic being introduced concurrently with that to which it primarily contrasts. Another was that the introduction of a new thesaurus topic, in discursive writing, tends to follow a clustering of re-allusions to a single one of the topics which have been introduced earlier, and which are themselves synonymous, in such a way as to force the selection of the new thesaurus-head. This result was reached independently by Woolner-Bird and by Needham and Jones (by analysis of Southern, Cultural Aspects of European Territorial Expansion.) A third fact which emerged was that, if the unit to be analyzed consisted of a chapter, rather than a paragraph (that is, of a piece of discourse with an order of, say, 20 enlarged

12. G.King, The Requirements of Lexical Storage (International Telemeter Corporation, 1957).

thesaurus-heads), a sub-class of these heads, say, 2 or 4, will have vastly more synonyms of themselves occurring in the chapter than will any of the others; so that this sub-class of heads, taken in a prescribed ordering, can be taken as a title for the whole chapter. A fourth fact, of very general interest, was that there are some thesaurus-heads which always have to be constructed to analyze discourse; that is, which occur so constantly that it seems almost impossible to think without them. One of these conveys the very idea of a synonym: "is, constitutes, appears to be, seems to be equatable with, shows itself to be, constitutes the fact that; namely, that is, in other words; could be called, could be treated as, could be considered as; this comes to saying, this comes to the same thing as saying. . " These and their like appear in every text; (including the present report). So do synonyms of the very general generic idea of causation: "causes, promotes, produces, leads to, determines, results in; the result is, the upshot is, in the end, we find that we can say that.. " So do synonyms for the very basic idea of appearing to be one thing, while turning out in fact to be another. (This generic idea precedes nearly every introduction of contrast.) Since these thesaurus topics so constantly occur, it might be argued that their constituent synonyms were functioning as a queerly determined class of syntactical operators, rather than as arguments. Moreover, since, in order to analyze the chapter of a book into its constituent thesaurus-heads, a distinction has to be established, and in a non-contentious manner, between new ideas (formalized by P), qualifiers, to be taken as a single element with what they qualify (formalized by Q's) and re-allusions to ideas previously mentioned (formalized by R's); and as all these have to be distinguished from Q's, or operators, it becomes clear that if Halliday, to translate syntax, has to construct a new type of universalized thesaurus, so also the thesaurus makers, in order to analyze the semantic patterns occurring in texts, have to construct a very basic, simple kind of syntax. All of which gives reason to hope that in some way (the members of the Unit do not yet see how) the interlingual program for translating syntax, and the analytic program for constructing texts from thesaurus-heads, or thesaurus-heads from texts, may all turn out to be different parts of the same program, in the end.

In conclusion, a final word must be added on one problem of thesaurus construction which the members of the Unit will have to face squarely if they are to construct a full-scale translation thesaurus. The creative ability of man is not so easily amenable to mechanization, in this field, as the Unit's early, gaily-reached results, would seem to imply. In other words, with every text we analyze it becomes increasingly evident that every discursive writer constructs his own thesaurus. How then is the Unit to construct a thesaurus which has any hope of applying to more than one text?

One immediate reply to this capital difficulty is by asking another question: "How, equally, does any linguist compile a dictionary which fully applies to more than one text?" In a paper on categorization of lexis, recently read to a meeting of the Language Research Group at Cambridge, R. A.Crossland suggested that a procedure of selection out of a thesaurus-head, alternative or preferably supplementary to any procedure based on contextual distribution, might be based on the traditional dictionary-maker's technique of classifying words as appropriate to particular general contexts or types of diction. [13] Such indication is given only sporadically and somewhat unsystematically in most existing dictionaries, but, with refinement, it might provide a technique for programing the computer to make an appropriate choice from among the possible alternatives in a thesaurus-head, especially when this is to be used in the final stage of translation. Two methods of providing this selection suggest themselves. Either information about the appurtenance of a word in a source language to different dictions ("high" or "low" style, the styles of various technologies, etc. [14]), is recorded and passed through the interlingual stage, though the computer in that stage translates just an approximate lexical equivalent (the key word of a thesaurus-head, perhaps). Or else, without the recording and transmission of such information, an appropriate equivalent, out of a head "labelled" according to the appurtenance of its constituent elements to different dictions, would be selected in accordance with general

13. Diction seems now to be virtually a synonym in philological discussion for "verbal or written style" (cf. Oxford English Dictionary).

14. Crossland noted the element of subjectivity involved in categorization not based on detailed analysis of contextual distribution within restricted textual material.

and immediate context, (either by the procedure described earlier, or by some other mechanizable procedure to be substituted for it), within the set of such heads constituting the "rough output."

If any of these suggestions proves fruitful, it would seem likely, on the face of it, that new thesauri will have to be prepared, or existing ones reorganized by "labelling" of items and no doubt by addition, deletion and rearrangement, for languages between which translation is envisaged. Also it might be useful to prepare thesauri on the basis of particular scientific or other specialized "dictions." These could be considered valid in practice for fairly extensive categories of writers, though in principle the argument that every writer has his own thesaurus, based on what he alone desires to write or has written, seems reasonable enough.

Whether the Cambridge Research Unit will really succeed in compiling such a gigantic, universally valid, thesaurus of thesauri is not yet clear. What is clear, in the sense that it is becoming established as a thesis supported by considerable factual evidence, is that when a human being thinks discursively he does use a thesaurus. Secondly, it is intuitively clear, in the sense that it follows from this, that somehow or other, human beings do succeed, in discursive argument, in communicating to one another the boundaries of their respective thesauri; for if they did not, there would be no argument. We know this; for when communication fails to take place, we say, "I cannot understand the writer; he is too allusive." What we say, in making such a comment, is the opposite of what we actually mean; because what we mean is that such a writer does not take the trouble to order and display the re-allusions to his main ideas sufficiently for us to "catch" his personal procedure of synonym creation; that is, sufficiently for us to ascertain his thesaurus. And when we say this, it is further intuitively clear that we must be referring to some objective communication-promoting procedure; some procedure which we use, without being aware that we use it, whenever we argue discursively with one another.

The task that confronts us, then, though formidable, is not hopeless. Objective synonym-creating procedures which can be employed, can also be discovered; and logicians, dictionary makers and descriptive linguists are just the men to discover them.

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.

Bibliography

A. A. Lyapunov and O. S. Kulagina 106
Ispol'sovanie vyčislitel'nych mashin dlya perevoda s odnogo yazyka na drugoy
Priroda. Avgust 1955, Ottisk IZ, No. 8, pp. 83-85

This is a superficial outline of the IBM demonstration of January 1954 at Georgetown.
G. H. Matthews

107
I. K. Bel'skaya, L. N. Korolev, I. S. Mukhin, D. Yu. Panov, and S. N. Razumovskiy
Certain Problems of Automatic Translation
Vestnik Adakemii Nauk SSSR. Vol. 26, No. 12, 1956, pp.24-33

The authors give a general description of the English-to-Russian translating program, mechanical dictionary design, and solution to the multiple meaning problem (by looking at context) used in the BESM.
G. H. Matthews

T. N. Moloshnaya 108
Certain Questions of Syntax in Connection with Machine Translation from English to Russian
Voprosy Yazykoznaniya, No. 4, 1957, pp. 92-97

In this article a description is given of investigations which are being conducted in order to formulate rules for machine translation. Although orthography, morphology and lexicology are cited as problems, attention is concentrated on multiple-meanings, idiomatics and syntax. The author emphasizes the fact that the formulation of rules for mechanical translation is based on convenience and practical considerations and may, therefore, require departures from established procedures of scientific linguistics. Sample analyses of English sentences are included. These analyses are based on word classes defined by Fries and syntactic models given by Jespersen.
J. R. Applegate

R. E. Wall, Jr. 109
Some of the Engineering Aspects of the Machine Translation of Language
Transactions Paper No. 56-693, AIEE

A nondetailed presentation of the general "state of the art" of the mechanical translation problem. The paper is addressed primarily to engineers and indicates some of the more promising methods of attack on individual problems in the four steps of the translation process: encoding (encoding text material into machine language), memory search (amount, nature, sequence, and indexing of material stored in the machine's memory), logical operations (the utilization of word context), and decoding.
E. S. Klima

J. Poulet 110
Grammaire universelle pour machines à traduire
Le Linguiste, No. 3, pp. 3-6, No. 4, pp. 3-8, No. 5, pp. 4-8, 1957

In these three articles the author describes a coding system which he proposes for use in mechanical translation. During the preliminary dictionary look-up routine of input language words, the words will be labelled according to this system so that necessary syntactic operations (change of word order, inflection etc.) can be performed on these symbols according to rules formulated for specific output languages. The fact that all necessary grammatical and syntactic information can be expressed in this code makes it applicable to all languages.
J. R. Applegate

L. N. Korolev 111
Coding and Code Compression
Doklady, AN USSR, Vol.113, No. 4, 1957, pp.746-747

This paper discusses alphabetic coding, dictionary coding and code compression.
K. C. Knowlton

N. D. Andreyev 112
Machine Translation and the Problem of an Intermediary Language
Voprosy Yazykoznaniya, No. 5, 1957, pp. 117-121

This article deals with the problem of defining an intermediate language which might be used to simplify the problem of machine translation if such translation is to be carried out on a large scale. Because a binary system of translation (from one language to one other language) requires a separate program for each pair of languages, the development of an intermediate language would reduce the number of separate programs required. Statistical studies are proposed to determine the optimum form of the intermediate language. Relative positions of adjective and noun as well as the relative positions of subject and predicate are cited as examples of the methods to be used in determining the optimum form.

J. R. Apple gate

M. A. K. Halliday 113
The Linguistic Basis of a Mechanical Thesaurus
Mechanical Translation, Vol.3, No. 3, pp. 81-88

The grammar and lexis of a language exhibit a high degree of internal determination, affecting all utterances whether or not these are translated from another language. This may be exploited in a mechanical translation program in order to cope with the lack of translation equivalence between categories of different languages, by the ordering of elements into systems within which determination operates and the working out by descriptive linguistic methods of the criteria governing the choice among the elements ranged as terms in one system. Lexical items so ordered form a thesaurus, and the thesaurus series is the lexical analogue of the grammatical paradigm.

Author

A.D. Booth 114
Mechanical Translations
Aslib Proceedings, Vol.9, No. 6, (June 1957) pp. 177-181

This article presents a discussion of a few ideas for the structure of a mechanical dictionary and its use in the translation of idioms and ambiguous forms.

G. H. Matthews

I. A. Mel'chuk 115
Conference on Problems of Development and Construction of Information Machines
Voprosy Yazykoznaniya, No. 5,1957, pp. 161-162

Brief summaries of papers presented at the conference are given and the conclusions reached by the participants are stated. The fact that it is impossible to solve practical linguistic problems for information and translation machines before theoretical problems of methods for syntactic and phonological analysis, etc. are solved is stressed. The author recommends closer cooperation among linguists, mathematicians, and engineers to further the development of linguistics as an exact science using the methods of mathematics and mathematical linguistics. The regular publication of the results of work in mathematical linguistics and mechanical translation is also urged.

J. R. Applegate

116
A. D. Booth, L. Brandwood, J. P. Cleave
Mechanical Resolution of Linguistic Problems
Academic Press, New York, Butterworths, London, 1958, 306 pages

This book contains an account of some of the results which have been obtained at Birkbeck College Computational Laboratory on the application of digital calculators to linguistic problems. The chapters are:

1. Historical Introduction
2. The Nature of Calculating and Data Processing Machines
3. The Analysis of Content and Structure
4. Stylistic Analysis
5. General Aspects of Language Translation
6. Programming Technique for Mechanical Translation
7. The Mechanical Transcription of Braille
8. French
9. German
10. Russian
11. Multi-Lingual Translation
12. Technical Details of a Proposed Translating Machine

Over half the book is devoted to chapter 9. This is a collection of the main problems to be faced in translating German together with some suggestions on how to overcome them. The approach of Oswald and Fletcher is taken as a basis.

V. H. Yngve

G. A. Miller and J. G. Beebe-Center 117
Some Psychological Methods for Evaluating the Quality of Translations
Mechanical Translation, Vol. 3, No. 3, pp. 73-80

The excellence of a translation should be measured by the extent to which it preserves the exact meaning of the original. But so long as we have no accepted definition of meaning, much less of exact meaning, it is difficult to use such a measure. As a practical alternative, therefore, we must search for more modest, yet better defined, procedures. The present article attempts to survey some of the possible methods: One can ask the opinion of several competent judges. Or, given a translation of granted excellence, one can compare test translations with this criterion by a variety of statistical indices. Or a person who has read only the translation may be required to answer questions based on the original. The characteristic advantages and disadvantages of each method are illustrated by examples.

Author

118

Robert E. Wall, Jr. and Udo K. Niehaus
Russian to English Machine Translation with Simple Logical Processing
Transactions Paper No. 57-1062, AIEE

Improvements in word-for-word translation of Russian-English material can be made by considering grammatical usage of the word in the sentence. Storing grammatical tags instead of English prepositions allows for certain logical processing which improves the quality of the translation, solving 70 to 80% of the grammatical problems. The logical processing considers only the tags from the dictionary and compares the tags of the first member of a grouping with those of the second and so on to the end of the grouping. This eliminates grammatical usage possible for the given member in isolation, but not possible in co-occurrence with a subsequent member of the grouping. As a measure of the quality of the translation after logical grammatical processing vs. that of a strict word-for-word translation, a logarithmic criterion might be the most realistic, where

translation ambiguity = log n
(n = number of possible sequences that can be formed considering all grammatical possibilities of the words.)

E.S.Klima

H. P. Edmundson and D. G. Hays 119
Studies in Machine Translation—2: Research Methodology
The RAND Corporation, Santa Monica, Calif., P-1251, December 15, 1957

The first in a series of papers describing the methods now in use at the RAND Corporation for research on machine translation of scientific Russian. At each stage of refinement, automatic computing machinery is used for some aspects of translation and for collecting and gathering data about other aspects. The latter, when analyzed, are used to improve the MT program. The method gradually reduces the extent of the work done by the human editor.

E. S. Klima

S. N. Razumovskii 120
On the Question of Automatizing the Programming of Problems of Translation from One Language into Another
Doklady. AN USSR, Vol.113, No. 4, 1957, pp.760-761

The translation process is described as consisting of three kinds of operations: logical, identity, and arithmetic. A scheme is presented whereby each of these kinds of operation may be written in symbolic notation. It is proposed that from this input, computer programs may be used to synthesize the logical operations in machine language, compress the completed program, and divide it into parts.

K. C. Knowlton

A. Koutsoudas 121
Mechanical Translation and Zipf's Law
Language, Vol.33, No. 4 (Part 1), October-December 1957, pp.545-552

A problem which arises in the course of research on mechanical translation is the prediction of dictionary size. This article investigates the relation between empirical frequency laws and the function V(n) — the expected number of different words in an n-word sample of text. It is found that the probability-law proposed by Joos (1936) yields results which do not check well with experiments, and it is concluded that some modification of it is necessary for the purpose of vocabulary prediction.

Author