THE PRAGUE BULLETIN OF MATHEMATICAL LINGUISTICS 35, 1981

# Automatic Translation from English to Czech

Eva Hajičová and Zdeněk Kirschner

After the first and limited experiments with the English-to-Czech translation from the beginning of the 1960's, a broadly conceived and detailed system of automatic synthesis of Czech sentences has been formulated. The work on an analysis of English has been resumed by the group of algebraic linguistics at the faculty of mathematics and physics, Charles University, in 1976. Thanks to a very effective cooperation with Canadian linguists and computer scientists (R. Kittredge, B. Thouin) it was possible to prepare three experiments, the first of which was accomplished in the years 1978-1979. Although with each of the experiments partially different aims have been pursued, almost the same general strategies and similar tactical devices have been adhered to.

The main difference between our approach to automatic analysis and that of the Canadian system consists in that we work with dependency grammar instead of that of immediate constituents. We introduce labelling the edges in the graphs, representing the structures analyzed, indicating thus the direction of branching on the one hand and assigning the functions to the dependent (dominated) sentence elements on the other. The automatic analysis of English elaborated in Prague uses Colmerauer's Q-systems; in several respects the analysis is brought to the tectogrammatical level at which semantic functions are represented to the measure that corresponds to further orientation of the particular experiment. Especially the prepositional phrases are analyzed not only syntactically, but also semantically, i.e. their function as an adverbial of a given kind (instrument, manner, place, direction, time, purpose etc.) is

identified.

In the first of the three experiments only relatively simple English input sentences (over a limited dictionary) were analyzed, yielding a representation of these sentences suitable as an input for the synthesis of the corresponding set of Czech sentences - which will accomplish the process of their translation. The programme of the analysis is divided into twelve steps which can be grouped into four main blocks: (i) premorphological processing of the input, (ii) morphological analysis, (iii) syntactical analysis, and (iv) transfer (the first part of which is implemented in Q-language, the second part in PL/1).

Morphological analysis is rather restricted in our first experiment; there was no need to test a relatively extensive procedure that had been already brought to perfection in the Canadian experiments and that, if necessary, could be taken over as a whole (which actually happened in. the second experiment). At the output of this block all elements (if, of course, they had been identified) appear as trees in what may be called canonical form; in case they underwent morphological treatment, they are followed by standardized forms of the endings that had been separated in the course of this treatment (e.g. S for plural of nouns or third person singular of verbs, ED for past tense, EN for past participle, etc.). In the subsequent stage, such pairs are interpreted: e.g. the rewriting rule N(A* (U*)) + S → N(A* (U*, *PL)) represents an instruction for the interpretation of the plural ending -s. Here, the N and S are constants representing the noun category and the ending -s respectively, A*, U* are variables (variables are marked by an asterisk following a letter; letters from the beginning of the alphabet - A to F - denote variables that stand for values (labels, etiquettes), variables represented by letters I to N can stand for trees, and letters U to Z are used in the representation of variables standing for lists of trees. A list can be empty or contain one tree only; the simplest case of a tree is a value).

Here, the variable A* stands for the lexical value or the unit in question, the variable U* denotes the list or semantic features assigned to it. The constant *PL denotes plural (semantic features, some grammatical information, etc. are usually marked by an asterisk preceding a letter or a group of letters, mostly mnemonic symbols, abbreviations and others like that).

Syntactic analysis is divided into several steps: first, nominal complexes are built up (by putting together nouns with their attributes standing to their left or right, and by putting together prepositional phrases); then verbal complex forms are identified and constituted. The verbal complements are attached to the verb form one by one: this is being done by filling in the "slots" contained in the dictionary interpretations of verbs and representing their frames. As a rule, the subtree governed by the numeral "1" indicates the properties a subject of the verb must possess, that dominated by "2" those of its direct object, etc. The last step attaches free modifications (local, temporal causal, etc.). The following two rules may serve as an example of a saturation of such verbal frame:

(1) N(A*(U*),V*) + AUX(B*(X*,T(C*),Y*),Z*)
V(B*(X*,T(C*),Y*), N(A*)(L(SUB),U*) V*),Z*).

(2) V(U*) + PP(BY MEANS OF,N(A*(X*),Y*))
V(U*,N(A*(R(ADV,*MNS),X*),Y*)).

They are applied to the string

(3) N(CHANGE(*A,*PL,*DEF)) + AUX(MONITOR(T(PRS),MDL(CAN),
*NEG,*PSV,1(*H,*C,*A),2(*A),*DUR)) + PP(BYMEANSOF,
N(MEASURE(*A,*INDEF),AD(OR(L(φATR),*STO,*SFTO),AD(SINGLE
(L(φQ))),AD(SIMPLE(R(φQ), *STO,*SFTO))))).

In this example the rules and the string were simplified. A stands for adjective, AUX denotes a verbal complex with "slots" for actants not yet filled, V a verbal complex in which the "slots" have" already been saturated, PP is a prepositional phrase; T stands for tense, PRS for present tense, MDL for modality, *NEG for negation, *PSV for passive; semantic features *A, *C, *H stand for the features "abstract", "concrete",

"human", respectively. The subtrees governed by L and R are labels on the edges: they specify the position (L - left, R - right) and the function of the given part of the sentences: the functions are marked by the sign "φ". In the above rules, L(φSUB) and R(φADV, φMNS) denote the subject standing to the left and the adverbial standing to the right of the verb respectively; *MNS stands for the feature "means", the function φQ denotes members of a coordination series. Informally, the above rules state that (in English) if a noun is followed by a verbal complex with unsaturated participants, this noun is inserted into the "slot" of subject of the verb in question; if a verbal complex with participants saturated is followed by a prepositional phrase with the preposition by means of, the head noun of this phrase is inserted into the complex as an adverbial of manner, specified by the feature "means".

The tree in (3) is transformed by the application of the rules (1) and (2) into the tree (4), which represents the structure resulting from the syntactic processing of the English input sentence "The changes cannot be monitored by a single of simple measure":

(4) S(V(MONITOR(T(PRS),MDL(CAN),*NEG,*PSV,xDUR),N(CHANGE (L(φSUB),*A,*PL,*DEF)),N(MEASURE(R(φADV,*MNS), *A, *INDEF),AD(OR(L(φATR),*STO,*SFTO),AD(SINGLE(L(φQ))), AD(SIMPLE(R(φQ),*STO,*SFTO))))))).

The root S serves only for the formal purpose of indicating that the analysis is finished; the representation of the complete structure of the sentence proper starts with the node V (representing the governing verb).

The so-called transfer articulates the resulting tree into particular subtrees governed by such categories as N, V, AD, etc., changes their order (placing the dependent elements in front of the governed ones), replaces the original symbols denoting grammatemes by indices used in the synthesis of Czech and substitutes for the English lexical units the corresponding Czech ones. The last part of the transfer, which adapts the output of the Q-language

programme to the notation and organization of the input for the programme of synthesis is written in PL/1.

Several modifications of the system are incorporated in the second experiment, which is now being prepared, and which has in view practical application in an automatic translation system for the INSPEC tape service (confined in the experiment to abstracts from the field of the production and application of integrated circuits). The overall strategy of the procedure is very similar to that used in the first experiment; however, several changes and improvements have been introduced, the more important of which concern the following points:

(i) the programme contains a full system of morphological analysis of English word forms (taken over, in the main, from the first Canadian project TAUM 1973, with kind approval of the authors of the system); thus, almost all irregular, anomalous, or rare forms can be analyzed and the normal (dictionary) forms reconstructed;

(ii) to reduce the scope of the main dictionary, a new component has been added: the so-called translational dictionary, the rules of which translate the most frequent classes of terms of international usage directly into Czech, mostly by changing suffixes and executing orthographical changes - e.g. APPLICATION into APLIKACE, PHILOSOPHY into FILOZOFIE, AMPLIFIER into AMPLIFIKÁTOR, OPERATIONAL into OPERAČNÍ, etc. Such words, provided that their grammatical and semantic properties are not idiosyncratic, need not be included in the main dictionary, which always has the unfortunate tendency to grow beyond measure; at the same time, it is a way how to deal with some words that could not be identified in the previous stage;

(iii) particular attention is paid to the syntactic analysis of nominal complexes in general and compounds in particular, especially with regard to the problems of conversion; a repertoire of semantic features gradually built on what can be called a highly schematic model of the universe of discourse helps to formulate some rules that

cover the regular or at least the most frequent phenomena in this domain. The corresponding section of TAUM grammar as well as that of our first experiment were very limited and simple; the texts analyzed by the second experiment grammar - abstracts from the field of electronics - are based on technical terms, abound in them, combine them in various ways, modify them, etc., and that is why they cannot be treated without an adequate apparatus of rules that solve the current problems leaving to the "lexicalist" solution as little as possible;

(iv) since the texts are rich in coordination structures, more sets of rules analyzing different types of conjunction both on the phrasal and sentential levels were included to operate at different stages of the process;

(v) wherever possible, Czech equivalents replace the English lexical values already at the initial stages of the analysis; sets of indices required for the synthesis of Czech are supplied already in the course of dictionary operations; the English semantic features as well as the "slots" in verbal frames are always deleted as soon as they have fulfilled their task;

(vi) the system gives preference to more general solutions whenever possible; e.g. rules are supplied that reconstruct elements deleted in the surface structure of sentences but necessary for the semantic interpretation of the sentence (and thus also for the proper choice of its equivalent construction in Czech);

(vii) more general or universal solutions are also preferred to meet another problem connected with the fact that English, owing to its rather poor morphology and to the almost complete lack of the means of determining referential relationships that is called grammatical concord, is a language more vague than Czech; extralinguistic knowledge plays a more important role in English than in Czech, where the elements bound together by referential relationship must agree in case, gender, number, and, with verbs, in person. A computer for which such an "extra-

linguistic knowledge" is, under present conditions, practically unattainable, will face difficulties if, e.g. it is to decide to which nominal complex the verbal attribute "using..." belongs in such a sentence as "These methods employ a Monte Carlo analysis in the parameter space using a simplicial approximation to the region of acceptability..." A layman can exclude the "space" as an agentive with the aid of the same means as the computer, which can also be endowed with the knowledge that under normal circumstances, "space" cannot "use" anything, but as for the other two candidates - "methods" and "analysis" - the decision will be difficult without at least some idea of what "Monte Carlo analysis" and "simplicial approximation" are. An experimental system working in laboratory conditions can afford to register all ambiguities and regard a multiple solution as a success - the more so, as the computer is often able to detect ambiguities where man fails to become aware of them; however a practical system must seek a way out and accept such a solution only in case of inevitability, and, in fact, regard it as a failure. The only chance is to make the output as linguistically ambiguous as is the original utterance in the source language and leave the decision concerning the correct interpretation to the reader. In quite a number of cases, such solution is possible: here, e.g., it can be done by translating the transgressive as "with using", in Czech "s použitím";

(viii) last but not least, a system that aspires to be applied in practice must confront the fact that time from time it can come across a phenomenon that cannot be handled by the means that stand at its disposal. As has been already pointed out (see the above paragraph (ii)), the simplest and probably the most frequent case will be a word that cannot be identified by any of the dictionary operations. For this case, a "universal" noun is prepared provided with a "universal" set of semantic features and indices and retaining its original lexical value. More difficult problems are to be expected with syntactic and other anomalies; e.g., a noun will be used figuratively so that the intersection of

the set of its semantic features and the set of the features required in the "slot" of the verbal frame remains empty; in such a case, the noun fails to become integrated in the verbal complex and the construction of the sentence tree cannot be accomplished. Here, a special stage - a special Q-system - is designed to solve the most frequent failures: a sort of "emergency" grammar. It goes without saying that the particulars of such a "rescue-device" can be drawn only after sufficient experience with the whole system in experimental operation has been accumulated.

The third experiment concerns automatic analysis of English sentences serving as the source of information for the system based on the method called TIBAQ (described by P. Sgall, in press). It is closely connected with the second experiment. The front end of the system TIBAQ then will be either a Czech text or an English text; the output structures of the analysis of English sentences are complemented and augmented in order to achieve tectogrammatical representations of the English input sentences that might supply the material for compiling the knowledge representation and serve as a base for the operation of inference rules in the same way as the Czech sentences at the input.

The synthesis of Czech output sentences in the first two experiments, as well as that of Czech answers in the experiment of question answering based on the method TIBAQ (i.e. "Text-and-Inference Based Answering of Questions"), follows closely the framework elaborated for the purpose of random generation of Czech sentences (see esp. Panevova", 1979). The procedure of transfer, mentioned above (the first variant of which was prepared by S. Machová), yields representations of sentences more or less identical with the output strings of the transducer III of the random generation. This means that the analysis of the English sentences does not go as far as to the tectogrammatical structure. As we have already mentioned, e.g. the different meanings of prepositional phrases are identified (in the scope necessary for the given experiments); on the other

hand, the syntactic roles of individual phrases are characterized only as far as the surface structure is concerned. It would be superfluous e.g. to reconstruct every nominalized clause in the English input sentences, since in most cases the nominalizations in Czech are similar to those in English (i.e. there are close Czech equivalents for before/ /after John's arrival, due to the richness of John, in spite of John having come late, etc., etc., though it is not always easy to find more or less exact semantic counterpart, cf. point (vii) above).

Thus only within the TIBAQ experiment the tectogrammatical level (with such syntactic units as the participants: Actor, Patient, Addressee, etc.) is to be achieved. In the machine translation system the synthesis of Czech sentences starts with the procedure corresponding to Transducer III, which takes surface syntactic representations as its input string. These representations are transferred by Transducers III and IV to the morphemic level; the meanings of place, time, cause etc. are changed here into the morphemic forms realizing them (prepositional phrases etc.); furthermore, the morphemic units of tense, aspect, gender, number etc. are chosen here. Then the rules of morphemic synthesis translate these morphemic representations into sequence of Czech word forms (with case inflections, personal endings etc.) corresponding to grammatical sentences; at last, the graphemic shape of a sentence is achieved, which expresses the meaning that was represented by the given input string of the transductive components.

The sequence of computer programmes performing this transduction (written in PL-1) is based on the formal pattern of pushdown transducers. The main programme of each step is constructed on the basis of the defining function of such a transducer, where by means of a single passing through the given representation of a sentence the changes necessary for the trandsduction to the next lower level are ensured, while every dependent word (rectum) is confronted here with its governing word (regens). The results of such

a confrontation (first of all, modifications of the dependent word according to relevant properties of its governor) are given in the form of large tables, represented as subroutines of the main programme, activated any time the two members of a syntactic pair are confronted, one of them being read at the input of the transducer and the other being at the accessible end of the pushdown store. The large size of the tables makes it necessary to have specific subroutines (a) for the identification of the types of word forms figuring as names of rows and (b) as names of columns and(c) for the identification of the result found in the table (at the intersection of the given row and column), i.e. the value of the function for the given values of its two arguments. These operations represent the most extensive part of the functioning of the computer (which also is connected with the largest requirements on its memory). The formal mechanism used is based on rather strong linguistic hypotheses:

(1) For the choice of means (i.e. units of the next lower level) realizing or expressing a given functional unit (of the higher level), it is enough to confront a pair of complex symbols (word forms) connected by syntactic dependency (with the exception of individual more complex constructions); the dependent item changes its characteristics according to relevant properties of its governing item.

(2) In the course of the operation of a transducer, it is sufficient to pass the representation of the sentence only once.

(3) When the dependent item is being transduced to the next lower level, its governing item has already been transduced to this level.

It has been possible to construct algorithms for the synthesis of Czech sentences, respecting the quoted assumptions. The formal shape of the rules makes it possible to interpret the rules as rendering the relationship of "function" (meaning) and "form" (means), well known from

the Prague school of structural linguistics; also such notions as primary and secondary functions of a form can be studied in a more explicit way when using such a formalism.

The procedure of morphemic synthesis has been described by Weisheitelová (1979); it has the formal shape of a finite transducer and has been also implemented in PL-1.

The whole system of programmes is rather complex; it certainly can be reshaped to work more quickly and to need less storage capacity than in the case of the first variant of the system. During the experiments and the first period of application it will also be possible to detect and remedy the lack of rules accounting for grammatical phenomena which were not accounted for in the first version.

Processing of texts at large scale will bring new empirical data for linguistic research and will make it possible to handle many of the open questions of the procedure of translation in a new way, since the automatic system can serve not only as a means of application of the results of linguistic analysis, but also as a useful aid in studying various empirical phenomena and correspondences between the two languages, their frequency in the given type of texts, etc. These new conditions will bring about a new relationship between linguistic theory and its applications in the domain of translation.

References:

Panevová J. /1979/, From Tectogrammatics to Morphemics. In: Explizite Beschreibung der Sprache und Automatische Textbearbeitung IV, Prague, 3-166.

Sgall P. /in press/, Towards a Fully Automatic System of Communication with Data Bases. In: Proceedings of Int. Conference on AI and Information-Control Systems of Robots, Smolenice 1980.

Weisheitelová J. /1979/, Morphemic Synthesis. In: Explizite Beschreibung der Sprache und Automatische Textbearbeitung V, Prague, 3-68.