

On a Dependency Analysis of English for Automatic Translation

Zdeněk Kirschner

If we disregard the often used classification of machine translation projects in terms of generations, and if we avoid the somewhat vague term "high quality translation", we can simply divide the field of automatic translation experiments into four relatively distinct domains:

- approaches designing computer aids in translation
- translation essentially based on one-to-one substitution of elements (words, syntagms, phrases, etc.) and/or restriction and preliminary adaptation of input texts
- translation based exclusively on the means of linguistic (morphemic, grammatical and partial semantic) analysis and synthesis
- translation based on what may be called automatic understanding of natural language.

The relativity of the differences consists in that the four approaches overlap - each, as a rule, using the means of its neighbours as exceptional and marginal or as auxiliary procedures - and that, in a sense, the second and, to a certain measure, even the third approach can be referred to as computer aids, too, since their output must be revised and given final shape by human translators. Thus among other things the types of approaches differ in the degree of necessity of human participation at various stages of the actual processing; in fact only the products of the latter two can rank among true automation projects.

The fourth approach differs from the rest in that it

draws full consequences from the fact that any understanding and, accordingly, any exacting processing of natural language is unthinkable without utilizing extralinguistic knowledge. It starts with a full-fledged linguistic analysis and takes recourse to linguistic means at other stages of the processing as well, but it must have a near-to-perfect model of the relevant part of the universe of discourse at its disposal, which, of course, represents a serious complication on the way from devices aiding the human translator to fully automatic systems independent of human intervention.

Our experimentation described below belongs to the third type of approach; however, it does not leave out of account the possibility of using its results and experience in preparing more ambitious experiments that can be classed with the fourth category - the sophisticated domain of the so-called automatic understanding of natural language, and it constantly pays due attention to theoretical aspects. Some concrete and practical assignments accepted in connection with, and within, our projects do not contradict or disqualify the essential character of our experimentation, which represents a long-term effort to study the problems of automatic processing of natural language in general and those of machine translation in particular, and to do it both in terms of theoretical, especially contrastive, examination of languages, and in terms of applied linguistics. At the same time our experiments must be regarded as a part of the general preparation not only for solving the tasks mentioned above, but for dealing with a number of analogous objectives, e.g., Czech to English translation, the automatic treatment of other pairs of languages, etc.

Up to this time, three experiments (APAC1, APAC2, APAC3) have been conducted by the linguistic group of the

Department of Applied Mathematics at the Faculty of Mathematics and Physics of Charles University in Prague. The first experiment was carried out in close cooperation with the Montreal University team, which under the guidance of professor R. Kittredge prepared a series of experiments (TAUM, 1973; Kittredge, Bourbeau and Isabelle, 1976), culminating in two translation projects, at least one of which since that time has been successfully operating in practice (the METEO system that translates meteorological reports from English into French). The analysis in our experiment was based on a relatively limited grammar designed by R. Kittredge, and extended by the author; it resembled the TAUM analysis of English, but the dependency structure realized with the aid of special subtrees (substituting labels on some edges of the dependency graph), signified an important difference. The output of the analysis was adapted to the structure and to the notation accepted by the third automaton of the transducing components of the programme of random generation of Czech sentences (see Panevová, 1979), by means of a special intermediate programme called transfer. With the generative (synthetic) procedures attached, our combined programme translated several sentences taken from a journalistic text on economics into reasonably good Czech.

The analysis programme, of course, was able to parse much more, and it actually analysed a great number of English sentences of several basic types, mostly variations based on the original sentences referred to above. It contained a very limited dictionary, and a simple morphemic analysis allowed to interpret a few regular inflectional items. The rules covered the basic modifications of nouns and regular creation of verbal forms with tense and modality included. Nominal composition was represented by one rule only, transforming the compound " $N_1 N_2$ " into the " N_2

of N_1 " structure. A set of rules handled most of the elementary syntactic structures based on filling the slots of verbal frames. Only one type of dependent clauses was solved in the rules (concessive clauses), in addition to some similar more or less idiomatic structures of minor importance, which was only natural for the first experiment.

The second experiment differed from the first one in several respects, the most important of which being what may be called purposeful semantic circumscription. The journalistic style and the publicist approach to a rather general field of economics in the input text processed in the first experiment offered minimum chance for defining the subset of language and reduced the possibility of constructing an adequate semantic analysis apparatus to little more than zero. Therefore, in the second experiment, the specific genre of abstracts and a highly specialized domain of microelectronics were chosen to serve as the main input material; at the same time, the possibility of extending or generalizing the grammar to apply to scientific and technological writings in the field in question was kept in view. Another difference consisted in a more objective approach: the construction of rules was based not only on a number of selected, sufficiently representative abstracts, but also on a more profound examination of the structures analyzed, which was supported by extensive excerption. The third difference can be seen in the stress laid on the reduction of dictionaries: while in the first experiment two parallel dictionaries were operative (the English-Czech dictionary formed a part of the transfer), and even some morphemic problems were solved at the expense of the dictionary, in the second experiment a complete morphemic analysis of English was employed, a set of rules covered the most frequent phenomena in English nominal composition, and a special device transduced the words of international

usage directly into Czech. The rest of the differences mostly concerned the scope of the grammar which, to characterize it in a general way, covered a greater number of relatively frequent phenomena in comparison with its predecessor; however, one more difference is worth special mention. The second experiment was not confined to testing the theoretical framework and the possibilities of attaching the transducing components of the random generation programme only, but it was additionally assigned to more concrete tasks: the analysis should serve the purposes of machine translation of the abstracts taken from tape-service materials, on the one hand, and potentially act as the front end of a natural language understanding system called TIBAQ (Text-and-Inference Based Answering of Questions - a project being prepared in the linguistic group) on the other.

The grammar, of course, did not cover the subset of the language in full. Some limitation had to be adopted to make the effort more purposeful and economical. The incompleteness of our grammar in some respects might be attributed to the fact that our attention concentrated on the most frequent structures and that the solution of principles was preferred to the solution of more or less marginal phenomena. The experiment could be regarded as an intermediate stage in a process advancing from a more general to a special orientation, from a more theoretical approach to an attempt at immediate practical applications. Therefore, it was brought close to linguistic meaning (tectogrammatical level), on the one hand, and at the same time, adapted to the target language structures as well as to some practical requirements of the implementation, on the other. In principle, continuation in both directions was possible: either stressing the theoretical aspects and strictly observing the principles of an objective repre-

sentation in terms of basic grammatical and semantic categories, irrespective of undecidable parallel outputs, of time and storage expenditure, etc., or making it potentially applicable in a translation project at the cost of compromises, limited range solutions, probabilistic estimates, etc. (This is not to say that the former approach is not applicable, or that it does not keep practical applications in view; such applications, however, would belong to the sphere of the so-called automatic understanding of natural language).

A step in the latter direction, i.e. towards practical application in the translation sphere has been made in the third experiment started in 1982 - APAC3. While firmly based on the APAC2 analysis grammar, it has departed from the previous standards and patterns in more respects, yet, not all of them meant the simplification and compromises alluded at above, which might seem surprising, if one failed to call back to mind what had been said about the essential purpose and character of our experimentation. The most striking change concerned the synthesis phase, where the originally conceived combination with the components of the random generation programme has been abandoned as technically difficult and inconvenient, and a completely new synthesis programme has taken up the outcome of the analysis to transform it gradually into corresponding and acceptable structures in the target language. Another major change worth mentioning consisted in a profound reorganization of the system of dictionaries, where the reduction and simplification strategy characteristic for the APAC2 experiment has been replaced by a theoretically more orthodox approach in which the semantic aspects are respected in a much more consistent way. Thus, the main dictionary falls into its analysis and synthesis divisions again, each operating in its respective domain, and the transducing dic-

tionary (constructing the target language equivalents directly from the source language forms of international words) has been transferred to function in its full scope at the level of morphemic synthesis, which is its proper place making possible its correct and most efficient application. However, a simplified transducing device has been retained in its original place in the analysis to serve as a special "emergency" device which helps to recognize and to semantically interpret an important part of expressions not found in the main dictionary.

APAC3 represents an attempt at a relatively complete system of automatic translation. At the end of 1982 it was capable of handling a fairly extensive selection of frequent structures found in the texts of the genre; it still has a limited dictionary, but its scope (a few hundreds of items) is enlarged by the transducing devices to cover, in case of need, thousands of technical terms. It will be developed in more directions - optimization of the programme, extension and generalization of the rules, extension of the dictionaries, etc., however, the description of its present shape may serve as an example of our experimenting in the field in question. Therefore, in what follows an informal account of the "philosophy" and structure of the project is given in the form of a brief outline neglecting particulars and minor problems.

In the framework of the functional generative description, a dependency grammar of the stratificational type has been implemented in a specific form largely determined by the purpose and by the possibilities given by the formalism employed: for the particular purpose of machine translation linearized tree structures represent the dependency relations with the aid of special nodes attached to each dependent element as the leftmost node governed by the lexical value; these nodes, at the same time, record

the ordering or the elements in that they are accompanied by symbols "L" (left) or "R" (right) to indicate the sense (direction) of the branching. Coordination relations are rendered by means of a special node "COOR" which stands in the position of lexical value, as the leftmost member of the set of the members of a coordination string, and governs the special nodes indicating the function and sense of branching shared by all the members. The other non-dependency relation - apposition - has been treated with a certain simplification as a normal dependency relation, but marked accordingly.

As a means of formal representation and a highly specialized tool for automatic treatment of hierarchical structures with which the structure of natural language can be classed, the so-called Q-systems (Colmerauer 1982) have been adopted. It is a means of implementing a non-deterministic transducer, which can be regarded as a higher level programming language, in which complex context-sensitive grammars can be written as systems of rewrite rules rendering transformations on strings of tree-graphs in linear representation. Such systems - relatively independent sets of rules, grammars - can form a sequence, a system of individual subsystems. Within each subsystem all possible combinations of the applications of the rules to the input string and to its subsequent transforms take place, but, to put it in a simplified way, only those results "survive" that form the simplest path from the beginning of the string to its end: the so-called purification procedure deletes any two or more partial substrings, or their transforms, spanned by one longer transform which represents a more accomplished, continuous (which is, as a rule, correct with a higher probability) interpretation of the corresponding part of the input string changed in this way gradually into one single tree.

The full combinatorics is indirectly controllable (e. g., by changing labels at the nodes of the trees resulting from the applications of the rules, imposing special conditions, introducing special markers, articulating the system into subsystems to interrupt the combinatorial process, etc.), so that the rules can be applied in a required order, if necessary. At any of the stages, the output of the preceding subsystem serves as the input of the immediately following one. In such a hierarchy of subsystems of grammatical rules a great number of possible parses of the given strings or substrings is tried and the most probably correct ones are chosen to undergo further treatment or to represent the results; the hierarchy can be regarded as an excellent instrument for the analysis and synthesis of natural language: all possible interpretations are tested; those that offer acceptable solutions (i.e. conforming to the rules) are automatically preserved (there can be more of them for one string; one of the merits of the Q-systems is that they make possible easy parallel treatment of more alternative structures) to form the point of departure for the subsequent processing.

The Q-language has, of course, its limitations, so that it must be used in combination with other programming languages, but, up to now, at least one of its advantages remains unsurpassed: being simple and clear, the Q-language provides for a lucid, highly transparent formal rendering of the structures of natural language, because it operates on linear (parenthesized) representations of tree structures, and produces patterns similar to the predicate calculus formulas (see Panevová and Oliva, 1982).

As regards the other general problems, space does not permit more than several sketchy observations.

The original triplet scheme - analysis - transfer -
- synthesis - has been abandoned already in the second ex-

periment. In APAC3 the transfer operations are carried out at different steps of the analysis, which in this way, becomes more target-language-specific. This reflects the fact that relationships between pairs of languages are specific, since there usually are more points in which the languages are, so to speak, incompatible. In some cases, it cannot be cured, but in others a more profound examination of the source language is possible and necessary to prepare measures providing sufficient information to be used in the synthesis. A universal analysis, the idea of which stands in the background of the abandoned scheme, is not available at the present time, and experiments of the APACn type may be regarded as steps that bring us closer to this ideal.

If we compare English and Czech, we can find more points in which one of the two languages offers or requires more information than the other. E.g., English, owing to its rather poor inflection and to its almost complete lack of the means of grammatical concord, is in this sense a language more vague than Czech, where the elements bound together by referential relations must agree in case, gender, number, and, with verbs, in person. A similar situation can be found in the sphere of aspect: while the English verb in general can be said to be rather inert or neutral as to aspect, which, in most cases can be explicitly stated only by means of some additional, special devices, the Czech verb, as a rule, contains explicit features expressing aspect; e.g., more verbal forms can be used for the same activity depending on whether it has been completed, whether it represents a habitual action, whether it is still going on, etc. On the other hand, in Czech no articles are used, there are only three tenses, etc.

In general, difficulties in machine translation appear in cases where there is less information available in the

source language than is necessary for the construction of the corresponding target language equivalents. This problem is, in its essence, always connected with solving ambiguities, though in some cases this need not be quite obvious. In this connection it must be noted that, in spite of a number of advantages, the genre of abstracts (summaries) of polytechnical texts by no means represents an easy object of linguistic and semantic analysis. The overwhelming omnipresent tendency to compress and to abridge the text as much as possible leads to a production of "concentrates" abounding in extensive nominal complexes, nominalizations of all kinds, especially condensation with the aid of verbal adjectives and *-ing*-forms, coordination at all levels, long enumerations etc. All this, and the fact that the authors assume that they are addressing an audience of experts and rely on their knowledge of the field in question, results in a great number of ambiguities, which can be resolved on the background of a solid knowledge of the universe of discourse only.

Since in our experimentation no satisfactory model of the universe of discourse was available, we have had recourse to a more or less traditional solution: a structured system of semantic features has been applied. Sets of semantic features organized partially in what with verbs has been called "frames", indicate general and special properties of individual words, and help in arranging the partial structures in overall patterns, in terms of the tectogrammatical representation. Four basic groups of semantic features have been used for the time being: features that help to recognize a metatextual framework in the abstracts, those concerning terminological expressions, helping to distinguish them from the rest and, in a measure, reflecting the position of individual terms in the system (e.g., most general categories, semiterminological expres-

sions, etc.), those indicating general properties or characteristics of concepts (abstract, concrete, human, action, property, etc.), and those that classify the words according to the role or function of their denotates in the objective field of technology and research (instrument, material, location, etc.). Mostly the same types of features have been employed in the frame structures as subcategorization features which refer to the environment rather than represent the properties of their bearers themselves. Thus, the required properties of the participants of verbs are stated in verbal frames, the features assigned to adjectives state the properties of the nouns these adjectives can modify, etc. Besides, grammatical requirements are contained in the frames as well. The structure of the frames as well as the choice of the features, etc., still leave much to be desired, but some experience with this device has proved that its use and further development is, to say the least, promising.

Such an apparatus helps in solving ambiguities, but it probably never will be able to do full justice to the complexities of the given universe, however circumscribed it may be. Therefore, multiple outputs of the analysis of an ambiguous sentence representing different structures possible seem to be inevitable in some cases, which, of course, cannot be regarded as a success in a translation project. However, with some types of ambiguities, fortunately very frequent ones, the structures come out as identical at the output of the synthesis. This is e.g. the case with the representation of the syntactic (and semantic) dependency of prepositional phrases: they very often can depend on more than one of the nouns depending on the finite verb, or form a participant in their own right; however, the resulting target language sentences are identical, being equally ambiguous as the source language structures. It

should be remarked in this connection that, according to our opinion, preserving, or, better, correctly reproducing, the ambiguity of the source utterance in its target version should be ranked as a success rather than as a failure. That is also why in APAC3 a conscious effort to comply with this principle wherever possible has been incorporated in the rules, which, of course, can by no means be regarded as neglecting ambiguities. It goes without saying that even so some ambiguities will "survive" in unsolved form, as multiple output.

So far some general characteristics of our approach and some observations on its "philosophy". What follows is a brief description of the structure of the programme as a whole. Two preliminary remarks are necessary: firstly, we shall confine ourselves to the "core" of the programme - the analysis and the synthesis - neglecting the auxiliary conversion programmes; secondly, the programme falls into 23 subsystems in terms of the Q-language, and the description will follow this scheme; however, this division only roughly corresponds to the logical structure of its contents, since other principles are involved, too. E.g. the storage and time requirements must be respected, the function of the division as a means by which the combinatorics and the order of application of the rules are controlled intervenes, etc. Last but not least, the "preferential" tactics repeating some important rules in a more "liberal" version in subsequent systems to intercept and analyse structures that failed to be analyzed in previous systems due to strict constraints repeats themes solved in principle elsewhere, in systems predominantly devoted to other problems. The boundaries of individual subsystems and the logical steps corresponding to the linguistic structure coincide only at some major structural turning points. As an example, a sentence has been chosen whose main metamor-

phases in the course of processing may serve as an illustration of the effects of the application of some rules.

The first 5 subsystems are devoted to dictionary operations and the morphemic analysis. Subsystem 1 contains the so-called dictionary of constants in which words and word groups that can skip the treatment in morphemic analysis are classified and assigned necessary information. Subsystem 2 contains formal preparation of the treatment in the morphemic analysis and main dictionary sections. Subsystem 3 includes the complete general morphemic analysis connected with the main dictionary: basic or dictionary forms of words are derived and looked up in the dictionary; words not found are decomposed to be subject to further processing. The morphemic analysis has been taken over from the TAUM experiments. Very ingenious and elegant, it covers in some 25 rules the morphemics of English offering to the dictionary treatment either basic forms directly or a choice between alternative forms of which the correct one is to be selected. That is also why in the subsequent 4th subsystem a special morphemic analysis is applied to deal with words that failed to be identified. It prepares them for the treatment in the "transducing dictionary" in the 5th subsystem. This solution, fully satisfactory at the present state of things, may prove inadequate or inefficient in the future, when the main dictionary must inevitably be divided into more subsystems; in each of these subsystems the morphemic analysis would have to be repeated, decomposing and recomposing words time after time, which, of course, would represent a serious burden for the system. Therefore, a new system of a universal morphemic analysis is considered, to make it fully independent of the dictionaries, which would also make any additional special analysis needless.

The 5th subsystem - the transducing dictionary - has a

special task here: it is concerned with assigning some classes of words distinguished by their end segments their most probable part-of-speech category, frame and/or set of semantic features. In this way, it functions as an "emergency" device handling an important part of words - mostly terms or terminological elements - not recognized in the dictionaries. The full version of the original transducing dictionary has been transferred to the 18th subsystem, i.e. to the level of morphemic and, in this case, also orthographical synthesis. The dictionary operations provide the words with information of systemic, grammatical and semantic nature. Some of the information is present in a coded form, only to be stated explicitly in some of the following subsystems. It should be added that an important division of the dictionary system - a dictionary of idiosyncratic compound expressions - is contained in the 6th subsystem: it handles compounds whose structure cannot be satisfactorily accounted for by general rules, or which behave in an anomalous way judged by the standards of both, the source language or its target counterpart.

The subsystems 6-13 can be roughly characterized as a gradual construction of nominal complexes. The 6th subsystem deals with the immediate outcome of the dictionary operations on the one hand, and prepares the subsequent ones on the other: words that had not been identified in any of the preceding operations are temporarily interpreted as names (i.e. proper names, acronyms, abbreviations, etc.); the detached morphemes (*-ing*, *-ed*, *-s*, *-ly*, etc.) are interpreted in connection with their bases as, e.g., *ing*-forms, past tense or past participle forms, third person of present tense with verbs or plural with nouns, adverbs, comparatives, etc. Subsystem 7 continues in interpreting the basic verb forms (including negation and modification by adverbs) and, as its main component, comprises a set of

rules concerning elementary syntax of nominal complexes, dealing with the most frequent and relatively unproblematic phenomena. Rules concerned with coordination of adverbs and adjectives follow. The 6th subsystem is predominantly devoted to nominal syntax, too. The elementary syntax of nominal complexes continues by another: set of relatively simple rules. A limited set of rules concerns the syntax of what had been called names (see above), and another one handles simple coordination of nouns. The elementary constructions of nouns modified by verbal adjectives are dealt with in another set of rules - the first in a series of three that, with growing complexity, analyze this type of modification in the 10th and 13th subsystems. Preparations for the analysis of indirect object constructions anticipate as early as that the verbal syntax. Subsystem 9 is again concerned mostly with the initial stages of the verb group analysis: some operations prepare the solution of relative clause constructions and, in connection with it, the interpretation of simple *-ing* constructions. Isolated names are classed as nouns.

Subsystem 10 concerns more complicated syntactic issues of nominal complexes: prepositional modification, nouns with special prepositional constructions (and analogous rules for verbs and adjectives), and modification of nouns by verbal adjectives; verb group syntax is represented by rules in which infinitives have their frames filled in. The analysis of one-member sentences is accomplished here, which, at the same time, signals that the basic syntax of nominal complexes is regarded as finished, too.

This is the proper place to come back to our promised example: a simple sentence, taken from a scientific journal - "An amplifier that activates a passive network to form an active analogue is called an operational amplifi-

er" - comes out of the 6th subsystem as a string of trees, in which some parallel structures occur (the nodes in the string at which the path starts, ends or divides into parallel branches are numbered). PHASE NO 6

-Ø1- \$(111) + ART + N(AMPLIFIER (/,*SG) , ; , *C, *TNST,*AG, *SG, Ø)+THAT(*)+V(ACTIVATE(/, *SG) , ; , 1(I), 2 (J) ,Ø)+ART+AD (PASSIVE(/) , ; , *A,*C, *MNR,*SG,1)+N(NETWORK (/,*SG) , ; , *A,*C,*STRM,*SG,Ø)+P(TO) -Ø2-
 -Ø2- V(FORM(/) , ; , 1(I), 2(L), 5,Ø) - Ø3-
 -Ø2- N(FORM C/, *SG) , ; , *A,*C,*STRM,Ø) -Ø3-
 -Ø3- ART+AD(ACTIVE (/) , ; , *A,*C,*MNR, *SG,1) +N (ANALOG(/ , *SG) , ; , *A,*C,*SG,)+V(BE(/,*SG) , ; , *CO) -Ø4-
 -Ø4-V(CALL(/) , ; , 1(I), 2(J), 3(N) , *EFE,*AUTH,*ED,Ø) -Ø5-
 -Ø4-V(CALL(/) , ; , 1(I), 2(J), 3(N) , *EFF,*AUTH,*EN,Ø) -Ø5-
 -Ø5-ART+AD(OPERATIONAL (/) , ; , *A,*C,*MNR,*SG,1)+N(AMPLIFIER (/,*SG) , ; , *C,*INST, *AG,*SG,Ø)+ -Ø6-

Note the parallel interpretations of the word FORM and the word form CALLED. In the string leaving the subsystem 10 the correct interpretations are accepted and the wrong ones are deleted. The graph (of tree graphs forming a string) does not contain any parallel structures, which is why only the entry node and the exit node are numbered (-Ø1- and -Ø2- respectively). It can be seen that the frame information with verbs is explicitly stated, that the modification of nouns has been analyzed in that the adjectival attributes have been attached to their head nouns as dependent structures, and that the frame slot of the infinitive (TO FORM) for the function "patient" has been occupied by the noun ANALOG modified by the adjective ACTIVE. PHASE NO 10

-Ø1-\$(111)+N(AMPLIFIER(*IDF,/, *SG) , ; , *C,*INST,*AG,*SG,Ø)+REL(THAT)+V(ACTIVATE(/, *SG) , ; , 1(*A,*C,*H), 2(*A,*C,*H,*OB), Ø,)+N(NETWORK(*IDF,/, *SG), AD(PASSIVE(L,\$ATR,/)), ; , *A,*C,*STRM,*SG,Ø)+INF(FORM(/), N(ANALOG (R,\$PAT ,*IDF,

```

/, *SG), AD(ACTIVE(L, $ATR, /)), ;, 1(*A, *C, *H) 5, Ø, #)+V
(CALL(*PSV, /, *SG), ;, 1(*A, *C, *H), 2(*A, *C, *H, *OB), 3(*A,
 *C, *H, *SS), *EFF, *AUTH, EN, Ø, #)+N(AMPLIFIER(*IDF, /, *SG),
 AD(OPERATIONAL(L, $ATR, /)), ;, *C, *INST, *AG, *SG, Ø)+.-Ø2-

```

The 11th subsystem continues in preparing the analysis of other indirect object patterns and *ing-form* constructions. The subsystem 12 contains a set of rules that create special subtrees representing in a schematic way full context of each component of the sentence at this particular stage: these subtrees are attached as a part of general information to all the members of the string, except for the remaining detached morphemes and similar formal items, which, however, figure in the context image as any other more complex sentence element. The 13th subsystem includes the third and most intricate set of rules devoted to the modification of nouns by verbal adjectives; they, among others, utilize the information on context provided by the preceding subsystem. As another part of the preparatory operations dealing with the relative clause construction, a set of rules marking off the scope of the relative clause is added. Subsystem 14 can be characterized as rules for sentence syntax. Along with some sets of rules devoted to some special problems, viz. prepositional phrases, relative clauses integrated as modification of the preceding noun, some more complicated *-ing* constructions, etc., the main blocks of rules deal with filling in the slots in verbal frames and manipulating clauses, their coordination, etc. Here, the sentence parse is accomplished and the analysis is finished.

The result can be seen on the output of the 14th phase: the sentence is represented as one tree structure dependent on a formal node labelled as S. Empty actor (agentive) "NIL" and patient AMPLIFIER depend on the finite verb CALL in passive voice, which demonstrates that the passive

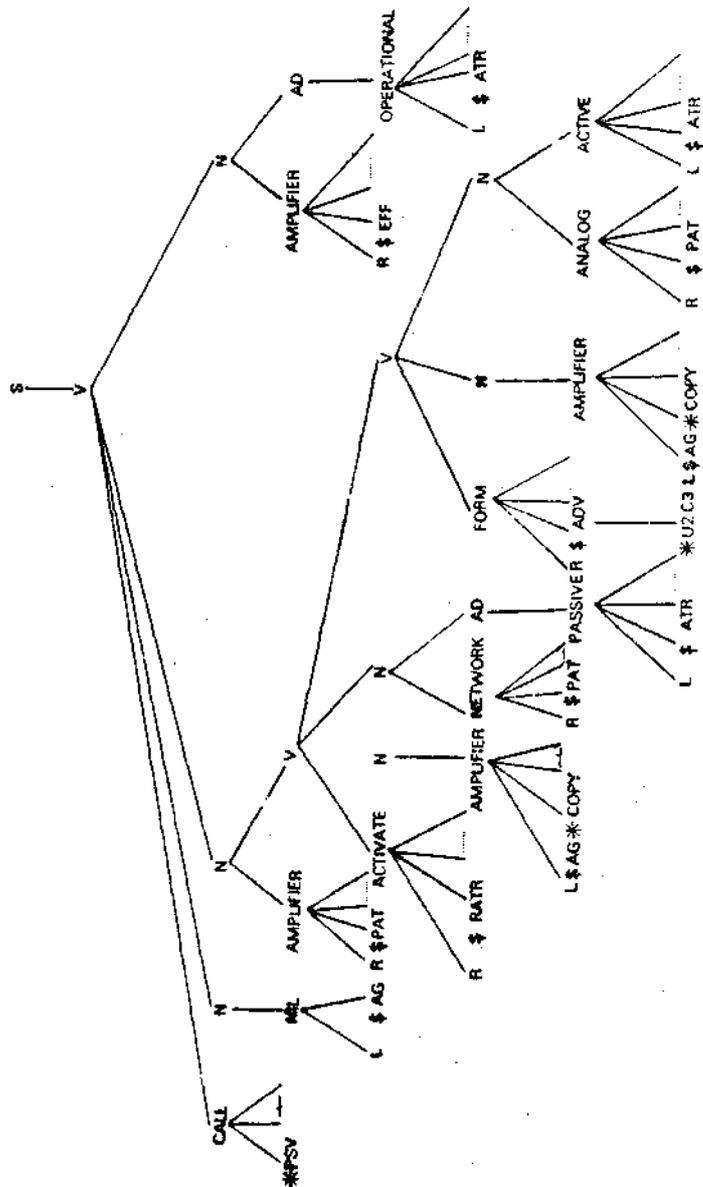


Fig. 1 A graphic representation of the output of the subsystem 14. Immaterial indices are left out.

structure is semantically interpreted: the third participant is the effect: OPERATIONAL AMPLIFIER. The patient is modified by a relative clause headed by the verb ACTIVATE with its participants - AMPLIFIER as a copy (dummy) actor, and PASSIVE NETWORK as a patient - depending on it: another clause, this time final, depends on the verb ACTIVATE, having as its finite verb FORM with a copy of the "copy" actor AMPLIFIER as actor and ACTIVE ANALOG as patient. All information that becomes superfluous at this stage is dropped.

```
-Ø1- $(111)+S(V(CALL(*PSV,/,*SG),N(NIL(L,$AG)),N(AMPLIFIER
(R,$PAT,*IDF,/,*SG),V(ACTIVATE(R,$RATR,/,*SG),N(AM
PLIFIER(L,$AG,*COPY,*IDF,/,*SG)),N(NETWORK(R,$PAT,
*IDF,/,*SG),AD(PASSIVE(L,$ATR,/J)),V(FORM(R,$ADV
(*U2C3),/),N(AMPLIFIER(L,$AG,*COPY,*IDF,*COPY,/,*SG)),
N(ANALOG(R,$PAT,*IDF,/,*SG),AD(ACTIVE(L,$ATR,/)))))),
N(AMPLIFIER(R,$EFF,*IDF,/,*SG),AD(OPERATIONAL(L,$ATR,
/))))))+.-Ø2-
```

Graphic representation is presented in Fig. 1.

The synthesis starts in the 15th subsystem by a gradual decomposition of the S tree. First, the basic word-order is reconstructed. The subsystem, taking into consideration the voice of the finite verb (decomposes the structure into the main (finite) verb and its participants, and orders these components according to the principles of Czech word order. The functional sentence perspective is taken into consideration. The 16th subsystem decomposes individual components separated in the previous treatment according to the information on left or right branching in the tree.

PHASE NO 16

```
-Ø1- $(111)+V(AMPLIFIER(L,$SUBP,*IDF,/,*SG))+$RATR(KTERE2
(5))+N(AMPLIFIER(L,$AG,*COPY,*IDF,/,*SG))+V(ACTIVATE
($RATB,/,*SG))+AD(PASSIVE(L,$ATR,/))+N(NETWORK(R,$PAT,
*IDF,/,*SG))+$ADV(*U2C3)+N(AMPLIFIER(L,$AG,*COPY,*IDF,
```

```
*COPY,/,*SG))+V(FORM(R,$ADVB(*U2C3),/))+AD(ACTIVE(L,
$ATR,/))+N(ANALOG(R,$PAT,*IDF,/,*SG))+V(CALL(*PSV,/,
*SG))+AD(OPERATIONAL((L,$ATR,/))+N(AMPLIFIER(R,$EFF,
*IDF,/,SG)).-Ø2-
```

The 17th subsystem contains the bilingual dictionary. An introductory rule isolates every lexical value from the rest of the information, which is stored as a separate tree accompanying it in the string. Then, the lexical values are interpreted: they are assigned Czech equivalents classified according to the part-of-speech category and the paradigm. E.g. the value FORM is assigned the following information:

```
SL (VYTVOR3(42P))
```

```
SL (TVOR3(42I))
```

```
JM (FORM FØ2)
```

which means that it can be interpreted as the imperfective verb TVOR3IT whose paradigm pattern is coded as 42, or the perfective verb VYTVOR3IT belonging to the same paradigm, or the noun FORMA, paradigm FØ2. Words not found in the dictionary are treated in the 18th subsystem, containing the full scope transducing dictionary mentioned above. Here all the words of international usage are transformed into their Czech equivalents and assigned the pertinent grammatical information. If we look at the 18th phase in our example, we must be aware of the fact that the Czech words PASI2VNI2 (PASSIVE), AKTIVNI2 (ACTIVE) and OPERAC3NI2 (OPERATIONAL) have been constructed from their English counterparts by the rules of the 18th subsystem. The reader will note that, again, in the example more parallel structures can be found - three with the original word FORM - two verbs and a noun, and two with the word CALL, interpreted as two verbs. With verbs, except for such verbs as ACTIVATE, which are neutral as to aspect in Czech, always two alternatives are given - a perfective and an imperfective one; the decision as regards the part-of-speech cate-

gory depends on the accompanying information: e.g., with FORM, the following information is headed by V, which means that only verbal equivalents may be chosen; the choice between the perfective and imperfective alternative depends on more relatively complicated rules which cannot be dealt with in this brief outline.

PHASE NO 18

- Ø1- \$(111)+JM(ZESILOVAC3(M07)) + N(L, SUBP,*IDF,/,*SG)+
\$RATR(KTERE2(5))+JM (ZESILOVAC3 (M07)) +N (L, \$AG, *COPY,
*IDF,/,*SG)+SL(AKTIVIZUJ (30N))+V(\$RATB,/,*SG)+PJ(PA
SIV2NI2(9) +AD(L,\$ATR,/) +JM(SI2T(F14))+N(R,\$PAT,*IDF,
/,*SG)+\$ADV(*U2C3)+JM(ZESILOVAC3(M07))+N(L,\$AG,*COPY,
*IDF,*COPY,/,*SG)-Ø2-
- Ø2- SL(TVOR3(421))-Ø3-
- Ø2- SL(VYTVOR3(42P))-Ø3-
- Ø2- JM(FORM(F02))-Ø3-
- Ø3- V(R,\$ADVB(*U2C3),/)+PJ(AKTIVNI2 (9))+AD (L,\$ATR,/) +JM
(ANALOG(M 2))+N(R,\$PAT,*IDF,/,*SG)-Ø4-
- Ø4- SL(NAZY2V(50I))-Ø5-
- Ø4- SL(NAZV(14P))-Ø5-
- Ø5- V (*PSV, / , *SG) +PJ (OPERAC3NI2 (9)) +AD (L, \$ATR, /) +JM (ZESI
LOVAC3 (M07))+N(R, \$EFF , *IDF,/, *SG)+. -Ø6-

In the 19th subsystem the forms are constructed and the information on concord is gathered. It should be observed in this connection that this subsystem as well as the subsequent three, though working at what may be called lower levels of the language, are by no means the simplest ones. Especially the 20th, 21st and 22nd subsystems dealing with the transfer of information on case, number, gender, etc., from the governors to the dependent elements and from the subjects to the predicates contain sets of rather intricate rules which act in cycles and move the complexes of data in both directions - left or right - depending on more conditions and circumstances. The last subsystem - the

longest one - is devoted to morphemic synthesis. Its outcome is a string from which only the intervening plus signs are to be deleted and on which some minor adjustments must be carried out to obtain the resulting Czech translation in an acceptable form.

PHASE NO 23

-Ø1- \$(111)+ZESILOVAC3+,+KTERY2+AKTIVIZUJE+PASI2VNI2 + SI2T3
+,ABY+TVOR3IL+AKTIVNI2+ANALOG+,+SE+NAZY2VA2+OPERAC3N
I2+ZESILOVAC3+.-Ø2-

As has been already said, much remains to be done, and the work we face will be very complicated and difficult. The first steps have brought us invaluable experience and encouragement, though not excessive optimism. In conclusion we present examples of the analysis and translation of some further sentences (input texts and outputs of phases 14 and 23).

Input sentences:

- (1)-Ø1-\$ (8Ø4)+THE+DIMENSIONS+OF+THE+BOARDS+ARE+MINIMIZED+
WHEN+MONOLITHIC+TECHNOLOGY+IS+USED+. -Ø2-
- (2)-Ø1-\$ (919)+THE+PEGAMAT+TESTING+EQUIPMENT+THAT+WAS+USED+
IN+THE+NETWORK+WAS+ACTIVATED+,+ALLOWING+THE+SYSTEM+
HIGH+STABILITY+. -Ø2-
- (3)-Ø1-\$ (926)+A+COMPUTERIZED+AUTOMATIC+OPTIMIZATION+STRA
TEGY+IS+DESCRIBED+USING+A+SYSTEMATIC+CONCEPTION+BA
SED+ON+STATISTICAL+,+CUMULATIVE+AND+TENTATIVE+PRO
CEDURES+.-Ø2-

Outputs of phases 14:

- (1)-/1-\$ (8/4)+S(V(MINIMIZE(*PSV,/),N(NIL(L,\$AG)),N(DI
MENSION(R,\$PAT,*DEF,/,*PL),N(BOARD(R,\$ATR(OF),*DEF,
/,*PL))),CLS(USE(R,\$ADV(*TCND,WHEN),*PSV,/,*SG),N
(NIL(L,\$AG)),N(TECHNOLOGY(R,\$PAT,/,*SG),AD(MONOLITHIC
(L,\$ATR,/))),&)))+.-Ø2-
- (2)-Ø1-\$ (919)+S(V(COOR(*RATR),CLS(ACTIVIZE(*PSV,*PST,/,
SG),N(NIL(L,\$AG)),N(EQUIPMENT(R,\$PAT,*DEF,/,*SG),

AD (TEST (L, \$ATR, *TV, *VAD2, /)) , N (PEGAMAT (R, \$APP, *NSK, /)) , V (USE (R, \$RATR, PSV, *PST, / , *SG) , N (NIL (L, \$AG)) , N (EQUIPMENT (R, \$PAT, *COPY, *DEF, /, *SG)) , N (NETWORK (R, \$ADV (*LOC (IN)) , *DEF, /, *SG)) , &)) , CONJ (*RATR) , CLS (ALLOW (/ , *SG) , PRN (THIS (L, \$AG, /, *SG)) , N (STABILITY (R, \$PAT, *NIO, /, *SG) , AD (HIGH (L, \$ATR, /))) , N (SYSTEM (R, \$ADR, *DEF, *TY (MØ1) , /, *SG)))))))) .-Ø2-

(3) -Ø1-\$ (926) +S (V (DESCRIBE (*PSV, /, *SG) , N (NIL (L, \$AG)) , N (STRATEGY (R, \$PAT, *IDF, /, *SG) , AD (COMPUTERIZE (L, \$ATR, *VAD, /)) , N (OPTIMALIZATION (R, \$ATR (OF) , *NN, /, *S.G) , AD (AUTOMATIC (L, \$ATR, /)))) , N (USE (R, \$ADV (*DPRV (S7)) , *NV, /, *SG) , N (CONCEPTION (R, \$PAT, *IDF, /, *SG) , AD (SYSTEMATIC (L, \$ATR, /)) , AD (BASE (R, \$ATR, *VAD 1, /)) , N (PROCEDURE (R, \$ADV (*RESP (NA6)) , /, *PL) , AD (COOK (L, \$ATR, AND) , AD (STATISTICAL (, , /)) , AD (CUMULATIVE (/)) , CONJ (AND) , AD (TENTATIVE (/)))))))))))) .-Ø2-

Output sentences:

- (1) -Ø1-S (804) +DIMENZE+PANELU2+SE+MINIMALIZUJI2+, +KDYZ3+SE+UZ3I2VA2+MONOLITICKA2+TECHNOLOGIE+.-Ø2
- (2) -Ø1-S (919) +TESTOVACI2+ZAK3I2ZENI2+PEGAMAT+, +KTERE2+BYLO+UZ3ITO+V+SI2TI+, +BYLO+AKTIVIZOVA2NO+, +COZ3-DOVOLUUE+SYSTE2MU+VYSOKOU+STABILITU+.-Ø2-
- (3) -Ø1-S (926) +POPISUJE+SE+KOMPUTERIZOVANA2+STRATEGIE+AUTOMATICKE2+OPTIMALIZACE+S+UZ3I2VA2NI2M+SYSTEMATICKE2+KONCEPCE+ZALOZ3ENE2+NA+STATISTICKY2CH+, +KUMULATIVNI2CH+A+TENTATIVNI2CH+PROCEDURA2CH+.-Ø2-