Pala, Karel

Mathematical linguistics at the ninth International congress of linguists

Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná. 1965, vol. 14, iss. A13, pp. 198-202

die wichtigsten und positivsten Züge der Entwicklung unserer Disziplin, die auch in den Verhandlungen des 9. Kongresses klar und deutlich zum Vorschein gekommen sind.

Adolf Erhart

Mathematical Linguistics at the Ninth International Congress of Linguists

1. The problems of mathematical linguistics were dealt with at the 9th International Congress of Linguistics in a special section, at whose meetings six papers were read. This figure is more apparent than real, since the process of assigning papers to this section was not altogether a thorough one. About twenty further papers including two of the lectures at the plenary session were of a character relating to mathematical linguistics and we shall consider several of them in this survey.

1.1 *Methodological Problems.* These problems were dealt with in the basic contribution by Spang-Hanssen, *Mathematical Linguistics — a Trend in Name or in Fact?* (61—67). Spang-Hanssen convincingly showed that the term „Mathematical Linguistics" (further ML) is not accurate and is too broad, since it may mean many different things. In the first place, we must realize that the antithesis ML and non-ML is artificial and in practice does not exist. There is only one science of linguistics, but in the course of solving its problems it makes use of various methods, both non-mathematical and mathematical.

In the opinion of Spang-Hanssen, if we use a mathematical apparatus (a model) in order to describe any linguistic phenomenon then we must distinguish: 1. Models whose nature is not axiomatically deductive and which are used in so-called quantitative (statistic) linquistics. Results gained by these methods have a numerical character (figures, e.g. the number of occurrences of a linguistic unit); 2. Models consisting of a certain set of axioms and conclusions, which can be directly interpreted in the concepts of empirical science, e.g. linguistics. Such models are supplied for example by the set theory. Spang-Hanssen considers that the application of axiomatic models is not so significant as the application of quantitative models and that the majority of the known structural approaches (Bloomfield, Hjelmslev, Chomsky) have recently only been rebaptized as ML.

He proposes the classification of research approaches into four types: quantitative and non-quantitative, each of which may be structural or non-structural. As Šaumjan pointed out in the discussion, and as contemporary development demonstrates, this classification is not altogether exact. More appropriate would appear to be the terminological distinction of quantitative linguistics, which includes the investigation of language by quantitative mathematical methods, i.e. by statistical methods, by the methods of probabilistic theory and of the information theory, and algebraic linguistics, including the investigation of language by methods of „non-quantitative mathematics", i.e. by methods of the theory of sets, mathematical logic, abstract algebra, the theory of automata, etc.[1]

Spang-Hanssen's assertions about the non-quantitative ML are, however, somewhat problematical. In the discussion Sigurd refuted Span-Hanssen's opinion on the possibilities and inadequacies of the set models. It can also be seen that non-quantitative ML depends not only on the theory of sets but also makes use of other mathematical disciplines, e.g. the theory of automata, the theory of algorithms, etc. Nor is it possible to agree with Spang-Hanssen when he places Bloomfield, Hjelmslev and Chomsky in the same group. It may perhaps be true of the first two, but not of Chomsky, who differs from his predecessors precisely because he consistently makes use of mathematical qualitative methods and because his work led to the development of the algebraic theory of grammar, which is today important both for linguistics and for mathematics (especially in the field of programming languages).

To sum up we may say that Spang-Hanssen's paper was very suggestive and showed that it is necessary to devote more attention to the methodology of linguistics in general.

2. In what follows we shall shortly refer to the papers in the field of quantitative linguistics, further to contributions from algebraic linguistics and in conclusion we shall deal separately with the lectures given in plenary session by N. D. Andrejev and N. Chomsky.

2.1 *Quantitative Linguistics.* The contribution of H. Kučera, *Statistical Determination of Isotopy* (713—721), though read in Section C (Application of Computers), nevertheless obviously belongs by its character to quantitative linguistics. The paper suggests a method of statistical phonological typology, in which the index of isotopy is used, i.e. a statistical measure based on the difference in the probability of appearence of comparable phonemes, and the measure of isomorphy, which is a measure of the phonological similarity of two phonemes from different languages based on the matrix of distinctive features of two phonological systems. The comparison was carried out on the material from current Czech and Russian (2 samples of 100,000 phonemes), the phonological transcript and further calculations were carried out with the IBM 7070 computer.

H. Karlgren's contribution, *Information Measures* (804—812) (according to the title in the list

of contents — in the text the title is *Information Estimates*) acquaints us with the investigation of language by methods of the information theory, carried out by a group of quantitative linguists in Stockholm. They carried out estimates of speech rates, counted the frequency of phonemes, letters, syllables, etc. in Finnish and Swedish, and statistically compared the length of translation with original, carried out prediction tests according to Shannon and also tests by an improved method, e.g. taking into account the effect of boundaries of words. Karlgren's paper is an inclusive and instructive survey of the possibility of applying methods of the information theory in linguistics.

A. R. Gammon, *A Statistical Study of English Syntax* (37—43), describes the statistical approach to problems of syntax, concretely, how to carry out segmentation of sentence on the basis of predictability of grammatical forms.

G. Herdan's contribution, *Mathematics of Genealogical Relationships between Languages* (51—58), is devoted to the assessment of the degree of relationships of languages on the basis of statistical ascertainment of terminological agreement; but was received with critical objections.

A number of other contributions belong to the field of quantitative linguistics, e.g. J. E. Grimes, *Measures of Linguistic Divergence* (44—50), L. Heilmann, *Statistical Considerations and Semantic Content* (427—432), and contributions from the field of lexicostatistics (glottochronology), with which for lack of space it is impossible to deal.

3. *Algebraic Linguistics.* It is characteristic that all the papers from this field were given in various sections, but non in that of ML where of course methodologically they belong. Thus the organizers of the Congress accepted the conception of ML as being equivalent to quantitative linguistics.

In his contribution *On the Fundamentals of Sentence Structure* (161—165), P. Siro endeavoured to indicate the very general and probably universal types of predicate in terms of the simple sentence model. The model is constructed axiomatically on a basis of two undefined concepts corresponding to the category of verbs and the category of substantives. All further simple sentence types can be deduced from this model. Siro basically uses a formal apparatus arising from the work of N. Chomsky.

In his paper *Mohawk Prefix Generation* (346—355), P. M. Postal attempts to show that the IC analysis is not sufficient for an adequate description of Mohawk sentences. Postal demonstrates that a description making use of the apparatus known from the generative grammar of Chomsky, i.e. a description using the explicit context-sensitive rules and transformation rules, is more adequate than a description based on IC analysis.

E. Bach's paper *Subcategories in Transformational Grammar* (672—678) represents an attempt to modify the theory of transformation grammar. It is a question above all of changes in the phrase-structure rules consisting of the introduction of upper and lower indices, the limiting of modifying lexical rules and a different placing of the vocabulary.

Schachter's contribution *Kernel and Non-Kernel Sentences in Transformational Grammar* (692—697) deals with the relationship between kernel and non-kernel sentences. Schachter showed that the relationships between these sentences are in some cases trivial, e.g. when some sentences can be derived in two ways: as kernel sentences or as non-kernel, using optional transformation. The choice between derivations is then based on criteria characteristic for the nature of language.[4]

Some further papers belong to the field of algebraic linguistics, e.g. W. S. — Y. Wang, *Some Syntactic Rules for Mandarin* (191—202) (a description of Chinese on the basis of the apparatus of generative grammars), a very interesting paper by Worth, *Suprasyntactics* (698—774), a contribution by K. Percival, *Word Order Rules in German* (600), and by R. P. Mitchell, *Properties of a Class of Categorical Grammars* (803).

Here too, though with some reservations, belongs the contribution of P. L. Garvin, *The Impact of Language Data Processing on Linguistics* (706—712) (further LDP), in which Garvin examines the part played by linguistics in LDP, i.e. on mechanical translation (MT), automatic linguistic analysis (ALA), information retrieval (IR), etc. Linguistic description must satisfy certain empirical requirements arising from LDP, which can be shortly formulated thus: 1. consistency, i.e. the requirement that linguistic information as the basic of a computer programme, should be explicitly formulated; 2. exhaustiveness, i.e. it is necessary for the linguistic description to be exhaustive both with regard to the machine vocabulary (the question of capacity of memory), and also with regard to the structure of the programme describing the language system, i.e. the programme must describe all the possible gramatical phenomena or else it must be possible to add them to the programme without difficulty; 3. simplicity, i.e, a requirement which can, e. g. be defined as the minimizing of the inventary of units or the minimizing of the number of rules. This requirement can however be defined exactly as an operational one, i.e.

with regard to the object we wish to attain, and in this sense we can also speak of effectiveness.

4. Further we wish to refer to the lectures of N. D. Andreyev and of N. Chomsky, which methodologically belong to the field of ML.

In his lecture on *Linguistic Aspects of Translation* (625—634), N. D. Andreyev poses six fundamental questions: 1. What has been contributed by machine translation (MT) to the general theory of translation? In comparing MT and human translation (HT) certain conclusions are reached: a human translator translates in such a way that he comprehends the input text and the output text, i.e. he correlates the text translated and the text arising from translation with his past and present, conscious and subconscious perception of reality. Andreyev terms this activity human heterolingual rendering (HHR). A machine translates in such a way that it turns from the input text to the ouput text without comprehending them, merely correlating the given text with the bi-codal dictionary stored in this memory, and with the indicated routine for transference from one code structure to the other. Andreyev terms the group of operations carried out by the machine translation. A beginner translates to a certain degree like a machine, and Andreyev terms this human translation. 2. What constitutes an invariant in the process of translation? An invariant in the course of translation is the numerical intermediary language (IL) specially constructed for the requirements of MT. IL allows us to explain the differences between MT and HHR; their invariants are quite different. With HHR the invariant is the message, which is rendered in two or more languages, a set of thoughts and concepts. With MT the invariant is the invariant text in IL, i.e. a certain string of numerical symbols in IL. If we compare the input or the output texts of natural languages — paralanguage (PL) — with the corresponding IL text, we can see the IL texts are not structurally identical with the text in PL, i.e., some elements of input PL are incongruent with regard to IL.

3. What are the methods of confronting the elements of different languages? According to Andreyev the space of a language has two axes (syntactic and paradigmatic) and three planes (morphological, syntactic and semantic.) This space is also included in IL and forms the basis on which we can develop the classification of incongruence., In IL the semantic units are semoglyphs, the syntactic relationships between them are explicitly expressed by tectoglyphs further relationships and morphological information are expressed by formoglyphs. This division enables us to classify incongruence in the whole space of language.

4. What are the ways of transition from input structures to output structures? A translation making use of IL has two basic phases: analysis, i.e. the transition from the input language to IL, and synthesis, i.e. the transition from IL to the output language, while input and output PL are described by means of a symbolic sign system, which is called the metalanguage. Each PL requires a special ML, while the IL is common to all.

5. What is the algoritmic linguo-typology? Andreyev suggests two algorithmic approaches (approximational, statistico-combinatorial), which enable us to examine the typological differences and agreements between languages. The most interesting of these is the algorithm of statistico-combinatorial modelling,[5] which works without any previous gramatical information, ascertains the type of language and analyses in detail the given language morphologically, syntactically and partly also semantically. The algorithm merely presupposes that the alphabet of the language analysed is given along with a sufficiently long text in the given language.

6. What is the future of translation? Andreyev sees the future of translation in the extension of translation on the basis of IL and working out a retrieval language (RL) which would be the logico-pragmatic code for information retrieval (IR) and sould serve for the accumulation of scientific information. Then it would be possible to carry out very quickly and with a very wide scope the translation and in working out of scientific and technical information (e.g. Pl → ML → IL → RL and in reverse). The discipline of translation also includes the formation of various kinds of languages designed for the communicational classes Man — Man, Machine — Machine, Machine — Man, Man — Machine.

A few remarks in conclusion: Andreyev's lecture on the one hand sums up the conception of the Leningrad MT group (1—4, partly 5), on the other it contains the announcement of future plans (6, partly 5), which so far can scarcely be discussed, until practical results are available. The algorithm of statistico-combinatorial modelling is particularly interesting, but at the same time it arouses several doubts, e.g. whether a purely statistico-combinatorial approach is sufficient and evident, or whether the heuristic processes can be completely formalized to such an extent.[6] So far only fragmentary reports of the practical testing of this algorithm are available, and the testing was carried out by hand, so that we are not yet entitled to come to final conclusions. As far as point 6 is concerned, the situation is still more complicated. So far only a few experiments have been carried out in the field of IR[8] and these were fundamentally not very successful. Other opinions of Andreyev are however confirmed, e.g. recently there have been very intensively

worked out languages of the type Man — Machine, i.e. programming languages for automatic computers (e.g. ALGOL, COBOL, FORTRAN, IPL, etc.).[9]

Great attention was paid to N. Chomsky's lecture *The Logical Basis of Linguistic Theory* (914—978).[10] In the first part, The Aims of Linguistic Theory, Chomsky first repeated his informal explanation of his conception of the algebraic theory of grammars already well known from his previous publications.[11] He explained the differences between two different models of generative grammars (GG), i.e. between the taxonomic model and transformational grammar (TG). The aim of the traditional grammars (to which TG is very close, to a certain extent it formalizes them) is to provide the user with the ability to understand at will any sentence in the given language, to form a sentence and use it correctly on suitable occasion, while relying completely on the language intuition and intelligence of the user of the grammar, who himself draws his own conclusions. The aim of linguistic theory according to Chomsky is (p. 923) "the precise specification of two kinds of abstract device, the first serving as a perceptual model and the second as a model for acquisition of language. The perception model A is a device that assigns a structural description D to presented utterance U, utilizing in the process its internalized generative grammar G, where G generates a phonetic representation R of U with structural description D ... The learning model B is a device which constructs a theory G (i.e. a generative grammar of a certain language) as its output, on the basis of primary linguistic data (e.g. specimens of parole) as input... We can think of general linguistic theory as an attempt to specify the character of the device B. We can regard a particular grammar as, in part, an attempt to specify the information available in principle (i.e. apart from limitations of attention memory, etc.) to A..."

The criteria for evaluation are given by three levels of adequacy of linguistic description: 1. level of observational adequacy (OA), reached by a grammar which correctly reflects the primary language data; 2. level of descriptive adequacy (DA) reached by a grammar which correctly reflects the linguistic intuition of the speaker and provides generalizations for the data observed which expresses the appropriate laws of the language; 3. level of explanatory adequacy (EA), attained by a linguistic theory which endeavours to provide a base independent on any language and enabling for the given language the choice of a GG, which would attain the level of descriptive adequacy.

On the basis of these criteria Chomsky shows that descriptive linguistics to a great extent dealt with the level of OA, whereas traditional grammars dealt with the DA level. The levels of adequacy are further examined in phonology, syntax and semantics. Only a few words are devoted to the question of linguistic comprehensiveness and objectivity. .

A great deal of space — almost half the lecture — is given to phonology. Here Chomsky fundamentally rejects "classical" phonology (phonemics), characterizing it as taxonomic and asserting that for GG attaining the DA level only the so-called systematic phonetics and systematic phonemics can be considered (i.e. fundamentally morphonemics). The substantiation of this is very exhaustive and supported by many examples. In conclusion Chomsky deals rather shortly with the question of perceptive and acquisition models. These questions belong partly to theoretical psychology, but it can be seen that only what is adequate from the linguistic point of view can be of any interest to psychology.

In this brief survey we can scarcely deal with Chomsky's exhaustive lecture in detail, and so we can made only a few fundamental remarks. First of all we must take into account the fact that Chomsky has recently considerably changed his conception.[12] According to the new conception the GG of any language contains three components: syntactic, semantic and phonological (the former GG were composed of syntactic and phonological components). The last two elements are purely interpretive. The syntactic component contains the base and the transformational subcomponent and its recursive rules are the source of the infinite generative capacity of the grammar. The base generates deep structures which enter into the semantic component, receive a semantic interpretation and by means of transformational rules are mapped onto surface structures, which are interpreted in the phonologic component. GG now contains in addition the semantic component, which interprets sentences semantically (ascertaining their semantic homonymity, synonymity or anomaly).[13] In the new conception, too, a considerable change has taken place in the role of the transformational rules and the transformational subcomponent, which now contains only singular transformation (as compared to the former generalized transformations) and whose role is to filter out the incorrectly formed deep structures generated by the base.

We must, however, ask ourselves the question, why did Chomsky change his conception so fundamentally? It seems that three groups of reasons operated here: 1. The former GG without the semantic element did not describe or explain adequately the structure of the language being

unable to deal completely with the semantic properties of the transformations; 2. There were difficulties with the formalization and the formal complexity of the generalized transformations and transformation markers, which in any case have never been solved completely; 3. Psychological aspects, clearly explained by Katz.[14] In conclusion it must however be remarked that the justification of the new conception of GG will be best demonstrated by the construction of the GG of a concrete natural language. The current stage of development of GG is so far characterized by a great number of theoretical deliberations and "indicated" GG, but an infinitesimal percentage of concrete work[15], which would tend to the formation of concrete GG, the adequacy of which could be controlled experimentally. It seems however that we have been waiting too long for such a concrete GG.

To sum up we may say that the papers in the field of ML given at the 9th Linguistic Congress demonstrated the vitality and fruitfulness of mathematical methods in contemporary linguistics and that the further development of linguistics including its relationship to other sciences, is inconceivable without new methological approaches.

Notes

1. P. Novák, *O terminologii matematické lingvistiky*, (On the Terminology of Mathematical Linguistics.) Čs. terminologický časopis 2, 1963, 234 n.
2. See e.g. N. Chomsky, *On Certain Formal Properties of Grammars*, Information and Control 2, 1959, 137—167; also *Syntactic Structures*, 's Gravenhage 1957, *Three Models for the Description of Language*, IRE Trans., vol. IT-2, No 3, 1956, 113—124, and several others.
3. Gammon's approach has several features in common with that of Andreyev, see n. 5.
4. The problems dealt with in Schachter's paper, and in Bach's too, are deal with distinctly differently in Chomsky's new conception, see op. cit. n. 12.
5. N. D. Andreyev, *Algoritmi statisticko-kombinatornogo modelirovaniya morfologii, sintaksisa i semantiki*, Materiali po matem. lingvistike i mashinnomu perevodu, II, Leningrad, 1963, 3—44.
6. B. V. Sukhotin, *Algoritmi lingvisticheskoy deshifrovki*, Problemi strukturnoy lingvistiki, Moskva, 1963, 75—101.
7. L. D. Andreyeva, *Statisticko-kombinatornoye videleniye paradigmi pervogo morfologicheskogo tipa v russkom yazike*, loc. cit. n. 5, 45—63.
8. J. W. Perry, A. Kent, *Tools for Machine Literature Searching*, New York, 1958; K. Čulík, B. Palek, *Automatizace referování*, (Automatisation of Information Retrieval), Metodika a technika informací, 1962, No. 3—4, 65 n.; some very interesting, even though very simple results are given by L. E. Pshenichnaya, E. F. Skorochoďko, *Sintez osmislennikh predlozhenii na ECVM*, Problemi kibernetiki 10, Moskva, 1963, 261—275.
9. See e.g. *Annual Review of Automatic Programming I, II, III*, Pergamon Press, Oxford, London, New York, Paris.
10. A corrected and expanded version of this lecture has now been published in the volume *The Structure of Language* (Readings in Philosophy of Language), ed. by J. A. Fodor, J. J. Katz, Prentice Hall Inc., Englewood Cliffs 1964, under the title *Current Issues in Linguistic Theory*, 50—118, We were unable to refer to this since it did not reach us until this report was finished.
11. See the work quoted in n. 2.
12. N. Chomsky, *Categories and Relations in Syntactic Theory*, M. I. T. 1964 (a paper sent to the conference on "Sign and System in Language" held in September 1964 in Magdeburg; further also J. J. Katz, P. M. Postal, *An Integrated Theory of Linguistic Desriptions*, M. I. T. Press, Cambridge, Mass,. 1964; P. M. Postal, *Nový vývoj teorie transformační gramatiky*, (New Development in the Theory of Tranformational Grammar), translated from English, SaS 26, 1965, 1—13.
13. J. J. Katz, J. A. Fodor, *The Structure of a Semantic Theory*, Language 39, 1963, 170—210.
14. J. J. Katz, *Mentalism in Linguistics*, Language 40, 1964, 124—137.
15. An honourable exception here is R. B. Lees, *The Grammar of English Nominalizations*, Baltimore, 1960.

Karel Pala

Dialectology and Linguistic Geography at the Ninth International Linguistic Congress

It is a satisfactory feature that so much time at the congress and so much space in the report were devoted to problems of dialectology, and not only to dialectology itself but also to linguistic geography and to languages in contact. The contribution by Pavle Ivić, *Structure and Typology of Dialectical Differentiation* (113—121) was the most provocative of discussion. The author endeavoured to determine features which are quantitative and thus measurable: 1. the differentiation density of the dialect, 2. the linear distribution of isoglosses (equal distances — a bundle