[*Prague Studies in Mathematical Linguistics* 1, 1966.]

MACHINE TRANSLATION IN PRAGUE

D. Konečná - P. Novák - P. Sgall

A group of linguists of the Charles University began to prepare machine translation of English technical texts into Czech in the autumn of 1957. A special Department was established for this purpose at the Philosophical Faculty in 1959. Later it was divided into two groups[1] working in close cooperation in machine translation, as well as in algebraic linguistics. The work of both groups is coordinated with that of the Research Institute of Mathematical Machines (RIMM), where the algorithms are being adapted for individual computers and programmed, and with the work of several Institutes of the Czechoslovak Academy of Sciences.

After the first experiment, very limited in scope,[2] had been performed on the Czechoslovak built computer SAPO at RIMM in January, 1960, another experiment was prepared, based on a running text of 40 English sentences taken from R. Shea, *Principles of Semiconductors.* The second experiment, too, was prepared in cooperation with RIMM, where it is presently being programmed for the Czechoslovak built computer EPOS I. The linguistic aspects of this second experiment are the main subject of this paper. Originally the tasks of the experiment had been considerably more extensive: it was supposed to include the translation from intermediate language into an information language calculus specialized in electronics, and putting and answering questions of a certain type concerning the data contained in the analyzed text.

The general features of this experiment can be described shortly as follows:

a) In the dictionary searching, a modification of the "direct method" has been used; the text has been coded in a special way, so that the code of a word itself gives the address either of the grammatical information concerning the given lexical unit (in case that no homonymy is present in the coded form), or of the group of words, which has become homonymous by the coding and which therefore has to be stored and identified in the overt form.[3]

b) The morphological analysis of the input text uses a simple table of regular English endings and the respective grammatical information; the irregular forms are, of course, identified in the vocabulary.

c) The analysis of idioms (phraseologisms) includes only examples actually met in the text, in an *ad hoc* manner.

d) In the syntactic analysis of English essentially the method described by Moloshnaya is used.

e) Some elements of an Intermediate Language (IL) were used. Further investiga tions are necessary, concerning the possibility of using such an IL in connection with other languages.

f) The synthesis of the output text in Czech has to be tested and completed in connection with the investigations concerning the generative grammar of the Czech language.

First, we shall specify the linguistic approach underlying our work. As formulated here, this approach has developed during the preparation of the experiment, and not all of its features have been fully applied in the algorithms.

The experience gained during this work led us above all to the conviction that theoretical work concerning machine translation has to include the study of algebraic linguistics, the central parts of which we see in the generative (and recognitive) description of language systems, where the mathematical specification of a language is connected with ascribing one or several structural characteristics to every sentence.[4] As to the concrete forms of this description of languages, the transformational grammar is known as the most adequate of the existing systems, even though some questions still remain open,[5] and we assume that further exploitation of the grammatical and semantic notions of classical linguistics can be made use of with advantage.[6]

Our view of the language system is based on the approach considering this system to be composed of levels, where the units of each two nearest levels are connected by the relation of "form-function" or "representation" (relation R). We distinguish, for the purposes of the present objectives of machine translation (i.e. with regard to printed texts) four such levels, denoted here by L_i.[7] L_1 *is* the graphemic level, with the alphabetic letter as the elementary unit, and the morph, a string of letters, as the complex unit. It may be useful, of course, to distinguish two different levels here, corresponding to phonetics and (morphophonemics, respectively. L_2 has the seme (e.g., nominative, future, feminine; or a stem, such as the stem of the English verb *go,* which is present in its preterital forms too) as the elementary unit[8] and "combinations" of semes as complex units. By a combination of semes we mean an element of the Cartesian product $S_1 \times S_2 \times \ldots \times S_n$ of some disjoint subsets of the set of semes, A morpheme is such a combination of semes, which is represented by morphs (e.g. the Russian morpheme Nom. Pl. Subst. Masc. is represented by *-i*, *-y* or *-a;* the lexical morphemes are considered here as combinations of one element only — elements of the Cartesian product $S_1 \times \ldots \times S_n$, where $n = 1$ — e.g. the English morpheme *go* is represented by the morphs *go* and *went);* another combination of semes is the formeme, which represents a unit of the next higher level, L_3 (e.g. the Russian formeme consisting of the semes *na* and Acc. represents the object in a construction as *smotrju na nego;* in most other cases, the formeme is a combination of one member).

L_3 corresponds partly to the syntax, as recognized by classical European linguistics. At this level, the sentence (as a complex unit) is composed of elementary units of three kinds:

a) semoglyphs,[9] which are represented by "stems" on L_2 (e.g. the several semoglyphs corresponding to the English Verb *go* — one of them corresponding to German *gehen,* another to *fahren* etc. — are represented by the stem, which, in its turn, is represented by one of two morphs — *go* or *went,* in L_1);

b) syntactic markers or sentence-parts as subject, object, adjunct, etc., represented in L_2 by formemes (e.g. the object is denoted by the combination of the seme of *na* with Acc. in Russian, if it is governed by a verb like *smotret');*

c) suffixes as "actual present", "gnom. pres.", "plur.", "indicative", represented, in L_2, by morphological semes.

Further, we use at this level auxiliary symbols, as ordinal indices, tectoglyphs (i.e. an ordinal index of the head word) and word classes.[10]

L_4 is intended to correspond to the semantic sentence structure[11] in a certain sense; instead of the syntactic markers we have here units as Agens (Actor) and Patiens, Action, Determinant; the units of other kinds (a, c) remain here the same as on the level L_3 (i.e. there is a one-to-one relation of representation between the units of the two levels), but the word-classes of this level are other than those in L_3. There the "syntactic derivation" was regarded as a relation between different word classes (in Russian, e.g. *rassmotrenije* is a noun in L_3, *rassmotrevšij* an adjective, *krasivo* an adverb, etc., whereas on the level L_4 the first two forms are classed as verbal forms, i.e. as grammatical forms of *rassmotret', krasivo* as a form of the adjective).[12] The distinction between active and passive, as well as between a nominal or participial construction and a clause is dropped in most cases on level L_4. Thus the pairs of sentences as (a) and (b), or (c) and (d), which can be regarded as paraphrases or synonymous sentences, have the same form on this level:

(a) *Mouton published Brown's book.*
(b) *Brown's book was published by Mouton.*
(c) *After having solved the first problem, he went on studying the second one.*
(d) *When he had solved the first problem, he went on ...*

We assume that the differences between natural languages decrease, as we proceed from the lower to the higher levels, i.e. from L_1 to L_4. Perhaps a level L_5 should be included into the system, which would correspond to (or include) the "topic-comment" (theme-rheme) articulation of sentences and phrases; it is probable, that on such a level the corresponding sentences of different languages would have the same form in most cases (as to scientific and technical texts). But such a system has to respect inter-sentence relations and other problems of great complexity, which are not to be managed under the present circumstances. As to European languages, the following hypothesis seems to deserve empirical and experimental testing: It is possible to proceed, in the analysis of the input text, from L_1 to L_2 (morphological

analyses) and then to L_3 and L_4 (syntactical analyses) and from the text of this L_4 to come over, in the synthesis, to the output text in L_1 i.e. in the graphemic level of another language; here, L_4 has to be the same for different languages and to function as an intermediate language. In many concrete constructions European languages differ, of course, on the level L_4 as well as on the lower levels; but — for the purposes of machine translation — there is a possibility of handling the "exceptional" constructions in a similar way as idioms are handled on lower levels.

The grammar of IL, based on this approach, is to be regarded as an individual language,[13] and not merely as a "net of correspondences" between languages. It can be specified by a generative grammar.[14] In the examples below, not all the elements of IL are used, and there are some auxiliary symbols not included in the more recent form of IL.

The sentence of IL can be regarded as composed, in a way, of semoglyphs. Every semoglyph, with the exception of that of the predicate, depends on some other semoglyph; there are several types of this relation of dependency: the depending member can be an actor (agens), a patiens, or a determinant of some other kind. (In the examples on p. 194, this dependency relation is denoted in the column V.) Further, there are two relations between the semoglyphs of a sentence: coordination (of several kinds: conjunction, disjunction, adversative relation etc.) and apposition; only two or more semoglyphs depending both (all) on the same member (or both on nothing) can be connected by one of these relations (denoted in the column III). Every semoglyph can be further joined and provided with several suffixes (cf. columns II, IV), corresponding mostly to morphological elements (endings, auxiliary words) in the natural languages.

The word of the text in IL remains as close to the one of the input text as possible, i.e. it is changed only in the cases where in the input text its position is conditioned by grammatical rules of the language and where the position of some word in this text therefore does not correspond to its position in the topic-comment dichotomy[15] (as far as this can be detected by the algorithm of the analysis). By algorithms analogous to ours the sentence (a) of p. 187 would be translated into Russian *Mouton izdal knigu Browna,* but (b) would become *Knigu Browna izdal Mouton* or *Kniga Browna izdana Moutonom,* i.e. the order of the lexical elements would remain the same; it is assumed, that in (a) the words *Brown's book* belong to the comment, but in (b) to the topic, and for such distribution the Russian equivalents (or their Czech parallels in case of our algorithm) are adequate.

The word-order in the IL does not depend on its grammar, but is determined by the order of the input text (or possibly, by the context criteria of the algorithm of the analysis). This enables us to handle the topic-comment dichotomy in the text without a complex system working with inter-sentence context. Of course, this approach is only a preliminary one.

Further knowledge of adequacy and economy of such an IL can be obtained only by further theoretical investigations and by processing large texts in different

languages. It is possible that this approach will not prove to be useful (or amendable). But even in that case, probably, such study of several systems of natural languages would be of interest for linguistics; the relations between the units and levels of different languages are investigated here empirically in a way useful for the comparison of the structures of different languages and for their typology. For an example of our method see Table I.

Explanations to the table: I — serial number, II — number, III — co-ordination, IV — kind of adverbial construction, V — semantic sentence part, VI — semantic word class, VII — tectoglyph, VIII — semoglyph; *D* — determinant, *Sb* — subject, *O* — object, *Atv* — complement, *S* — noun, *V* — verb, *A* — adjective, *T* — end of the sentence, *X'* — *X* pronominal.

At the beginning of analysis, a line with *n* places for adding certain characteristics is ascribed to each word form of the input sentence. In the course of the analysis individual places are filled in on the basis of the data given in the dictionary, older data are modified or several lines with indications for several word forms in the text are joined in one line.

One of the important parts of the analysis is represented by a block of syntactic analysis, in which, according to an organized list of so called configurations, the syntactic structure of the translated sentence is determined by a stage-by-stage reduction of a series of syntactic characteristics corresponding to a series of the word forms of the translated sentence.[16] Thus for instance the following reduction rules (with a special block for coordination) must be used in the analysis of sentences, being a part of the compound sentence, mentioned above:

i	*j*	*k*			
3	1		→	1_j	*j:i*
2b	2 + ed		→	$2pas_i$	
D	1		→	1_j	
MV	1		→	1_j	*j:i*
1_r	*F*	P_m	→	1_i	*i:k*
2-	*F*	1		$2—_i$	*i:k*
1	2-			$2—_j$	*j:i*
$2pas_i$	1			$2—_i$	*i :j*

Explanations: *i ,j, k* — ordinal indices of the word forms; The first line of the above Table means: If there occurs the sequence of word forms with the syntactic characteristics 3 1 in the sentence, the sequence is to be reduced into 1, while the ordinal index *j* is to be filled in the column of tectoglyphs (col. VII) of the word form with the ordinal index *i;* 1 — noun, 3 — adjective, 2 — finite verb, 2b — be, 2+ed — past participle of transitive verb, *J* — subordinating conjunction, *D* — article, *MV* — mathematical expression, *E* — end, *E/Č* — conjunction expression or a comma, *F* — preposition, P_m — personal pronoun in accusative, 2 — intransitive verb, l_r — noun with rection.

The sequence of syntactic characteristics of the word forms of the English sentence given in Table I is as follows:

J 3 *E* 3 1 2 - *F D* 3 1 *E/Č* 1 *F* P_m 2b 2+ed *D MV* 1/3 *T*

(The fragmental dash divides possible alternatives.)

*

Even the best existing descriptions of Czech syntax (as e.g. Šmilauer's *Novočeská skladba),* giving for every sentence part the list of all possible forms, do not usually state the conditions under which the given form is preferred to a synonymous form, or when the given form is the only correct one (grammatically correct). The investigation of those conditions has become necessary in connection with the problem of machine translation (esp. of MT using an Intermediate Language) and generative grammar of individual languages; the restriction of those conditions has its general importance for an exact description of the language as well.

When the algorithm of syntactic synthesis for the experiments on the computer EPOS was under construction, such an investigation was just beginning. An assumption was taken that given a list of conditions under which various forms to express the given meaning may be used, we may try to determine "the basic form", i.e. such a form the conditions for the use of which cannot be determined while for other synonymous forms the conditions can be stated.

Let us give an example. We have to determine the form expressing the necessity of an action, the agent and object of which can be determined. We may select the basic form from the three forms: l) dependent clause, 2) infinitive, 3) noun; we know (e.g. from Šmilauer's Syntax) that these are the forms of subject depending on the predicate of the type *je nutné 'it is necessary';* in the case of another form of predicate, we should consider another set of forms expressing the action.

Svoboda states[17] that the form infinitive is not correct in the case when the agent is expressed or when the object is not expressed.

Such indicative sentences as:

(a) *Je nutné opravit úlohu od učitele.*
(b) *Je nutné opravit*

may even be considered non-grammatical; grammatical sentences would be in this case

(a) *Je nutné, aby učitel opravil úkol 'The teacher should correct the homework'*
(b) *Je nutná oprava 'The correction is necessary'*

Thus we know a condition for the forms (1) and (3); the algorithm may contain rules ensuring that form (l) will be used when the agent of action is expressed, form (3) where the object of action is not expressed, and in other cases, form (2) — infinitive — will be used, which was chosen as a basic form.

The above solution implies that in the case where the agent is expressed, the object will be expressed as well; sentences as: *Je nutné, aby učitel opravil* cannot exist. If we abandon this assumption and want to select such a form which may be used in all possible combinations of expressing and not-expressing the agent and object (here the form (3) would be convenient — see *Je nutná oprava úlohy učitelem, Je nutná oprava úlohy, Je nutná oprava učitele, Je nutná oprava)* and if we put into the algorithm only the form chosen in this way, it would lead to unacceptable "standardization" of the output text. Three forms would be reduced to a "universal" one, while the reduction of several possible forms to a universal form (acceptable under all conditions) is possible only in some instances. Let us quote now a description of a sentence *"Jestliže se oblasti typu n a typu p vyskytují v témže krystalu, hranice mezi nimi se nazývá přechodem p-n"* in L_3 and L_2 (see Table II where only the data necessary for the application of the given rules are stated). The rules for creating L_3 and L_2 (only those, the use of which is necessary for the given sentence) are the following:

The rules for generating L_3

1. ØP → *pv*
2. *DVcon* → *pj*
3. *OS* → *sbs*
4. *X* → *x*
5. -X' → *x*

The rules for generating L_2

1. $m_1 ds \rightarrow S_{v6}$
2. $m_6 ds \rightarrow S_{mezl\ 7}$
3. $ds \rightarrow S_2$
4. $as \rightarrow S_7$
5. $sbs \rightarrow S$
6. $sbs + px \rightarrow S_i + V_i$
7. $xs + da \rightarrow S_j + A_j$

Note: 1. Column *b,* containing tectoglyphs and some semantic data, is a part both of L_3 and L_4. Small letters in L_3 are analogous to capital in L_4 (with the exception of *p* and *j* — see later). The data giving the form of the stem and the gender of the noun stem are filled in from the dictionary before using the rules forming L_2 (m = masc., f = fem.). The unmarked morphological categories (3rd person, sg., ind., pres., act., neut., nom., sg., neutr., posit.) are not recorded.

2. *x* — indicates any sentence part and any word class; *p* — predicate; *j* — definite form of verb with the conjunction *jestliže 'if';* in L_2 : *S* — subst., *V*— verb., *A* — adj., *Z* — pronoun *(S* may be replaced by *Z* according to the dictionary). The rules 6 and 7 see later.

The algorithm of syntactic synthesis for the experiment on EPOS starts with the word with a null tectoglyph. This word will be the predicate of the main clause (see the rule L_3 1). Then the forms for expressing other actions are considered. In some cases, only the universal form is reached (rule L_3 2 — each action, which is a condition for another action, will be expressed by a dependent clause with *jestliže* [if]), in some cases, the basic form is selected *(je nutné změnit proud* — rules see

above). When stating the form of the action (in the case when it has been determined that it will be expressed by the predicate of the dependent clause), one of the following forms must be chosen: active, passive and reflexive. In the combination "the agent is not expressed while the object is expressed" OS → *sbs* (L_3 3), the object becomes the subject of the sentence; the reflexive form is the basic form, the active form is conditioned by the expression of agent, while the passive form is conditioned by the verbs, for which the use such a form is prescribed by the dictionary.

There are several rules, by application of which the semantic sentence parts are changed into syntactic ones: the rule *(S) OS → ds* and *(S) SbS → ds* (*S* in brackets = if depends on *S)*, the pair of rules *(S) DAd → da, (S) DA → dad,* that is rules, expressing the traits of formal syntax of Czech as we know them e.g. from Šmilauer's syntax (in those cases we have not a more precise description) and finally the rule *X → x,* which includes the cases in which semantic classification corresponds to the syntactic one (which concerns, e.g., determining circumstances *(D Ad),* corresponding to the adverbial construction expressed by an adverb *(d ad)* in Czech). Further on, the algorithm contains rules ascribing to the formal sentence parts the word classes, the data for morphological synthesis, which can be gained by examining which formal sentence parts the given word expresses (rule L_2 1—5) and the well-known formal syntactic rules of grammatical concord (L_2 6 and 7) and rection, the latter being obtained from the dictionary (taking account of the dictionary, we also correct the indication about word class— see line *8,* column *e).*

In the course of preparations of the experiment on the computer EPOS, the basic experience was to be gained concerning the arrangement of the rules and the conditions for using a given form for a given meaning.

Notes

[1] One of them is the linguistic group of the Centre of Numerical Mathematics, Faculty of Mathematics and Physics, the other the Section of Algebraic Linguistics and Machine Translation at the Department of Linguistics and Phonetics, Philosophical Faculty.

[2] K. Korvasova, *Note of Experiment of Mechanical Translation of the Computer SAPO,* Information Processing Machines 8, ČSAV, Praha 1962, B. Palek-P. Pit'ha- P. Sgall, *Mathematical Linguistics in Czechoslovakia,* Prague Bulletin of Mathematical Linguistics 1, 1964.

[3] K. Korvasová- B. Pale k, *A Problem of Czech Coding,* Information Processing Machines 10, ČSAV, Praha 1964.

[4] N. Chomsky, *Syntactic Structures.* The Hague 1964; *On the Notion "Rule of Grammar",* Structure of Language and its Mathematical Aspects, PSAM 12, Providence 1961, 6—24; *Logical Foundations of Linguistic Theory,* Proceedings of the Ninth. International Congress of Linguists, The Hague 1964; Y. Bar-Hillel, *Some Recent Results in Theoretical Linguistics, in* Logic, Methodology and Philosophy of Science (E. Nagel, P. Suppes, A. Tarski eds.), Stanford 1962, 551 — 557, *Four Lectures on Algebraic Linguistics and Machine Translation,* Jerusalem 1963.

[5] H. Putnam. *Some Issues in the Theory of Grammar,* Structure of Language and its Mathe-

matical Aspects, PSAM 12, Providence 1962, 39 ff.; H. B. Curry, Some Logical Aspects of *Grammatical Structure,* ibid., 65 ff.

[6] Cf. P. Sgall, *'Generative Beschreibung und die Ebenen des Sprachsystems* (paper read at the Symposium in Magdeburg, 1964).

[7] Cf. P. Sgall, *Zur Frage der Ebenen im Sprachsystem,* Travaux linguistiques de Prague I, Praha 1964.

[8] V. Skalička, *Zur ungarischen Grammatik,* Praha 1935, 13.

[9] N. D. Andrejev, *Mašinnyj perevod i problema jazyka-posrednika,* Voprosy jazykoznanija 6, 1957,5, 117ff.

[10] Cf., for the Czech language, V. Šmilauer, *Novočeska skladba.* Praha 1947.

[11] M. Dokulil - F. Daneš, *K tzv. významové a mluvnické stavbé věty,* in O vědeckém poznáni soudobých jazyků, Praha 1958, 231 ff.; F. Daneš, *A Three-Level Approach to Syntax.* Travaux linguistiques de Prague I, Praha 1964.

[12] J. Kuryłowicz, *Esquisses linguistiques,* Wrocław— Kraków 1960.

[13] N. D. Andrejev-S. J. Fitialov, *Jazyk-posrednik mašinnogo perevoda i principy jego sostavlenija,* in Tezisy soveščenija po matematičeskoj lingvistike, 15—21 aprelja 1959 g., Leningrad 1959, 53 ff.

[14] P. Sgall, *Prevodni jazyk a teorie gramatiky,* SaS 24, 1963; a summary in English with examples, in Computational linguistics 2, Budapest 1964.

[15] V. Mathesius, *Čeština a obecný jazykozpyt,* Praha 1947; J. Firbas, *K vyjadřovani akutálniho členěni v češtině,* in O vědeckém poznáni soudobých jazyků, Praha 1958, 250 ff.

[16] Recently, our group has been engaged in a modified type of predictive analysis.

[17] Cf. K. Svoboda, *Infinitiv v současné spisovné češtině,* Praha 1962, 5 and 8.