Automatic Syntactical Analysis of the Czech Language

JARMILA PANEVOVA,
PETR SGALL

Charles University,
Prague. Czechoslovak
Socialist Republic

A strategy of automatic syntactico-semantic analysis of Czech is described, based on the dependency relations and having as its output the semantic (tectogrammatical) representations of sentences as proposed in the framework of functional generative description. The approach is illustrated by means of an example of a partial algorithm, demonstrating the use of linguistic criteria for the determination of particular functions. The system is being implemented in Q-language.

The rapid development of the projects dealing with automatic understanding of natural languages entails the necessity of explicit descriptions of syntactical analysis systems. In comparison with the majority of first projects in mechanical translation we are dealing here with a new stage in linguistic analysis in the form of more or less formalised models which can be realised with the aid of computers.

We have described elsewhere [1—5] the theoretical background of a synthesis of Czech sentences, which is tested in a system of random generation of sentences by computers, and serves also as a part of several systems designed for applications (mechanical translation, question-answering systems). This description has also served as a starting point for the formulation of a procedure for automatic syntactical analysis of Czech. The output of the complete syntactical analysis (SA) is understood as being at the same level as semantic representations or tectogrammatics, the structure (and inventory of units) of which has been systematically studied in connection with the formulation of the procedure of automatic synthesis and generative description.

Within the presently prepared system of automatic answering of questions in Czech this level is used also for the formulation of rules of inference leading from an assertion (or from the conjunction of two assertions) to its consequences and operating on tectogrammatical representations of sentences. These rules make it possible for the system to answer questions which do not directly correspond to the input sentences (or to their paraphrases), but can be answered only on the basis of the consequences of some of the sentences that were included in the input text. In this way it is possible to formulate a fully automatic system of question answering, based only on input texts, automatic analysis and rules of inference, without any prerequisites such as a data base, or any other procedural adjustments in communication 'for the needs of computers'.

It is a difficult task to formulate an automatic grammatical analysis that would not be constructed merely *ad hoc* for a certain set of texts, but would be founded on a sound theoretical base granting a wide use. This task has been made considerably more feasible by the fact that the generative procedure (and synthesis) has already been formulated and widely checked. It is, of course, not possible simply to reverse the procedure of synthesis, since questions of ambiguity, which have to be solved in an efficient analysis, are not relevant to synthesis. Nevertheless, the task has been facilitated to a high degree by the possibility of using in analysis a set of units of individual levels and relations between these units (elementary and complex units of the levels of tectogrammatics, surface syntax, morphemics) which has already been established with respect to the procedure of synthesis.

It is especially important that we already know what form the output of SA should take: it is a semantic (tectogrammatical) tree representation of the sentence the nodes of which are labelled by complex symbols composed of three kinds of elementary symbols (semantemes): lexical, syntactical (the participants, or deep cases, and free adverbials, i. e. types of the dependency relations), and morphological (such as plural, preterite, etc., called grammatemes).

Also the lower level representation forms of the sentences and the assortment of the units of these levels have been studied and checked in connection with the procedure of synthesis: the relationships of 'realisation' of higher level units ('functions') by lower level units ('forms') have been established for every pair of adjacent levels.

There are now two tasks that remain for the formulation of automatic analysis:

(a) to find and explicitly state contextual criteria for all cases of lexical and grammatical ambiguity, accounting for their interplay in an adequate and efficient way:

(b) to choose the technical means necessary for the implementation of a system of analysis.

Point (a) was elaborated for Czech during the 1970's in the form of a preliminary model which should cover syntactical phenomena frequently occurring in various technical texts: the lexical phenomena with which we are concerned have a rather limited scope and concern texts on electronics.

As for point (b). the technical means chosen is the Q-language. elaborated by a group of Canadians working with machine translation (Colmerauer, Kittredge, Thouin, Chandioux, Isabelle and others). This language, designed for transformations of trees into trees, is used with a program written in FORTRAN 4, by the Quebec group, which interprets and executes the sets of rules formulated in the form of Q-systems,

The input of Q-systems is the output of morphemic analysis (MA) of the Czech language [6]; in such an inflectional language as Czech morphemic analysis plays a great role, but the wide range of homonymy of single word forms must be considered. In any case the SA must be built on the morphemic description of Czech.

The result of MA received a form which can be combined with a lexicon the entries of which have a form suitable to Q-language. Every lexical unit found in the output of MA and connected with certain morphological data is identified in the 'syntactic' lexicon (in the form of a Q-system). The new characteristics due to this lexicon are, e. g., the valency of the word with a marker of the obligatoriness or optionality of individual modifications and with the surface form of modifications required by the governor, the 'semantic part of speech', the semantic features, an indicator of the synonymous lexical units and of the counterpart of different aspect for verbs.

A question-answering system based on natural language and supposing an immediate man-machine communication must be prepared for questions of various surface shape, which must — in the case of their equivalence— be reduced to one deep (tectogrammatical) representation.

The relationship of synonymy between two or more surface lexical units having identical syntactic properties is denoted as SYN 1. One of the lexical units is denoted as primary; it is then substituted for the other(s) by means of lexical rules. The synonymy in a broader sense (not requiring an identity of syntactic properties) is denoted as SYN 2. This relationship concerns e. g. the deverbative noun and the corresponding surface verb, or two synonymous verbs differing as to the form of their participants. The characteristics of SYN 2 are maintained and in the final part of SA both constructions with the relation SYN 2 are unified into a single tectogrammatical representation (e. g. 'Pro *udrženi* lineární závislosti se používá X.' — 'Aby *se udržela* lineární závislost používá se X.' = 'To maintain- a linear relationship X is used').

We present here some examples of the lexical rules in Q-language:

```
N(ELEMENT(*N(12, 13, 16), 5)) =
    = N(PRVEK(*N(12,13,16),*K,*SEMN,5))
V(JMENOVAT(OS(3),SG,PRAES,AKT,NEDOK,7)) =
    = V(NAZY2VAT (OS (3),SG,PRAES,AKT,NEDOK,
    OPAT(4),OEFF(1,7),*R,*ST,*SEMV,7))
N(APLIKOVA2NI2(*S(17),12)) =
    = A(APLIKACE(*F(17),FPAT(2),*A5,*SEMV,
    SYN2(UZ312VAT),12)).
```

It must be noted that in Q-language, for example, the string A(B, C(D)) denotes a tree, the root of which is labelled A, whereas B and C are the (labels of the) daughter nodes of the root and D depends on C. As for the individual symbols: N — noun, V — verb, *N — masculine inanimate, 12 —genetive sg., 13 — dative sg., etc. The last digit (e. g. 5 with the word ELEMENT) is the serial number of the word in the sentence, *K — semantic characteristic (concrete), *A5 —nouns of events, *SEMN—semantic noun (i. e. a noun on the tectogrammatical level), *SEMV—semantic verb, OS(3) — 3rd person, AKT — active voice, and NEDOK — impf. aspect. The general form of a particular part of speech can be illustrated by the example of the nouns:

```
N(A*(B*(V*),A*1))+A*(U*) =
= N(C*(D*(V*), U*, A*1)).
```

This is only a scheme, not a rule in Q-language; A*, C* are variables for lexical units, B*, D* are variables for gender, V* is a variable for the list of cases, U* is a variable for the characteristics obtained from the syntactic lexicon, and A*l is a variable for the serial number of the word; OPAT (4) denotes that a word has an obligatory patient (object in the accusative, etc.). In normal cases (without SYN 1) the relations A*=C*, B*=D* hold.

A simple algorithm connected with the complex lexical units is then applied. As to SA itself, the procedure describing the structure of simple noun phrases (NP) has been completed and some of the major questions concerning the determination of the structure of complex NP's have been solved. One of these problems consists in the necessity to distinguish between adverbials (depending on a verb, adjective or adverb) and nominal adjuncts in such cases as *psal tužkou na papir (he wrote with a pencil on (a) paper),* cf. the well known example of ambiguities in 'He saw a man in the park with a telescope'. Questions connected with such syntactic ambiguities were discussed by Panevová [7] and have been analysed for different types of tests, since in every technical domain different preferences for certain prepositions, conjunctions, cases, etc. can be found.

It should be recalled that SA is understood not only as determining the relationship of dependency (for every node of the dependency tree, except for its root, one and only one governing node must be identified), but also as determining their roles on the tectogrammatical level, e. g. agentive, patient, addressee, effect, cause, purpose, direction (with a differentiation according to the meaning of prepositions—*in, on, behind, beside,* etc.).

A set of partial algorithms has been prepared by a group of algebraic linguistics of Charles University, by means of which the tectogrammatical roles of several tens of morphemic means (cases, prepositions, subordinative conjunctions) are identified, using various contextual criteria. These partial algorithms (prepared by J. Panevová, A. Bémová, E. Buráňová, K. Králíková, P. Pitha and others) are now being combined into an integrated syntactico-semantic analysis of Czech.

If we assume the question of the determination of the dependency relationship as having been solved, we can now give the following example of determining functions of the prepositional group (PG) *bez* + genetive; the rules are expressed in a readable, non-coded form*:

Instructions for the analysis of *bez* (without) + genetive:

	YES	NO
1. Is *bez* (without) followed by *ohledu na, zřetele* na?	7	2
2. Does the PG depend upon a noun?	8	3
3. Is the PG a regular participant of the governing verb (see the data given in the lexicon)?	9	4

* This partial algorithm, as well as the above mentioned division of lexical units into semantic groups, was worked out by E. Buráňová.

4. Does the given clause contain a modal verb or one of the expressions *lze, je možne* (it is possible)? 10 5

5. Is *bez* preceded by the conjunction *i* (even)? 11 6

6. Does the given clause contain a verb in the conditional mood (not combined with the conjunctions *aby, kdyby)?* 12 8

7. The PG has the function of a compound preposition introducing an adverbial of regard.

8. The PG functions as an adjunct of accompaniment.

9. The PG functions as an object.

10. The PG functions as an adverbial of a real condition.

11. The PG functions as a concessive adverbial

12. The PG functions as an adverbial of an unreal condition.

In a similar way the identification of the roles (functions) of other PG's is performed. We may state in general, that the main points underlying the SA of PG's (and also of subordinate clauses, which play similar roles as PG's, as well as of NP's without prepositions) are: the valency of the governor, the semantic characteristics of the governor or of the analysed noun (inside NP or PG); only accidentally a broader context (e. g. the presence of other modifiers of the governor) is to be examined.

An algorithm identifying (in a somewhat simplified way) the individual words as belonging to the topic or to the focus of the sentence (or also of an embedded subtree) has also been formulated*.

In such a way a corpus of polytechnical texts was empirically analysed and generalisations were made similar to those mentioned above.

Another problem is the choice of a successful strategy for the syntactic parser. Some features connected with our strategy belong to the inherent features of Q-language. This language is a good means for such linguistic tasks as. syntactic analysis; it is simple and transparent enough to be used by linguists themselves as a programming tool.

From the linguistic point of view, it is important that a level of analysis has been achieved which specifies not only surface structure of sentences, but also the semantic roles played by the individual sentence parts.

Many ambiguities have been found and analysed in the syntactic structure of Czech, most of which have their more or less exact counterparts in other languages. The experience of the present approach to syntactic analysis may thus be useful in the general research into the strategy of natural language parsers and recognition routines. Since the dependency syntax, which has been well established in continental linguistics for decades, stands close to the framework of categorial grammar, interesting results for theoretical linguistics may be also achieved.

REFERENCES

1. Sgall, P.; Nebeský, L.; Goralčiková, A.; Hajičová, E. *A functional approach to syntax in generative description of language.* New York: American Elsevier, 1969.

2. Sgall, P.; Hajičová, E. A 'functional' generative description. *Prague Bulletin of Mathematical Linguistics,* 1970, **14**, 3—38; see also *Revue Roumaine de Linguistique,* 1971, **16**, 9—37; a complemented version see in: *Functional generative grammar in Prague.* Ed. by W Klein and A. v. Stechow. Kronberg/Taunus, 1973, *1—52.*

3. Hajičová, E.; Sgall, P. A dependency based specification of semantic representations. Presented at COLING 78, Bergen, Norway (in press).

4. Panevová, J. Verbal frames revisited. *Prague Bulletin of Mathematical Linguistics,* 1977, **28**, 55—71.

5. Panevová. J. Funkce a formy ve stavbě české věty. Prague: Academia (in press).

6. Weisheitelová, J. Morphemic synthesis. In: *Explizite Beschreibung der Sprache und automatische Textbearbeitung, V.* Prague (in press).

7. Panevová. J. Non-agreed attribute from the analysis viewpoint for machine translation purposes. *Prague Studies in Mathematical Linguistics,* 1966, **1**, 219—239 (in Russian).

8. Sgall, P.; Hajičová, E. Focus on focus. *Prague Bulletin on Mathematical Linguistics,* 1977, **28**, 5—54; 1978, 29, 23—42.

• For a discussion underlying this identification see [8, especially the sections 3.7 and 3.81