[From: *Beiträge zur Linguistik und Informationsverarbeitung* 8, 1966]

Computational Linguistics in Czechoslovakia

Von P. SGALL und E. HAJIČOVÁ, Prag

The aim of this short survey is to supply short information about the studies in computational linguistics in Czechoslovakia. Since the authors work at Charles University in Prague, a more detailed description of studies conducted there will be given with a short bibliography of the most important works of Czech linguists working in that field, esp. those written in English, Russian, German or French.

I

Linguistic studies have a long tradition in Czech linguistics and therefore there is no wonder that new theoretical developments and methods have been studied deeply and in many respects.

Let us only recall the pre-war activities of the so-called Prague School; in many former works of its members there can be found outstanding and revealing ideas, which anticipated much of what is just now being studied by the workers in the field of mathematical linguistics, both in the respect of its statistical (quantitative) and theoretical (algebraic) branches (cf. [47]).

Since the very beginning of their existence the respective groups of computational linguistics at Charles University (one attached to the faculty of Philosophy, the other to the faculty of Mathematics and Physics, both under the leadership of P. Sgall) have focused their attention on the development of contemporary theoretical linguistics; foreign works using a mathematical approach for the aims of linguistic description have been commented and reviewed (cf. articles by L. Nebesky, P. Novák, B. Palek, P. Sgall — esp. in *Slovo a slovesnost* and in *The Prague Bulletin of Mathematical Linguistics*). An attempt of an axiomatic formulation of the relation of "form" to "function" in language has been made [22]; several models of formal grammars have

been proposed starting from a certain relation on the set of strings and constructing a kind of generative grammar, substantially different from existing phrase structure grammars. The generative procedure is based on the structural analogy of strings [23], [26]. A certain model of sentence analysis has been studied [24], [25]; based on the knowledge of the set of "subjects" of the given sentence, a dependency tree is assigned to that sentence and the possibilities and limits of such a procedure are stated and analysed.
After the first period, the research performed by the group has been concentrated on one main aim: we are now preparing an outline of a generative description of the Czech language, which could be used, too, as a basis of the algorithms of synthesis for MT (in the translation into Czech). The general features of this description can be characterized as follows [41], [42], [43], [44]:

1. The description uses several linguistic levels (a phonetic level, a phonemic or morphophonemic level, a morphological level, a level of sentence parts corresponding roughly to the traditional notions of dependency syntax, and a tectogrammatical level based on the "semantic sentence structure" analysed by several members of the Prague School of structural linguistics [5], [3], [4]), where the representations of a sentence on two adjacent levels are in the relation of "form" to "function" (representation or realization in the terminology of S. M. Lamb's "Stratificational Linguistics").

2. The description is a sequence of several components; the first is a generative grammar which generates strings of the tectogrammatical level, the following are transducers which translate the generated strings step by step into other levels.
In that way, the output string of every component (i. e. a certain representation of the sentence on a given level) is handled as an input string of the following component and is transduced by it to one or more other representations until the phonetic representation is reached at last.

3. As to the weak generative power, the first component should be weakly equivalent to a context-free phrase structure grammar, the transducers should have the form of pushdown store machines for the translation from the tectogrammatical to the morphological level) and of finite state automata (for the translation from the morphological to the phonetic level).

That is to say, this system belongs to that of the "two major avenues for extension of syntactic models", (as distinguished by D. G. Hays in his Dependency Theory. A Formalism and Some Observations, *Language* 40, 1964, 522), which extends grammars by combining them in sequences (as opposed to the transformational grammar).

4. Sufficient strong generative power is gained by distinguishing the level of sentence parts and the tectogrammatical level: in such a system a non-homonymous sentence obtains two different structural characteristics, both pertinent to its syntactic structure (similar to the underlying P-marker and the T-marker in transformational grammar).

Monographical research is being carried out, concerning several parts of the description of Czech syntax and respective empirical questions, as, for instance, the problem of co-ordinate constructions.
It is obvious, that the relations of coordination and apposition cannot be handled in a simple way either in a dependency grammar or in the transformational grammar [37]; the possibility of including them into the system just described is being studied. A description of Czech co-ordinate constructions on the tectogrammatical level has been proposed [35]; this outline is based on material originally collected and studied for the work on a procedure for an automatic analysis of Czech co-ordinative conjunctions in MT [36].
Large empirical work [33] also stood at the beginning of the study of another interesting problem, namely the questions of nominalization in Czech. In the framework of the generative system described above, the question of dependent predication (i. e. the predication besides the main predication in the sentence, which can be rendered in various ways: dependent clause, semi-sentential — infinitive or participial — construction, nominal construction) is being studied. The attention is focused on the verbal categories rendered by the different degrees of condensation (tense, relative tense, aspect, etc.) and the conditions for the choice of a given type of construction are being formulated [34].
Many theoretical studies are devoted to the questions of dependency grammars as such [27]. By means of an elaborated formalism [30], the author reexamines the notion of syntactic dependency (and its asymmetry or orientation), of the traditional sentence parts, etc. In this connection, some methodological questions concerning mathematical models of language, their syntax, semantics and pragmatics are discussed [29], and some proposed systems (Sørensen, Šaumjan) are critically reviewed by the same author.

Some basic problems of statistical linguistics are studied, too, cf. esp. the brief statement of the role of analytical and synthetical formulae in this area [28], and a critical analysis of L. Dolezel's model of stylistical encoding [30a].

II

In the domain of Machine Translation, our groups are concerned, first of all, with the translation from English to Czech. The history of our efforts in that respect is not a long one, a partial binary algorithm for English-to-Czech translation was prepared in the years 1958—59 and tested on the high-speed digital computer SAPO during January 1960 [16]. For the experiment, a short passage from the report by E. Fischer-Jörgensen at the 8th International Congress of Linguists in Oslo (1957) was chosen and several sentences were composed of the given vocabulary but in different configurations.

Another experiment of English-to-Czech machine translation has been prepared in cooperation with the Computing Machines Research Institute and is to be tested on the Czechoslovak Computer EPOS. Its algorithm is again based on a limited text (of about 40 sentences containing 400 words) dealing with electronics. The project contains some basic elements for an intermediate language, the general framework of which has been formulated [40]; [39] based on the similar assumptions as in the Czech generative system described above.

Our present activities are concentrated on applying the experience gained in previous experiments to a more general, not so strictly text-oriented translation program.

A dictionary for texts on electronics is being prepared with the aid of punch-card machines [9]. More than 100 000 running words from English texts chosen from the field of electronics were punched on punch-cards together with data of their grammatical characteristics and their respective Czech equivalents (from the corresponding Czech translation of the given text). Thus the frequency count of all words occurring in the analyzed texts has been gained not only as to their English forms but also as to the grammatical characteristics and Czech equivalents. This material is being studied in order to investigate the problems of multiple meaning as well as other questions of particular interest for MT. Since the English context has also been partly coded on the cards, the contextual criteria for choosing the particular Czech

equivalent of the English word can be studied on this material as well. As for the morphological parts of the algorithm of Czech synthesis, they have in principle been completed [17], [18], [38]. The algorithm of the synthesis of Czech declension has been tested on computer LGP-30. It has four stages:

1. the algorithm itself, which solves certain shifts in the type of declension that are not suitable for entering the dictionary (such as the change of "hard" declension into "soft" with the transition from positive to comparative by the same adjective, etc.), and some graphemic (phonological) alternations which need not burden the tables;
2. the tables for finding the required ending and stem-alternation according to the demanded form, and the tables of irregularities, where complete irregular forms are stored;
3. the dictionary of declension types and
4. the algorithm which classifies individual flexible words (except verbs) to the respective declension types (paradigms) according to certain features of their inflection. The function of the machine consists in forming the word-forms, when the stem and the declension type are given.

As to syntax, the algorithm of the synthesis of Czech will be formulated in connection with the above-mentioned system of generative description. Questions concerning the analysis of English are studied, mainly in the framework of predictive analysis, as proposed by the Computational Laboratory of Harvard University [14]; [15].
In connection with the research on MT, some problems of information retrieval have been examined, too [31].

III

In the Czechoslovak Academy of Sciences there also appears an increasing interest in mathematical linguistics. In 1961 a department of mathematical and applied linguistics was founded at the Institute of the Czech Language of the Czechoslovak Academy of Sciences, chiefly concerned with quantitative problems, the linguistic aspects of information theory, statistical stylistics, etc.; Shannon's method has been used there for the study of the entropy of the system of Czech graphemes [6] and a model

of stylistical encoding has been proposed [7], [8]. In 1962 a department of mathematical linguistics was also founded at the Institute of Slovak Language at the Slovak Academy of Sciences in Bratislava. It has a research program comprising the questions of quantitative and structural analysis of the Slovak language [10]. Serious attention is also being paid to the questions of mathematical linguistics by the department of linguistics of the Institute of Languages and Literatures of the Czechoslovak Academy of Sciences [20]; [21] and by the laboratory of mechanolinguistics at the Institute of the Czech Language, which has also been preparing a large application of punch-card machines in linguistic research [45], [46]. A group of mathematical linguistics is about to be founded at the Faculty of Philosophy of Brno University.

The questions of mathematical linguistics have been discussed at several national conferences and meetings on linguistics and at the conferences on cybernetics (1962, 1954, cf. the papers published in the Proceedings of the 1st Czechoslovak conference on cybernetics and the papers from the 2nd conference published in the journal *Kybernetika*). An international colloquium on algebraic linguistics was held in Prague in September 1964 [32]; the papers and reports submitted there were published in the journal *Kybernetika,* too.

The application of new methods in linguistics has also been drawing the attention of Czechoslovak mathematicians and logicians [1], [2], [9], [10], [11], [12]. At the Faculty of Philosophy of Charles University the two-year post-graduate course is being held for the third time, regular lectures and seminars of mathematical methods applied in linguistics are held at the University in Prague, Brno and Bratislava. The courses have been rendered a very important support through the cooperation with the Mathematical Institute of the Czechoslovak Academy of Sciences (K. Culík, Fr. Zítek) and with the Institute of the Theory of Information and Automation (J. Nedoma, L. Tondl).

A more comprehensive picture of the state of research in algebraic and quantitative linguistics is supplied by the first volume of *Prague Studies in Mathematical Linguistics,* published at the end of 1965 in Prague (edited by L. Dolezel, P. Sgall, J. Vachek), written in English and in Russian.

Literatur

[1] ČULÍK, K., Some Axiomatic Systems for Formal Grammars and Languages. Preprint of the Proceedings of the IFIP Congress 1962, München, 134—137.

[2] —, Использование абстрактной семаники и теории графов в многоязычных первеводых словарях. Проблемы кибернетики 13, 1965, 221-231.

[3] DANEŠ, F., Syntaktický model a syntaktický vzorec [Syntactic Model and Syntactic Pattern], in: Čs. přednášky pro V. mezinárodní sjezd slavistů v Sofii, Praha 1963, 115—124.

[4] —, A Three-Level Approach to Syntax, Travaux linguistiques de Prague 1, 1964, 225—240.

[5] DOKULIL, M., DANEŠ, F., K tzv. významové a mluvnické stavbě věty [On the So-Called Semantic and Grammatical Structure of Sentences], in: O vědeckém poznání soudobých jazyků, Praha 1958, 231—246.

[6] DOLEŽEL, L., Předběžný odhad entropie a redundance psané češtiny [A Preliminary Estimate of Entropy and Redundancy of Written Czech], Slovo a slovesnost 24, 1963, 165—175.

[7] —, Un modèle statistique du codage linguistique, Études de linguistique appliquée III, 1964, 51—63.

[8] —, Probabilistic Automation in a Model of Language Encoding, Preprints of 1965 Intern. Conference on Computational Linguistics, New York.

[9] HAJIČOVÁ, E., The Problems of MT Dictionary (prepared).

[10] HORECKY, J., K definici morfémy [On the Definition of Morpheme], Acta Universitatis Carolinae, Slavica Pragensia 4, 1962, 145—150.

[11] JAURISOVÁ, A., JAURIS, M., Užití teorie množin v jazykovědě [Application of the Set Theory in Linguistics], Slovo a slovesnost 21, 1960, 9 ff.

[12] —, Formal Methods for Syntactic Analysis, Acta Universitatis Carolinae, Philosophica et Historica 5, 1962, 17—30.

[13] JELINEK, J., BEČKÁ, V., TEŠITELOVA, M., Frekvence slov, slovních druhů a tvarů v českém jazyce [A Frequency Count of Words, Word Classes and Word Forms in Czech], Praha 1961.

[14] JELINEK, J., NEBESKY, L., A Syntactic Analyser of English, PBML[1] 2, 1964, 22—33.

[15] JELINEK, J., A Syntactic Analyser of English II, III, PBML 3, 38—59, PBML 4 (in print).

[16] KONEČNÁ, D., První pokus se strojovým překladem v Československu [The First Experiment of MT in Czechoslovakia], Naše řeč, 1960, 109 ff.

[17] —, Flexe českého slovesa z hlediska strojového překladu [The Flection of Czech Verb from the Point of View of MT], Acta Universitatis Carolinae, Slavica Pragensia 2, 1960, 85—96.

[18] —, Анализ форм чешсного глагол, PBML 2, 1964, 34—51.

[19] KORVASOVÁ, K., PALEK, B., Některé vlastnosti entropie českého slovníku [Some Features of the Entropy of Czech Dictionary], Slovo a slovesnost 23, 1962, 58—66.

[20] LEŠKA, O., KURIMSKY, A., in: Славянска филология, 1, Отловори на въпросиете за научната анкета по езикознаюие, Sofia 1963, 190—191

[21] - , относительно енропии в языке 1. PBML 2, 1965, 15—21.

[22] NEBESKY, L., SGALL, P., The Relation of "Form" and "Function", in: Language (Summary), PBML 1, 1964, 29—37.

[23] NEBESKY, L., On a Formal Grammar, PBML 1, 1964, 24—28.

[24] —,Об одной формализации разбора предложения, in: Математическая лингвистика. Москва 1964, 145—149.

[25] —, к одной модели анализа предложения, PBML 2, 1964, 3—10.

[26] —, ξ-Grammar, in: Prague Studies in Mathematical Linguistics, Praha (in print).

[27] NOVÁK, P., Některé otázky syntaktické analýzy (z hlediska strojového překladu) [Some Problems of Syntactic Analysis (From the Point of View of MT)], Slovo a slovesnost 23, 1962, 9—20.

[28] —, Two Types of Formulae in Quantitative Linguistics, PBML 2, 1964, 11—14.

[29] —, Mathematical Models of Linguistics Objects, in: Prague Studies in Mathematical Linguistics, Praha (in print).

[30] —, Dependency Theory in Syntax (A Formal Analysis) (prepared).

[1] PBML means *The Prague Bulletin of Mathematical Linguistics*, published by the respective groups at the Charles University in Prague.

[30a] —, K jednomu modelu stylistické složky jazykového kódování [On one model of the stylistic component of the language encoding], Slovo a slovesnost 27, 1966, 1 (in print).

[31] PALEK, B., некоторые лингвистические проблемы информационного Acta Universitatis Carolinae, Slavica Pragensia III, 1961, 197—208.

[32] —, Colloquium on Algebraic Linguistics and Machine Translation, Prague, PBML 2, 1964, 64—74.

[33] PANEVOVÁ, J., разбор електротехнических текстов, PBML 4 (in print).

[34] —, K otázkám druhé predikace v generativním systému [On the So-Called Second Predication in the Generative System], Acta Universitatis Carolinae, Slavica Pragensia 8 (in print).

[35] PITHA, P., SGALL, P., Koordinační konstrukce v generativním popisu [Coordinate Constructions in Generative Description], Acta Universitatis Carolinae, Slavica Pragensia 8 (in print).

[36] PITHA, P., K analýze koordinačních spojek při SP z češtiny [On the Analysis of Coordinate Conjunctions in Czech for MT], Acta Universitatis Carolinae, Slavica Pragensia 7, 87 ff.

[37] —, On the Programme of Coordinate Conjunctions in the Analysis of Czech, in: Prague Studies in Mathematical Linguistics, Praha (in print).

[38] SGALL, P., Soustava pádových koncovek v češtině [The System of Declension Endings in Czech], Acta Universitatis Carolinae, Slavica Pragensia 2, 1960, 85—96.

[39] —, К вопросу синтаксисе языка-посредника. материалы по математической лингвистике и машинному переводу Ленинград 1963, 86—91

[40] —, Převodní jazyk a teorie gramatiky [Intermediate Language and the Theory of Grammar], Slovo a slovesnost 24, 1963, 114—129 (an English summary of this article was published in Papers of the American Documentation Institute Convention, Chicago 1963).

[41] —, Zur Frage der Ebenen im Sprachsystem, Travaux linguistiques de Prague l, Praha 1964, 95—106.

[42] —, Generative Beschreibung und die Ebenen des Sprachsystems, submitted at the Magdeburg's Symposium "Zeichen und System der Sprache", 1964.

[43] —, Generation, Production and Translation, preprints of 1965 Intern. Conference on Computational Linguistics, New York.

[44] —, Generative Description and the Czech Declension (in print).

[45] ŠTINDLOVA, J., Sur le classement inverse des mots et sur ce qu'on appelle "dictionnaire inverse", Cahiers de lexicologie, Besançon, 1960, 2, 79—86.

[46] —, Stroje na zpracování informací a jejich význam pro jazykovědu [Information Processing Machines and their Significance for Linguistics], Slovo a slovesnost 22, 1961, 208—215.

[47] VACHEK, J. (ed.), A Prague School Reader in Linguistics, Bloomington 1964.