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

SOURCE-LANGUAGE SPECIFICATION WITH TABLE LOOKUP AND HIGH-CAPACITY DICTIONARY*

by

LEW R. MICKLESEN

(IBM Research Center
Yorktown Heights, New York)

AT the present state of the art of machine translation most energy is being expended on the elaboration of schemes for the hierarchical specification of source-language texts. It is presumed that such schemes, although they graphically illustrate the law of diminishing returns, are well worth the effort they entail because they will provide the elusive acceptable translations sought by machine methods. At the present time this hierarchical specification has progressed as far as the level of the sentence where attention is understandably concentrated because of the many problems and because of the critical linguistic interest in the sentence. It seems perfectly reasonable to conceive of this entire progressive specification solely as an optimally ordered series of lookups in a high-capacity, rapid-access dictionary. The photoscopic disc memory developed at the IBM Research Center answers this description and is actually being incorporated into a complete table-lookup system for the automatic specification of Russian sentences for Russian to English machine translation. The reader will immediately observe that in MT operations, just as in essentially linguistic work, there are no autonomous levels; there exist very complex, inescapable interrelationships. But just as in a complete grammar where these complex interrelationships are specified in terms of sentence building and morphological and phonological regularities are formalized at different levels, so also the steps in an MT procedure yield to a kind of leveling.

In an exclusively table-lookup procedure for the automatic translation of languages, the source-language specification begins with the construction of a dictionary consisting of absolutely unique entries. If the memory device employed were of unlimited capacity, this task would be relatively simple but extremely tedious, especially for highly inflected languages. All words and/or phrases would be stored in their several forms. This method

*This work was supported in large part by the Rome Air Development Center, Griffiss Air Force Base.

presents two dubious advantages because it forces the attention of the dictionary compilers on every inflected form of every word. It practically insures the detection of every case of homography and the careful and detailed grammar coding of every individual form. It seems that some real grammatical soul-searching would be required for even a native speaker of Russian to discover cases of homography between the fairly uncommon forms XORÓWEH*, instrumental singular feminine of adjective XOROWI1 (= good), and XORÓWEH, first person singular present tense of the verb XOROWET6 (= to grow better-looking), and SIN4, feminine predicative form of the adjective SINI1 (=blue), and SIN4, present gerund of the verb SINIT6 (= to turn blue), unless he wrote out and gave some thought to these very words. On the other hand, however, this method is contra indicated if only because the physical production of these entries is so time consuming.

This method is disadvantageous also because it forces dictionary compilers to worry unnecessarily about whether a given form is or is not used and forces them to include in the dictionary in their entirety forms whose incidence in any body of texts would be very low. In the final analysis, there is no practicably infinite memory, and economy must be intelligently instituted if processing reliability and grammar coding need not be prejudiced.

If, then, the storage of all inflected forms of a language presents some serious difficulties for reasons of economy and unnecessary decisions, linguistic units which at one and the same time permit economy of storage and uniqueness of entries must be chosen. Stems which usually coincide with actual grammatical stems and which, in conjunction with appropriate sets of endings, permit the recognition of all nominal, adjectival, or verbal forms, constitute the only choice although they must be created with care because shorter entries naturally reduce specificity. Here is a good example of the strict interrelationship of levels of specification. This first level of specification is concerned only with recording Russian words economically and uniquely. The problem here is with stems and words, but grammatical information plays an indispensable role in decision making. Whether or not a stem can be utilized at all for a member of a paradigmatic form class depends on the requirements for absolutely unique dictionary entries. For example, the two stems DN (= day) and DN (= bottom) are graphically identical; therefore the full paradigmatic forms of one or the other must be

*To facilitate typing and because actual or intended dictionary entries are used throughout the paper, all Russian linguistic forms will be written in the IBM 407 code representation. The coded form of Russian alphabetic symbols is immediately recognizable except in the following cases: Ж = J, Й = 1, Ц = Q, Ч = C, Ш = W, Щ = 5, Ъ = 7, Б = 6, Э = 3, Ю = H, Я = 4.

stored in the dictionary. Considerations of economy dictate that the full forms of DN (= bottom) be stored because this stem is valid only for the singular, whereas the "day" stem can be utilized for both singular and plural.

Whether or not a given stem can be used to identify ***all*** members of its paradigm depends on whether there exists another stem that would effect a longer match on any paradigmatic forms of the former stem. This complication can best be illustrated with the stems DEL (= matter/affair) and DELA (= do/make). Any paradigmatic forms of DEL that contain the sequence of symbols DELA must be included in toto in the dictionary to obviate so-called short matches on the part of the stem DELA. Thus, the words, DELAM, DELAMI, DELAX are specified by memorizing them in full. The word DELA, which exactly matches the "do/make" stem, can be neatly specified by creating the entry DELA#, utilizing the space sign acquired from the input text.

Stems longer or shorter than what would be generally regarded as grammatical stems are created out of considerations of unique dictionary entries (longer stems) or of greater economy through derivation (shorter stems). Verbal entries frequently illustrate both types of artificial stems. The verb DELIT6 (= to divide) suggests a stem DEL, which, plus an appropriate set of endings, would account for all the inflected forms of the verb; but because it is homographic with the nominal stem DEL (= matter/affair), it must be substituted by three stems: the grammatical stem, DELI, and two artificial stems DELH and DEL4. Thus the basic entries for the verb "divide" are rendered unique in terms of the nominal stem "matter/affair". If there is no conflict of stems, stems shorter than grammatical stems can be utilized for increased economy. In the case just cited above, if the stem DEL had not been ambiguous, it could have served as the verbal stem; and the grammatical stem DELI and the two longer stems DELH and DEL4 would have been superfluous.

The possibility of exploiting derivation, both intra- and inter-form-class derivation, introduces an even greater degree of economy. Stems shorter than grammatical are frequently employed for this purpose. Intra-form-class derivation is practically exclusively represented by verbal entries. The non-linguistic stem PEREASSIGN (= to reassign) is an excellent example of a stem that will effect identification of all paradigmatic forms of PEREASSIGNOVAT6 and PEREASSIGNOVYVAT6. Inter-form-class derivation by means of steins shorter than grammatical does not seem practicable because it involves too much pre-processing that is likely to yield too little economy. A possibility, but only a remote one, is the storage of a "short" noun stem like KOPEE (= kopeck.), which will combine with -K to produce the genitive plural of the noun and with the form -CN- to produce all forms of the adjective. Some details and implications of derivation in the table-lookup method will be discussed in succeeding sections of this paper.

The preceding discussion about creating unique dictionary entries for inflectional and derivational processes has tacitly assumed, but never explicitly stated, the fundamental feature of the whole dictionary system, the principle of longest match. This principle states that, in the process of reading the dictionary, the computer will always try to find an entry in the dictionary that will match on the longest sequence of symbols in the input text. The preceding Russian examples should have demonstrated the application and usefulness of the principle as well as the caution to be observed in working with it. The next paragraph introduces an area where the principle of longest match operates most dramatically.

Thus far the discussion has centered about the economical and unequivocal identification of stems and words. Since longer entries are more nearly assured of uniqueness and lose the advantage of economy, beyond the word in the phrasal level these considerations rapidly fade into the background, and the specification of source-target semantics becomes practically the paramount consideration. All types of phrases may play a role in semantic specification, but problems of possible discontinuity, inflection and ambiguity suggest that only those phrases enjoying 100-percent predictability should be subjected to a simple lookup, i.e., words and stems where there is no discontinuity and where inflection is easily ascertained. This subject will be broached again in a few paragraphs and should be clarified at that time. Suffice it to say for the present that the aforementioned constraints demand phrases which are uninflected, non-discontinuous, and semantically unambiguous. Such requirements can be fulfilled only by cliches, formulas, and proverbial expressions:

VYN6#DA#POLOJ6	right of way
GUBA#NE#DURA	nobody’s fool
DEWEVO#I#SERDITO	cheap but good
ESLI#BY#DA#KABY	if ifs and ans were pots and pans
SREDI#BELA#DN4	in broad daylight

Very few phrases of this type seem to exist in a highly inflected language. The first and the last examples show that archaic forms like "POLOJ6" and "BELA" are of considerable assistance in locating this kind of phrase.

The concept of word or stem and their specification in terms of the primary dictionary lookup by means of the principle of longest match has been explained. The nest level of specification is concerned with linking each stem with its particular set of inflectional endings and with the particular, presently recognizable, inherent grammatical information associated with each ending of the set. This specificational level touches only paradigmatic forms, whereas the previous level involved both paradigmatic and non-paradigmatic. It should be worth observing at this point that this level of specification is definitely the second. It can be completely formulated only after the nature of the stem dictionary is known. The inherent grammatical information spoken of is grammatical information conferred upon full words outside of context and is opposed to what will be termed contextual grammatical information in succeeding paragraphs. Inherent grammatical information proceeds from both stem and ending. It is maximally general (as opposed to contextual grammatical information) and yields profitably to classification. A properly constructed classification of paradigmatic stems and their endings leads to mutually exclusive subsets of endings for each form class. It also produces a classification scheme which permits all inherent grammatical information to be associated exclusively with the ending. This promotes some ease of handling data and, as it turns out, is of great utility in increasing economy when taking advantage of intra- and inter-form-class derivation. This capability will be illustrated in due time, but first it is imperative to expound the basic classification.

A brief survey of this classification problem would seem to reveal a task of undue proportions, but closer examination discloses that stems longer than grammatical stems and, therefore, stems that are not immediately predictable, occur only among the verbs because of the fact that the complex endings of the verbal system permit various stems with variable combining power. On the other hand, the less highly inflected nouns and adjectives can be stored practically only in their really grammatical stem forms except when the requirement of uniqueness compels storage of complete words. It is to be understood that the presence of such complete words in the dictionary in no way harms the effectiveness of a general stem classification because by virtue of the principle of longest match they would be identified before the stems could operate on them. For example, the stem DEL belongs to the following nominal declensional pattern:

		N	G_1	G_2	D	A	I	L_1	L_2	NP	GP	DP	AP	IP	LP	CF
DEL	N75	-O#	-A		-U	-O#	-OM	-E		-A	-#	-AM	-A	-AMI	-AX	-O

*The alpha-numeric sequence N75 has been termed a "confix", i.e., contextual prefix, which directs the search mechanism of the dictionary from the stem to the appropriate ending.

This classification can remain in effect even though the special entries DELA#, DELAM, DELAMI, and DELAX are required by the presence in the dictionary of the verbal stem DELA. Accordingly, the nominal and adjectival stem classifications are relatively easy to establish although the patterns of homography are exasperatingly varied and actual examples of theoretical possibilities are frequently impossible to discover. But once one has control of all factors involved, the reasonable theoretical possibilities can be provided for by an open-ended classification. These classifications can be presented in chart form to expert grammarians who can thus clarify with maximum efficiency the massive vocabularies required by automatic language translation. It is worth notice that such a system relieves the classifiers of the exacting demands of grammar coding which becomes an entirely separate operation. See ***Chart 1*** (nominal) and ***Chart 2*** (adjectival) for illustration of a suggested format, bases of classification, and types of inherent grammatical information. The last column makes provision for combining forms so that extemporaneous compounds may be recognized. Note how simply the homography of combining forms with other forms can be solved by adding a "space" symbol to full-word forms in order to increase their specificity.

It was stated previously that the verbal classification would create a special problem because only here are there significant numbers of unusual stems. It is true that these unusual stems are most frequently not grammatical stems, but they prove to be highly predictable because beyond the stem-final consonant of any verbal stem the vowels -A, -4, -E, -O, -U, -H and the consonants -T, -M, -N, -5, -L, -W, -V may occur as the final symbols of augmented stems forced into existence by the possibility of shorter matches. Thus, the augmented stem DEL4 (= divide) is a frequent imperfective stem type that can be profitably combined with three endings -#, -T, and -5 to produce respectively the present gerund, the 3rd person plural present, and all forms of the present active participle. In like manner, the augmented stem KINU (= throw) is a frequently occurring perfective stem type that combines with the endings -#, -T#, -L, -VW, -V/-WI#, and T6 to yield respectively the 1st person singular present, the 3rd plural present or the masculine short form of the past passive participle, the other forms of the past passive participle, the finite forms of the past, all forms of the past active participle, the two forms, of the past gerund and the infinitive. The augmented stem VIDIM (= see) can be used with -# and with most adjectival endings to identify respectively the 1st person plural present or the masculine short form of the present passive participle and all other forms of the present passive participle. It would be practically impossible to predict which augmented verbal stems would or would not occur in a dictionary of 100,000 stems and in combination with what ambiguities inherent in verbal

CHART No.1

SUGGESTED NOMINAL CLASSIFICATION

EXAMPLES	Confix	N	G_1	G_2	D	A	I	L_1	L_2	NP	GP	DP	AP	IP	LP	CF
EJ- BRAT-	NØ1	-1 -# -6	-4 -A		-H -U	-4 -A	-EM -OM	-E#		-Y -I -64	-EV -6EV -E1 -OV	-4M -64M -AM	-EV -6EV -E1 -OV	-4MI -64MI -AMI	-4X -64X -AX	-E -O
UCITEL- DOKTOR	NØ3	-6 -#	-4 -A		-H -U	-4 -A	-EM -OM	-E#		-4 -A	-OV -E1	-AM -4M	-OV -E1	-AMI -4MI	-AX -4X	-E -O
SLUCA- LIST-	N18	-6 -1 -#	-A -4		-U -H	-6 -1 -#	-OM -EM	-E#		-Y -I -64	-EV -6EV -E1 -OV	-AM -4M -64M	-Y -I -64	-AMI -4MI -64MI	-AX -4X -64X	-E -O
UGOL-/UGL- SUK-/SUC-	N19	-#	-A		-U	-#	-OM	-E	-U	-Y -I -64	-OV -6EV	-AM -64M	-Y -I -64	-AMI -64MI	-AX -64X	-O
MAT6-/MATER- LOWAD-	N53	-# -6	-I		-I	-# -6	-6H	-I		-I	-E1	-4M	-E1	-4MI -6MI	-4X	-E -6
LEDI- MADAM-	N54															
OBLAK- SLOV-	N75	-O#	-A		-U	-O#	-OM	-E		-A	-# -OV	-AM	-A	-AMI	-AX	-O
UCILI5- KOP6-	N76	-E#	-A -4		-U -H	-E#	-EM	-E#		-A -4	-# -E1 -EV	-AM -4M	-A -4	-AMI -4MI	-AX -4X	-E

CHART No.2

SUGGESTED ADJECTIVAL CLASSIFICATION

EXAMPLES	CON FIX	MASCULINE							FEMININE							NEUTER						PLURAL							PREDI-CATIVE				ADVERB	IMPERSONAL PRED.	COMPARATIVE$_1$	COMPARATIVE$_2$	COMBINING FORM
		N	G	D	A$_1$	A$_2$	I	L	N	E	D	A	I$_1$	I$_2$	L	N	G	D	A	I	L	N	G	D	A$_1$	A$_2$	I	L	M	F	N	P					
KRUT- PLOX-		-01	-OGO	-OMU	-OGO	-01	-YM,-IM	-OM	-A4	-01	-01	-UH	-01	-OH	-01	-OE	-OGO	-OMU	-OE	-YM,-IM	-OM	-YE,-IE	-YX,-IX	-YM,-IM	-YX,-IX	-YE,-IE	-YMI,-IMI	-YX,-IX	-#	-A	-0#	-Y	-0#	-0#	-EE	-E1	-0
MOLOD- PR4M-		-01	-OGO	-OMU	-OGO	-01	-YM,-IM	-OM	-A4	-01	-01	-UH	-01	-OH	-01	-OE	-OGO	-OMU	-OE	-YM,-IM	-OM	-YE,-IE	-YX,-IX	-YM,-IM	-YX,-IX	-YE,-IE	-YMI,-IMI	-YX,-IX	-#	-A	-0#	-Y	-0#		-EE	-E1	-0
DELOV- NOCN-		-01	-OGO	-OMU	-OGO	-01	-YM,-IM	-OM	-A4	-01	-01	-UH	-01	-OH	-01	-OE	-OGO	-OMU	-OE	-YM,-IM	-OM	-YE,-IE	-YX,-IX	-YM,-IM	-YX,-IX	-YE,-IE	-YMI,-IMI	-YX,-IX									
SVEJ-		-I1	-EGO	-EMU	-EGO	-I1	-IM	-EM	-A4	-E1	-E1	-UH	-E1	-EH	-E1	-EE	-EGO	-EMU	-EE	-IM	-EM	-IE	-IX	-IM	-IX	-IE	-IMI	-IX	-#	-A	-0#	-I	-0	-0	-EE	-E1	-E
XOROW-		-I1	-EGO	-EMU	-EGO	-I1	-IM	-EM	-A4	-E1	-E1	-UH	-E1	-EH	-E1	-EE	-EGO	-EMU	-EE	-IM	-EM	-IE	-IX	-IM	-IX	-IE	-IMI	-IX	-#	-A	-0#	-I	-0#	-0#			-0
SIN-		-I1	-EGO	-EMU	-EGO	-I1	-IM	-EM	-44	-E1	-E1	-HH	-E1	-EH	-E1	-EE	-EGO	-EMU	-EE	-IM	-EM	-IE	-IX	-IM	-IX	-IE	-IMI	-IX	-8	-4	-E	-I	-E#		-EE	-E1	-E

endings. The only recourse is to provide for all mathematical possibilities. Such provisions are included in the following fragment from a suggested verbal chart to facilitate verbal classification in a table-lookup system. See ***Chart No.3.***

It has been mentioned in this section that the association of all presently recognizable inherent grammatical information, including even the form-class specification, with the ending greatly facilitates the exploitation of both intra- and inter-form-class derivation and significantly increases the capacity of the dictionary. Intra-form-class derivation would seem to be more easily exploitable than inter-form-class derivation because the kernel meaning usually remains the same and with it the target-language equivalent, which may be modified in some regular way. The primary candidates for exploitation are productive and/or thoroughgoing derivational processes. Within the form classes of nouns and adjectives the most wide-spread derivational processes are the production of diminutives and augmentatives. For example, diminutives are generated regularly from the following nouns:

LES	(forest)	=	LESOK
VETER	(wind)	=	VETEROK
OGON6	(fire)	=	OGONEK
DEREVO	(tree)	=	DEREVQO
OKNO	(window)	=	OKONQE
RUJ6E	(gun)	=	RUJ6EQO
GOLOVA	(head)	=	GOLOVKA
MAWINA	(machine)	=	MAWINKA
PEC6	(stove)	=	PECKA

And augmentatives can be formed regularly from these nouns:

DOM	(house)	=	DOM15E
GOROD	(city)	=	GOROD15E
VEDRO	(pall)	=	VEDR15E
SILA	(strength)	=	SIL15E
JARA	(heat)	=	JAR15E

Similarly, the following adjectives may form regularly both diminutives and augmentatives:

CHART NO.5

SUGGESTED VERBAL CLASSIFICATION

EXAMPLES and SPECIFICATIONS	Con-Fix	PRESENT									IMPERATIVE		PAST							
		1S	2S	3S	1P	2P	3P	Part. Act.	Part. Pass.	Ger.	S	P	M	F	N	P	Part. Act.	Part. Pass.	Ger.	
KAPRIZNIC- (IMPERFECTIVE) 1P = Pres.Part. Act.Masc. 2P = Imper.D	V01	-LH -U -4H -HH -H -AH -UH -NU	-IW6 -AEW6 -UEW6 -NEW6 -EW6 -4EW6 -HEW6	-IT -AET -UET -NET -ET -4ET -HET	-IM -AEM -UEM -NEM -EM -4EM -HEM	-ITE -AETE -UETE -NETE -ETE -4ETE -HETE	-AT -UT -AHT -UHT -4T -HT -4HT -HHT -NUT	-A5 -U5 -AH5 -UH5 -45 -H5 -4H5 -HH5 -NU5		-A -A4 -U4 -4 -44 -H4	-6 -A1 -U1 -N1 -1 -41 -H1	-6TE -A1TE -U1TE -NITE -1TE -41TE -H1TE	-L -AL -OVAL -# -IL -4L -EVAL	-LA -ALA -OVALA -ILA -4LA -EVALA	-LO -ALO -OVALO -ILO -4LO -EVALO	-LI -ALI -OVALI -ILI -4LI -EVALI	-VW- -AVW- -OVAVW- -W- -IVW- -4VW- -EVAVW-	-N- -LEN- -AN- -EVAN- -EN- -AN- -OVAN- -T-	-V -IVWI -4V -OVAVWI -VWI -AV -4VWI -EVAV -IV -AVWI -OVAV -EVAVWI	-T6 -AT6 -OVAT6 -IT6 -4T6 -EVAT6
VID- VIJ- (IMPERFECTIVE) 1P = Pres.Part. Act.Masc. 2P = Imper.P	V02	-LH -U -4H -HH -H -AH -UH -NU	-IW6 -AEW6 -UEW6 -NEW6 -EW6 -4EW6 -HEW6	-IT -AET -UET -NET -ET -4ET -HET	-IM -AEM -UEM -NEM -EM -4EM -HEM	-ITE -AETE -UETE -NETE -ETE -4ETE -HETE	-AT -UT -AHT -UHT -4T -HT -4HT -HHT -NUT	-A5 -U5 -AH5 -UH5 -45 -H5 -4H5 -HH5 -NU5	-IM- -4EM- -HEM- -AEM- -UEM	-A -A4 -U4 -4 -44 -H4	-6 -A1 -U1 -N1 -1 -41 -H1	-6TE -A1TE -U1TE -NITE -1TE -41TE -H1TE	-L -AL -OVAL -# -IL -4L -EVAL	-LA -ALA -OVALA -ILA -4LA -EVALA	-LO -ALO -OVALO -ILO -4LO -EVALO	-LI -ALI -OVALI -ILI -4LI -EVALI	-VW- -AVW- -OVAVW- -W- -IVW- -4VW- -EVAVW	-N- -LEN- -4N- -EVAN- -EN- -AN- -OVAN- -T-	-V -IVWI -4V -OVAVWI -VWI -AV -4VWI -EVAV -IV -AVWI -OVAV -AVAVWI	-T6 -AT6 -OVAT6 -IT6 -4T6 -EVAT6
NOSI- (IMPERFECTIVE) 1P = Pres.Part. Act.Masc. 2P = Imper. P.	V10		-W6	-T	-M	-TE			-M-		-#	-TE	-L	-LA	-LO	-LI	-VW-		-V -VWI	-T6
MERZNU- (IMPERFECTIVE)	V12	-#					-T	-5												

TEPLY1	(warm)	=	TEPLOVATYl	(warmish)
BELY1	(white)	=	BELOVATY1	(whitish)
SINI1	(blue)	=	SINEVATY1	(bluish)
TOLSTY1	(thick)	=	TOLSTU5I1	(very thick.)
BOL6W01	(big)	=	BOL6WU5I1	(very big)
GR4ZNY1	(dirty)	=	GR4ZNH5I1	(very dirty)

It is perhaps needless to observe that diminutives and augmentatives are too expressive for technical literature. The few that may occur with any frequency can best be handled by storage in toto. Other intra-form-class derivational possibilities within the nouns or adjectives are characterized either by low semantic predictability or by zero or low productivity. Although the formation of many feminine nouns in -KA, -WA from corresponding masculine nouns suggests some interesting possibilities:

WVED	(Swede)	=	WVEDKA
QYGAN	(gypsy)	=	QYGANKA
ANGLICANIN	(Englishman)	=	ANGLICANKA
PIANIST	(pianist)	=	PIANISTKA
STUDENT	(student)	=	STUDENTKA
KASSIR	(cashier)	=	KASSIRWA
KONDUKTOR	(conductor)	=	KONDUKTORWA

The verbs, however, offer a splendid opportunity of increasing dictionary capacity because aspectual derivation is all-pervasive in the system, is productive, and is semantically predictable to a great extent. The principal aspectual relationship, imperfective-perfective, is the only one that can be profitably exploited; iterative verbs are too infrequent and the indeterminate-determinate relationship concerns too few verbs and too many irregularities. Both perfective and imperfective verbs can be formed from a common artificial stem either by specific sets of endings or by the use of confixes plus specific sets of endings. The following examples should amply illustrate this capability. All forms of the aspectual pair ZAGORET6 / ZAGORAT6 (= to become tanned) can be identified from the common stem ZAGOR, which would be associated in a tactile entry with the confix V72 directing the search to the following set of unambiguously perfective and imperfective endings:

PRESENT

	LS	2S	3S	IP	2P	3P	Part. Act.	Part. Pass	Ger.	etc.
IMPERFECTIVE	-AH	-AEW6	-AET	-AEM	-AETE	-AHT	-AH5		-A4	etc.
	-4H	-4EW6	-4ET	-4EM	-4ETE	-4HT	-4H5		-44	
	-H	-IW6	-IT	-IM	-ITE	-AT				
						-4T				
PERFECTIVE									etc.	
	-U	-EW6	-ET	-EM	-ETE	-UT				
						-HT				

It may occur to the reader that another confix could be embedded in a set of exclusively perfective endings that would recognize all forms of ZAGORET6 and in the case of forms of ZAGORAT6 would direct the search to a set of exclusively imperfective endings. It turns out, however, that an exhaustive set of perfective endings contains endings also common to ZAGORAT6; and, according to the principle of longest match, this verb would be identified as perfective in all its forms. In order to solve this problem a special set of perfective endings would have to be constructed for a verb stem like ZAGOR; so the scheme outlined above is entirely feasible.

Specifically perfective or imperfective confixes, however, offer real advantages. They permit the storage of fewer stems and fewer endings. For example all forms of the aspectual pair UBAHKAT6/UBAHKIVAT6 (= to lull) can be matched by the single stem UBAHK plus a series of confixes and two sets of endings. The entry,

UBAHK V4Ø

directs the search to a set of perfective endings including the confix V4ØIVA. If the text word is a form of UBAHKIVAT6, this confix then will direct the search to the appropriate set of imperfective endings where all the requisite inherent grammatical information can be acquired. Likewise, in the case of the aspectual pair IZBEGNUT6 / IZBEGAT6 (= to avoid), all forms can be produced from the single entry:

IZBEG VØ2

which directs the search to a set of imperfective endings containing the confix VØ2NU, which, in the case of all forms of IZBEGNUT6, will direct the search to the correct set of perfective endings.

Inter-form-class derivation is a common phenomenon in Russian as it is in all languages. It is very tempting, indeed, to increase the capacity of the dictionary by taking advantage of the most easily processed and most productive derivational patterns. This procedure entails some careful and fairly extensive pre-processing, most of which, however, can be accomplished for a table-lookup procedure in a straightforward manner during dictionary construction. The easiest solutions are arrived at in instances where the English equivalent for both form classes involved in the derivation remains the same. The noun-adjective pairs ABSOLHT - ABSOLHTN and NEON - NEONOV will serve to illustrate the point. The adjectives are formed from the nouns, both of which belong to the declension with confix NØ1, so the adjectival formants -N- and -OV- will appear as the following entries:

NØ1N	A13
NØ10V	A25

Note that an extra specification would be necessary so that a given noun using formant -N-, say, could select the proper set of adjectival endings. This extra information is likely to be very limited if the derivational process is productive. The location of the inherent grammatical information in the stem dictionary insures that the proper grammar tag is selected whether the input word is a noun or adjective. Where either noun or adjective is combined idiomatically with some other specific element or elements as in phrases:

ABSOLHTNY1 3FIR	absolute ether
ABSOLHTNO CERNOE TELO	ideal black body

nothing is lost because the correct form-class information, vital to phrasal recognition, stands ready for use. The requisite grammatical information for matching of the phrase is augmented by matching on a coded representation of the basic Russian elements of the phrase. More discussion will be devoted to this point later in the paper.

A complication may seem to arise when additional English morphemes are needed to render the proper English equivalent for the derivative. Thus, the English form "-ic" should augment the word "alcohol" in the derivation of ALKOGOL6N- from ALKOGOL-, and the English form "-al" should augment the word "pentagon" in the derivation of the Russian adjective "PENTAGONAL6N-" from the noun "PENTAGON".

This apparent problem can be resolved by processing such English morphemes along with all other determinations of English output, i.e., in a final lookup devoted entirely to English synthesis based on information developed during the processing of the Russian sentence. The adjectival form-class specification should prove to be sufficient information to accomplish selection of the English adjectival formant so that it may be suffixed to the basic semantic equivalent for the noun-adjective stem. If this kind of operation is to be a basic capability of a table-lookup system and if one final lookup is the most effective spot for determination of target-language equivalents and if specific target-language formants can be selected at this time with no apparent difficulty, then a target-language equivalent for the adjective, say, entirely different for the equivalent for the noun, should be capable of being chosen merely on the basis of the form-class information. Such a derivational pair would be exemplified by the Russian noun-adjective pair ZRITEL6 / ZRITEL6NY1 where the nominal equivalent is "spectator" and the adjectival equivalent is "visual". This selection process is similar to, but even easier than, a syntactic situation exemplified by the Russian word TEXNIKI = technology/technicians, where only extensive processing involving considerations of gender, number, and case may lead to an unequivocal choice of one equivalent or the other.

Thus it seems that a table-lookup system can cope with any sort of derivational problem arising in MT operations. The only constraint is a rather vague but practical one and poses the question: just at what point does the amount of pre-processing of MT linguistic data become impracticable? Syntactic considerations may clearly favor as much morphological processing as possible, but this point must be reserved for discussion below.

The combined power of a table-lookup system and a high-capacity rapid-access memory is graphically illustrated in their treatment of all types of phrases. In a single-pass system which recognizes but does not record grammatical features, only non-discontinuous phrases can be identified. Non-paradigmatic phrases require only single entries:

V ANTRAKT	=	during the intermission
V KANUN	=	on the eve
V KOI VEKI RAZ	=	once in a great while
VSPLOT6 I OKOLU	=	all around

while paradigmatic phrases may require several entries:

KO3FFIQIENT#POGLO5ENI4	=	absorption	coefficient
KO3FFIQIENTØ#	"	"	"
KO3FFIQIENTØØ#	"	"	"
KO3FFIQIENTØØØ#	"	"	"

BEGAH5ØØ#LUC	=	scanning beam
BEGAH5ØØØ#LUC		" "

The Ø's employed in the above entries will match on any character and will permit recognition but not necessarily complete or proper translation of all inflectional endings. On the other hand, a single-pass system that both recognizes and records grammatical features by virtue of extensive preprocessing permits some rather interesting manipulations although again several entries may be required. Phrases like KO3FFIQIENT#POGLO5ENI4 (absorption coefficient) and LES#NA#KORNH (standing timer) where the variable portion, the ending, occurs within the phrase may be processed by essentially removing the invariable portion of the phrase, giving it a preferential translation, and then determining the grammatical and semantic properties of the variable portion. The series of entries required for this type of treatment of these two entries reads as follows:

KO3FFIQIENT#POGLO5ENI4	$\rho_{11}\rho K$	δ11absorption coefficient
KO3FFIQIENTØ#POGLO5ENI4	$\rho_{12}\rho_K$	δ11absorption coefficient
KO3FFIQIENTØØ#POGLO5ENI4	$\rho_{13}\rho_K$	δ11absorption coefficient
KO3FFIQIENTØØØ#POGLO5ENI4	$\rho_{14}\rho_K$	δ11absorption coefficient

$\rho_{11}\rho_{K\#}$	δ11	grammatical	tag
$\rho_{12}\rho_{KA}$	δ12	"	"
$\rho_{12}\rho_{KU}$	δ12	"	"
$\rho_{12}\rho_{KE}$	δ12	"	"
$\rho_{12}\rho_{KY}$	δ12	"	"
$\rho_{13}\rho_{KOM}$	δ13	"	"
$\rho_{13}\rho_{KOV}$	δ13	"	"
$\rho_{13}\rho_{KAM}$	δ13	"	"
$\rho_{13}\rho_{KAX}$	δ13	"	"
$\rho_{14}\rho_{KAMI}$	δ14	"	"

The first group of entries gives a partial specification of the complex ending in terms of number of characters and in terms of the stem class or specific stem, by means of a confix, e.g., $\rho_{12}\rho_K$ where ρ_{12} equals the number of symbols in the ending and ρ_K refers to the stem. The δ11 then shifts out the ending, and yields the correct grammatical information. Essentially, the same kinds of entries would specify all grammatical variants of the phrase LES#NA#KORNH. Note, however, that the special confixes created are specific only for stems of indicated lengths and ending affiliations. The above example indicates how a specific phrase may be identified. This method serves admirably for the recognition of phrases whose elements form classes as, for example, in the case of the chemical nomenclature for salts. The following sets of entries will provide the proper translation and grammatical information for chloride salts of sodium and potassium:

1)	XLORID	ρ_1	δ∅
2)	$\rho_1$000000#NATRI4	ρ_2	δ∅sodium#
3)	$\rho_1$000000#KALI4	ρ_2	δ∅potassium#
4)	ρ_2XLORID	ρ_3	chloride
5)	ρ_3#	ρ_4	grammatical tag
6)	ρ_3A	ρ_4	" "
7)	ρ_3U	ρ_4	" "
8)	ρ_3OM	ρ_4	" "
9)	ρ_3E	ρ_4	" "
10)	ρ_4NATRI4		
11)	ρ_4KALI4		

Entry 1) links XLORID with the remaining portions of the phrases. Entries 2) and 3) translate the remaining portions of the phrases in their proper order. Entry 4) directs the search back to the first portion for proper translation in proper order. Entries 5) through 9) match specific endings for grammatical information. Entries 10) and 11) remove the final portions from the search area so they are not unnecessarily translated.

This kind of operation makes the additional requirement that the English equivalent for the uninflected portion of the phrase always precede the English equivalent for the inflected portion because English plural information must be suffixed to the end of the phrase. Because of this requirement Russian phrases like:

NEODNOZNACNOST6 RELElNOl XARAKTERISTIKI (ambiguity of relay characteristic)
DOPUSTIMA4 OBLAST6 OTKLONENI1 (admitted region of deviations)

would have to appear in a machine translation as "relay characteristic ambiguity" and "admitted deviation region". But, because of the versatility of English nominalizations, it seems difficult to locate a phrase of this type whose translation is not wholly satisfactory.

A nominal phrase like BEGAH5I1#LUC in this second single-pass system would still have to appear in two entries, but each would be accompanied by the appropriate confix N18 for the noun LUC in order to isolate the proper grammatical information. The entries would appear as follows:

BEGAH5ØØ#LUC	N18
BEGAH5ØØØ#LUC	N18

This kind of nominal phrase, too, may have elements that belong to classes. Again a certain type of chemical salt offers an interesting example of rearrangement combined with recognition of grammatical information.

1)	BROMIST	ρ_1	δØ
2)	BROMN	ρ_2	δØ
3)	ρ_1ØØØØØØØØØ#JELEZ	ρ_3	ferrous#
4)	ρ_1ØØØØØØØØØØ#JELEZ	ρ_3	ferrous#
5)	ρ_2ØØØØØØØ#JELEZ	ρ_3	ferric#
6)	ρ_2ØØØØØØØØ#JELEZ	ρ_3	ferric#
7)	ρ_3BROMISTØØ#	ρ_4	bromide
8)	ρ_3BROMISTØØØ#	ρ_4	bromide
9)	ρ_3BROMNØØ#	ρ_4	bromide
10)	ρ_3BROMNØØØ#	ρ_4	bromide
11)	ρ_4JELEZ	ρ_5	
12)	ρ_5O		grammatical information
13)	ρ_5A		" "

etc.

Entries 1) and 2) establish the valence of the metal. Entries 3) through 6) yield the proper translation of the metal in the proper order. Entries 7) through 10) give the proper translation to the acid radical. Entry 11) shifts out the metal without translation. Entries 12) plus provide the requisite grammatical information. This system also suggests that some additional inter-word grammatical processing can be done in the case of nominal phrases where the juxtaposition and agreement of adjective and noun solve some of the ambiguity inherent in the noun. A case in point here is the nominal phrase CETNOE#CISLO, where the nominal form CISLA may be either genitive singular or nominative and accusative plural. The following entries would be necessary to solve the ambiguities:

CETNOØ#CISL	N85
CETNOØØ#CISL	N85
CETNYØ#CISL	N89
CETNYØØ#CISL	N89

The first two entries are linked with a neuter singular only confix, the last two entries are linked with a confix directing the search to a neuter-plural plus neuter-instrumental-singular paradigm.

As is already evident, the processing of phrases in this manner is fraught with some difficulties. By clever manipulation grammatical information vital to further processing of phrases can be extracted but at the expense of multiple entries. Another complication is caused by the fact that many phrases, even highly idiomatic phrases are ambiguous. Even if such phrases are non-paradigmatic and require only one dictionary entry such as:

PRI#3TOM	in#this#case//at/with/before/in-time-of#this
CTO#KASAETS4	as#regards//which/what/that#touches
STALO#BYT6	therefore/began-to-be
V#DAL6NE1WEM	in#the#future//in/at/on#further

there arises the unpleasant prospect of creating new form classes or unusual combinations of form classes as well as unusual members of existing form classes in order that such phrases can be further processed in hopes of resolving the ambiguity. And, of course, if any one of the elements of such a phrase, especially if it were not located at either one of the extremities, had to be singled out for some particular syntactic processing, this goal either could not be realized at all or could be accomplished only through the agency of complex tags.

An even more serious argument vitiating the processing of phrases at this point is that most phrases may be discontinuous, albeit more or less unusually. Even an adjective-noun phrase can present some serious problems of discontinuity. For example, the nominal phrase 5AVELEVA4#KISLOTA (= oxalic acid) looks quite innocuous in this respect, but it was found embedded in the following prepositional phrase:

. . . SO 5AVELEV01 I NEKOTORYMI DRUGIMI ORGANICESKIMI KISLOTAMI...
.. . with oxalic and certain other organic acids...

where only some extensive and selective processing could isolate the phrase for idiomatic translation. And the instances are legion where the intrusion of a particle would destroy the usual continuity of a phrase as in the following:

CTO#JE#KASAETS4#NAWEGO#OPYTA...As regards our experiment, however...

NESMOTR4#DAJE#NA#PRIMENENIE ...even in spite of the use...

Several times already it has been stated that the real power of an MT system combining a high-capacity dictionary with table-lookup process resides particularly in the capability of processing efficiently the vast number of phrases absolutely essential to automatic language translation. Several preceding paragraphs have been devoted, perhaps unnecessarily, to proving that the counter-intuitive processing of phrases at the second level of specification, that concerned with inherent grammatical information, is just not feasible even though there might be a saving in processing time and the avoidance of some unwanted and complicating ambiguities. In other words, phrasal structures are essentially contextual problems and should be treated accordingly in the third, the last, and the most difficult level of specification, that dealing with contextual grammatical information. At this level the words of given source-language sentences, already partially specified by the identification of inherent grammatical information, are further modified and exactly specified, it is hoped, by matching them selectively against table entries representing all necessary, if this can be determined, specifications of the Russian grammatical and semantic relationships involved in the formula, S=NP+VP in order to obtain intelligible and accurate translations into English. The purpose of this paper is not to spell out all the features of so extensive a requirement, only to indicate how the general ingredients for the intended solution of this requirement are supplied by one approach to the problem. It is not out of place, however, to mention the general desirable features of any automatic translation system particularly from the linguistic standpoint: as deep a grammar of both source and target as possible; a vast memory for storage of all types of phrases aimed essentially at solution of the semantic problem; a processing routine that is maximally powerful, i.e., only weakly constrained; a recognition procedure that records a history of its activities and resolves its self-generated ambiguities, particularly those resulting from the weak constraints on the whole routine.

Clearly, then, this third specificational level is the culmination of the two preceding levels and is all-inclusive, embracing all that has been pre-generated in the areas of lexicography, grammar, and semantics. The principal point to be stressed here is the versatility of a capacious dictionary and a table-lookup method. A large dictionary in combination with exploitation of inflectional and derivational processes can confer a tremendous grammatical and semantic capability on a machine system. For example, the verbal pair VZVINCIAT6 /VZVINTIT6 (= to excite) can be matched in all forms by the stem entries:

VZVINT and VZVINC

and the verbal confix V41IVA (for recognition of all the imperfective forms). But now if this verb may be linked syntactically with the noun QEN (= price), it acquires the meaning "inflate". If VZVINT / VZVINC + QEN occurred in only one syntactic linkage, one entry with the proper grammatical and semantic data would suffice. But if these two lexical items could occur in several syntactic linkages, and they certainly can, viz:

VZVINCIVAET	QENY	inflates prices
VZVINCIVAHTS4	QENY	prices are inflated
VZVINCIVANIE	QEN	inflation of prices
VZVINCENNYE	QENY	inflated prices

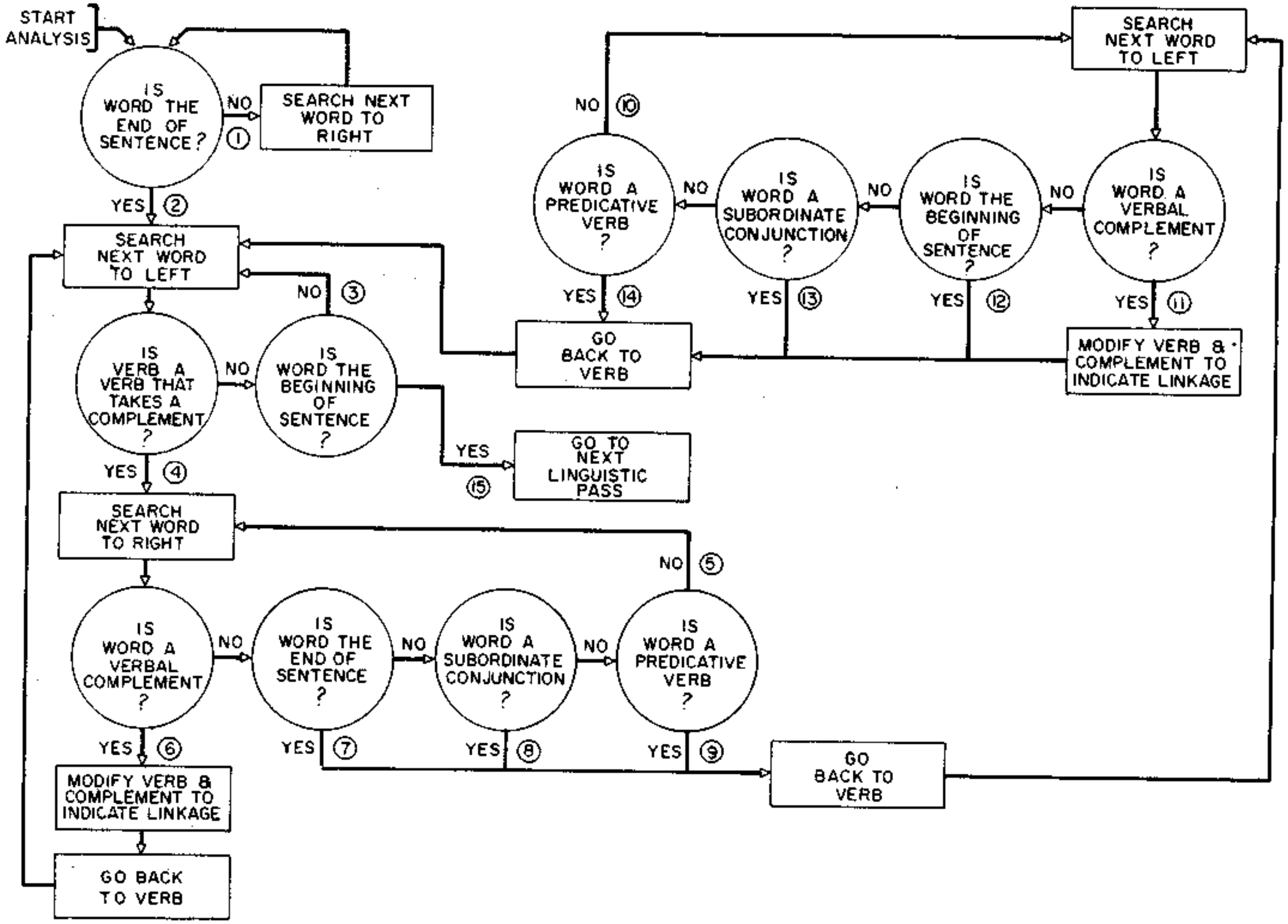
then one might create specific table entries for each of the syntactic linkages cited or one might make a table entry:

VZVINC QEN

that would apply to all the pertinent syntactic linkages and would modify only the semantic tags. In this way the proper kernel meanings of the words in terms of English could be indicated economically, and the relevant grammatical information would be utilized to modify the basic meanings as the result of a final lookup to supply all the English.

A very important part of the contextual level of specification is its implementation in terms of processing instructions for the table-lookup method. The table-lookup method has all the capabilities of a computer algorithm. This fact and some indication of the nature of processing instructions in the form of table entries is graphically shown on ***Chart No.4*** constructed from a series of fifteen table entries which govern the machine search for a predicative verbal and its complement. The table entries and the chart were created for illustrative purposes only. Even so, they would

CHART No.4 VERBS AND THEIR COMPLEMENTS


START ANALYSIS
IS WORD THE END OF SENTENCE?
NO ①
SEARCH NEXT WORD TO RIGHT
YES ②
SEARCH NEXT WORD TO LEFT
IS VERB A VERB THAT TAKES A COMPLEMENT ?
NO
IS WORD THE BEGINNING OF SENTENCE ?
NO ③
YES ⑮
GO TO NEXT LINGUISTIC PASS
YES ④
SEARCH NEXT WORD TO RIGHT
IS WORD A VERBAL COMPLEMENT ?
YES ⑥
MODIFY VERB & COMPLEMENT TO INDICATE LINKAGE
GO BACK TO VERB
NO
IS WORD THE END OF SENTENCE ?
YES ⑦
NO
IS WORD A SUBORDINATE CONJUNCTION ?
YES ⑧
NO
IS WORD A PREDICATIVE VERB ?
YES ⑨
NO ⑤
GO BACK TO VERB
SEARCH NEXT WORD TO LEFT
IS WORD A VERBAL COMPLEMENT ?
YES ⑪
MODIFY VERB & COMPLEMENT TO INDICATE LINKAGE
NO
IS WORD THE BEGINNING OF SENTENCE ?
YES ⑫
NO
IS WORD A SUBORDINATE CONJUNCTION ?
YES ⑬
NO
IS WORD A PREDICATIVE VERB ?
YES ⑭
NO ⑩
GO BACK TO VERB

serve their purpose for a tremendous number of Russian sentences, but they are not in their present form as powerful as they should be for the most effective routine of this type. It is obvious that they are designed for a multipass sentence recognition scheme. Because the chart is easily intelligible while the entries are not, the chart is presented first in its entirety and then only three of the entries in an abbreviated form.

The abbreviated entries are supplied for steps 2, 4, and 6 of the chart. Briefly, step 2 locates the end of the sentence and initiates a backward search. Step 4 locates the first predicative verbal encountered in a backwards search from the end of the sentence and starts the search forward for the verbal complement, and step 6 matches the predicative verbal with its complement. All three entries exhibit the standard form of table entries:

α (ARGUMENT) τ (FUNCTION)

where α = Begin Entry
τ = Argument - Function Boundary, Match Indicator

Step 2. $\alpha\ \rho_{\dot{a}}\ D_{eos}\ \tau\ \epsilon_1\ \rho_f$
where $\rho_{\dot{a}}$ = Prefixed Argument Modifier
D_{eos} = Argument Data (End of Sentence)
ϵ_1 = Index to Adjacent Word
ρ_f = Function Data Prefixed to Next Argument

Step 4. $\alpha\ \rho_a\ D_v\ \tau\ \epsilon_2\ \rho_f$
where D_v = Argument Data (Predicative Verbal)
ϵ_2 = Locate Succeeding Word Address

Step 6. $\alpha\ \rho_a\ D_v\ \epsilon_3\ D_c\ \tau\ \epsilon_1\ D'_v\ \epsilon_3\ D'_c\ \rho_f$
where ϵ_3 = Index to Succeeding Word Address
D_c = Argument Data (Verbal Complement)
D' = Function Data

The several entries cited together with their brief explanations should give some notion of how even machine instructions can be incorporated into a table-lookup procedure for automatic language translation. The chart prepared from the table entries should demonstrate in general terms the close and not surprising relationship between algorithms and table entries. Whether one processing method or the other is superior for automatic language processing remains to be seen and may depend solely on the hardware involved.

This paper has attempted to outline the necessary stages of source-language specification in terms of a thoroughgoing application of the table-lookup method for Russian-English machine translation. The third and vastly complex level, that pretending to establish source-language contextual grammatical and semantic relationships, could be discussed necessarily only in the broadest terms. It was possible only to sketch out present notions of handling a few general problems of context analysis. The degree of success of automatic context analysis is primarily a function of the depth developed in the source-language grammar and then of its clever implementation by sophisticated machine methods.