JPRS: 3502

13 July 1960

SOVIET DEVELOPMENTS IN INFORMATION PROCESSING AND MACHINE TRANSLATION

Photocopies of this report may be purchased from:

PHOTODUPLICATION SERVICE
LIBRARY OF CONGRESS
WASHINGTON 25, D. C.

U. S. JOINT PUBLICATIONS RESEARCH SERVICE
205 EAST 42nd STREET, SUITE 300
NEW YORK 17, N. Y.

F O R E W A R D

This publication was prepared under contract by the UNITED STATES JOINT PUBLICATIONS RESEARCH SERVICE, a federal government organization established to service the translation and research needs of the various government departments.

JPRS: 3502

CSO: 3901-D/15

SOVIET DEVELOPMENTS IN INFORMATION PROCESSING AND MACHINE TRANSLATION

FOREWORD

This translation series presents information from Soviet literature on developments in the following fields in information processing and machine translation: organization, storage and retrieval of information; coding; programming; character and pattern recognition; logical design of information and translation machines; linguistic analysis with machine translation application; mathematical and applied linguistics; machine translation studies. The series is published as an aid to U. S. Government research.

Previously issued JPRS reports on this subject include:

JPRS: 68, 241, 319, 355, 379, 387, 487, 621, 646, 662, 705, 729, 863, 893, 925, 991, 992, 1006, 1029, 1130, 1131 and 1132, 1133, 3225, 3300, 3356 and 3433.

TABLE OF CONTENTS

VII. PROBLEMS OF MACHINE TRANSLATION

Slovo a slovesnost, No 3,
1959, pages 208-210

Petr Sgall

One subsection meeting at the congress was devoted to problems of machine translation. The unexpectedly large number of delegates participating and the wide discussion demonstrated the interest with which Soviet and other linguists are following these questions.

The first report was given by the American delegate L. Michelsen of the University of Washington (cf. Russian-English MT, American Contributions to the Fourth International Congress of Slavicists, Moutin and Co., 's-Gravenhage 1955, 21 pp). He pointed out the advantages of the procedure chosen for working on automatic translation from Russian to English at this university, wherein most attention is directed to lexical questions. Work is also being done on the solution of very complicated problems of polysemy, and ways are being sought for solving these problems mechanically, on the basis of a large starting dictionary, without laborious and slow collection and examination of individual documents. At the same time, of course, until these problems have been worked on, the machine prints out many meanings for each instance. For example the Russian phrase SREDI NAYBOLEE VAZHNYKH NEDOSTATKOV V DEYATEL'NOSTI was translated into English thus: Among/in/middle most (of) important of defects/lacks in/to/at/on/of/like activity(s). In conclusion the speaker also mentioned new technical developments permitting better utilization of an automatic computer for translation purposes, particularly King's photoscopic memory.

V. Yu Rozentsveyg, of the Institute of Foreign Languages in Moscow, is chairman of the Soviet Society for Machine Translation. (The Society publishes a mimeographed bulletin containing reports on various current linguistic problems as well as work on problems of machine translation.) In his report Rozentsveyg pointed out that no existing linguistic methods -- even descriptive techniques, which are closest to these problems -- are able to analyze linguistic facts with a precision sufficient for the requirements of machine translation. Most Soviet workers therefore approach these problems from new theoretical angles. Linguistic phenomena are handled by the methods of set theory (Cf. O. S. Kulagina, A method of defining grammatical concepts on the basis of set theory, in Problemy kibernetiki I /Problems of Cybernetics I/, edited by A. A. Lyapunov, Moscow, 1958, pp 203-214) and other mathematical procedures, in order to set up unambiguous analytic criteria. Many institutes of the Academy of Sciences USSR have worked out sets of rules for machine translation. (Such institutes include the Mathematics Institute imeni Steklov; the Institute of Linguistics; the Laboratory of Electromodeling of the All-Union Institute of Scientific and Technical Information; the Institute of Precision Mechanics and Calculating Techniques where, unlike other Soviet centers, traditional methods are

used and principal attention paid to detailed lexical analysis of the linguistic material; the university in Leningrad, with its own machine-translation laboratory; the university in Gor'kiy; and institutes of the Academies of Sciences of the Armenian and Georgian SSR's.) These rules concern translation from various languages (English, French, Chinese, Japanese, Hungarian) to Russian. Work is being done on other languages, and also on translation from Russian into other languages, all on the basis of technical text. There is also some prospect of mechanizing the linguistic processing of the program for machine translation. Problems of machine translation are bringing Soviet linguists into the overall picture of automation of intellectual processes. They are aware that mathematical processing of language is important not only for machine translation but also for the operation of information machines, problems of coding, etc.

A young worker at the Linguistic Institute, Academy of Sciences USSR, I. A. Mel'chuk, gave an (unprinted) report on progress in his institute, and concentrated his remarks on a method which means not merely preparing machine translation from one language to another, but will permit translation from any language to any other language through an artificial auxiliary language. This intermediate language (yazyk-prevodnik) is envisioned as an abstract system of relations, and its model is made up empirically from a maximum number of languages. This means the careful determination and classification of various syntactic relations and methods of expressing them in various languages, followed later by similar work on wide-ranging semantic problems. (Since the Soviet linguists have not yet included Czech in their program, our group of machine-translation workers accepted their suggestion that we work on some problems of Czech from this point of view.)

In the discussion N. D. Andreyev acquainted those present with a somewhat different procedure used at Leningrad University: the workers there see the auxiliary intermediate language not as an abstract system of similarities but as a real language with its own structure. The ultimate goal here is the formation of a real machine language which can be expressed by various symbols (numbers) and, after the phonological aspect of language has been mastered, by sounds as well. By comparing the most divergent languages the Leningrad group is attempting to arrive at the most general methods. The 80 collaborators are engaged in setting up the rules for machine translation from more than 20 languages.

Further discussion showed the great importance of machine-translation theory for linguistics. Professor A. A. Reformatskiy emphasized the importance of this new field for linguistic typology. V. V. Ivanov stated that work on an auxiliary intermediate language will mean for synchronic problems of typology what the discovery of systematic similarities among related languages (the reconstruction of a proto-language) meant for historical and comparative linguistics. The great opportunities for applying mathematical methods in linguistics were mentioned by A. S. Chikobava, who also said that the material nature of the linguistic

sign is important for this "mathematization of linguistics." I. I. Revzin stressed the greater difficulty of English-Russian than Russian-English machine translation (because of the richer morphology of Russian) and the fact that the theory of machine translation also helps to clarify the concept of language as a set of subsystems. T. M. Nikolayeva mentioned, among other items, the possibility of a special artificial intermediate language for a group of related languages, such as the Slavic languages. After hearing Professor Finkel, who warned against overestimating the new methods and stressed the impossibility of machine translation of poetry, the chairman, Professor P. S. Kuznetsov, closed the meeting with a few words on the value of work on machine translation and related disciplines, which are giving linguistics a number of new ideas resulting from the union of theoretical investigation with practical requirements.

The meeting demonstrated the great range and basic importance of the problems which linguistics faces through the new methods which are developing so quickly in connection with the requirements of machine translation. The statements of numerous young linguists, principally from the USSR, showed that with only brief experience in these problems they are working with great energy and many brilliant new ideas, and using various working methods, which permit them a broad over-all view of these new problems.

MACHINE TRANSLATION IN THE USSR

(EXPERIENCE AT THE CONFERENCE ON MATHEMATICAL LINGUISTICS IN LENINGRAD IN 1959)

Slovo a slovesnost, No 4, 1959, pages 277-285

Bohumil Palek

Applied linguistics is a very important linguistic discipline today, and its importance is growing year by year. An important position in it is occupied by machine translation, not only because of its economic significance but also because it has a great influence on linguistic research. It requires of linguists a careful classification of all linguistic phenomena and permits conclusions which would otherwise be impossible. Its importance was also emphasized by the fact that at the Eighth International Linguistic Congress in Oslo in 1957 and at the Fourth International Slavistic Congress in Moscow several reports were devoted to this field. (Cf. B. Havranek and K. Horalek, The Eighth International Congress of Linguists in Oslo, Slovo a slovesnost, 19, 1958, p 47; and Petr Sgall, Problems of Machine Translation, The Fourth International Congress of Slavists in Moscow, Slovo a slovesnost, 20, 1959, p 208-210.) Furthermore several special journals are now being published: Mechanical Translation (MT); Byulleten' ob"yedineniya po mashinnomu perevodu /Bulletin of the Society for Machine Translation/; and the publication of the International Journal of Machine Translation and Mathematical Linguistics is in preparation. And a number of symposia have been published: Byulleten' ob"yedineniya po mashinnomu perevodu 1-7 (BMP 1-7); Mashinnyy perevod /Machine Translation/, a symposium of articles on machine translation, Institute of Precision Mechanics and Calculating Techniques, Acad Sci USSR, Moscow 1958 (MP); Materialy po mashinnomu perevodu I /Material on Machine Translation I/, published by Leningrad University, Leningrad 1958 (MMP I); Voprosy statistiki rechi /Problems of Language Statistics/ (report material), published by Leningrad University, Leningrad, 1958 (VSR); Problemy kibernetiki I /Problems of Cybernetics I/, published by the Acad Sci USSR, Moscow, 1958 (PK I); Tezisy konferentsii po mashinnomu perevodu /Papers from the Conference on Machine Translation/ (15-21 May 1958), I MGPIIYa /First Moscow State Pedagogical Institute of Foreign Languages/, Moscow, 1958 (Tez. I); Tezisy soveshchaniya po matematicheskoy lingvistike /Papers from the Conference on Mathematical Linguistics/ (15-21 April 1959), Leningrad University, Committee for Applied Linguistics, Leningrad, 1959 (Tez. II).

The conferences held almost every year in the USA, and recently in the USSR as well, are very important for the development of machine translation.

The two most recent congresses of the CPSU have stressed the development of automation and new techniques as a basic and important factor in the economy. This has of course affected work on machine translation. The Soviet specialists have received full support and recognition in their work.

Work on machine translation in the USSR was undertaken in 1955 in the Institute of Precision Mechanics and Calculating Techniques (ITM i VT) and in the Mathematics Institute imeni Steklov of the Acad Sci USSR. English-Russian translation was worked on during 1955 and by the end of that year the first translation was done on a BESM machine. It was only later that the Linguistic Institute and the First Moscow State Pedagogical Institute of Foreign Languages (I MGPIIYa) began work. Several centers are now working in the Soviet Union. One of the largest is the ITM i VT center, where technicians are working with a large group of linguists on machine translation from English, Chinese, Japanese, and German to Russian.

This group is interested principally in the practical requirements of machine translation and in the detailed analysis of languages. Another group is centered around the Mathematics Institute imeni Steklov of the Acad Sci USSR. It is working on theoretical problems, investigating new descriptions of language, and performing practical activity. Under the direction of A. A. Lyapunov algorithms have been worked out for French, English, and Hungarian (see details below in Mel'chuk's article). Collaborating with this group is another from the Linguistic Institute, directed by A. A. Reformatskiy. And they are in contact with workers in the Laboratory of Electromodeling of the All-Union Institute of Scientific and Technical Information, Council of Ministers USSR. In developing algorithms efforts are made that programming be standardized and automated. In addition it is being proposed that the compilation of algorithms be automated such that the computer will receive a relatively long text with a general program according to which the algorithm of the particular language will be set up automatically. In the Leningrad machine-translation laboratory, under N. D. Andreyev, thanks to the well-developed oriental studies in Leningrad, work is being done mainly from oriental languages -- Arabic, Turkish, Indonesian, Hindi, Japanese, Burmese; but work is also being done in European languages such as German, Spanish, etc. Intensive work is being done on the problems connected with an intermediate language, models of language, and information languages In addition to these large centers, which are issuing symposia and have gone far in machine translation, new groups have developed whose work is just beginning: work has begun at the universities in Gor'kiy, Yerevan, and Tbilisi. There is interest in other centers, too, in machine translation; e.g., in Kiev they are interested in Slavic languages, particularly Czech.

The development of applied linguistics can also be seen in the extent to which it has been represented in conferences on cybernetics and those specifically on machine translation. The first mention of machine translation was made at a meeting of the Academy of Sciences USSR on the automation of production, held 15-20 October 1956. A report given by D. Y. Panov (Panov, A. A. Lyapunov, I. S. Mukhin, Automation of Translation from One Language to Another, symposium from the meeting of the Academy of Sciences USSR on Scientific Problems of Production, 15-20 October 1956, Plenarnyye zasedaniya /Plenary Meetings/,

Moscow, 1956) dealt with the general problems of machine translation. In the intensive development after 1956 applied linguistics occupied an important place at subsequent cybernetic conferences (Cf. I. A. Mel'chuk, Conferences on Problems of the Development and Construction of Information Machines, VYaz, 1957, No 5, p 161; Scientific-Technical Conferences on Cybernetics, PK I, p 266). The meeting held 28-31 May 1957 in Moscow was attended by more than 500 specialists. It was organized by the Laboratory of Electromodeling, Acad Sci USSR. Mathematical linguistics was discussed at the plenary meeting: a report by V. V. Ivanov on the creation of a language suitable for the machine, on the languages of science, and on the need for their standardization; and a report by V. N. Uspenskiy on the logical and mathematical problems connected with machine translation. The theoretical section was devoted primarily to mathematical linguistics: 21 papers out of 22. The most varied subjects in mathematical linguistics were discussed, from general problems of machine translation (A. A. Lyapunov) through information languages (V. K. Finn) and methods of investigating language and creating a grammatical scheme (O. S. Kulagina) to concrete problems of the "metalanguage" of chemistry (G. M. Vleduts, V. K. Finn) and geometry (N. M. Yerolayeva, Y. A. Shikhanovich) and setting up algorithms for individual languages (T. N. Moloshnaya, L. A. Bezzabotna).

Since that meeting it has been clear that mathematical linguistics would continue to be important. The cybernetic seminars, held at Moscow University since 1955, had great influence on this development. Similar seminars were begun a year later at Leningrad University. An important position in mathematical linguistics is occupied by statistics. On 1-4 October 1957 a meeting was held in Leningrad on language statistics (cf. V. A. Uspenskiy, Conference on the Statistics of Language, VYaz 1958, No 1, p 170). It was attended by representatives of various academic and military institutions. There were two questions on the program: (1) The use of statistical methods in analyzing the spoken and written language, and the development of problems connected with the transfer of information; and (2) The relationship between structural and statistical methods in linguistics. An important factor in the coordination of work in mathematical linguistics in the USSR was the formation of the Committee on Applied Linguistics, Acad Sci USSR, in January 1958 (cf. V. V. Ivanov, Committee on Applied Linguistics, VYaz, 1958, No 3, p 136). In addition to such outstanding linguists as L. R. Zinder, A. A. Reformatskiy, V. Yu. Rozentsveyg, V. V. Ivanov, N. D. Andreyev, and others, committee members include military specialists such as V. I. Medvedev, etc.; mathematicians and technicians R. L. Dobrushin, Y.S. Bykov, L. A. Varshavskiy; psychologists and neurologists A. R. Luriya, I. I. Zhinkin, etc. Chairman is the Leningrad phonetician L. R. Zinder. Committee activity is divided into seven sections of applied linguistics: linguistic statistics, the application of mathematical methods in linguistics, the development of algorithms for machine translation, experimental phonetics, physiological and psychological methods in applied linguistics, transcription and transliteration, and artificial languages

and coding. Problems of the comprehensibility of speech were dealt with in a symposium on 26-30 May 1958 in Moscow (cf. L. R. Zinder, Symposium on Problems of Language Comprehensibility, VYaz, 1958, No 5, p 151), attended also by the Czech specialists J. Vachek and J. Slavik, whose papers were received with enthusiasm by the Soviet specialists.

The development of mathematical linguistics is clearest in machine translation. One can learn something of its development from the conferences held in Moscow in 1958 and Leningrad in 1959.

The first conference on machine translation was held in 1958 by the I MGPIIYa. It was attended by 340 specialists from various centers, and 70 papers were read (cf. V. V. Ivanov, Conference on Machine Translation VYaz, 1958, No 5, p 149). On 15 and 16 May 1958 a plenary meeting was held at which the most important papers were read. From 17-20 May the conference continued in two sections: theoretical, and on machine-translation algorithms. At the plenary meeting general problems of mathematical linguistics were discussed. It was emphasized that mathematical linguistics is important for modern linguistics principally because it requires a sharpening of linguistic concepts. On the other hand linguists must know how to apply mathematical methods to language. The plenary session also heard papers from outstanding representatives of individual centers working on machine translation. In the theoretical section various papers were given. This section analyzed problems of phonology, semantics, the relation of structuralism to machine translation, etc. The conference resolution emphasized the importance of machine translation for the national economy and technical progress, which is one of the most important points made at the XXth Congress of the CPSU. It was further pointed out that mathematical linguistics is very important for solving problems of the transfer of information, compiling grammars, etc.

The conference decided that an All-Union Conference on Mathematical Linguistics should be held in 1960. This decision was carried out one year earlier. From 15-21 April 1959 an unusually successful conference was held in Leningrad. It was also attended by a four-member group of machine-translation workers from Czechoslovakia: P. Sgall and P. Novak spoke on the Prague language typology and on its relationship to language models. Other countries represented at the conference were China, Rumania, and Poland. A total of 486 workers from various regions of the Soviet Union were present. In addition to such outstanding scientists as Academician S. L. Sobolev, Academician B. A. Larin, Corresponding Member A. A. Markov, Corresponding Member L. D. Kantorovich, Professors Zinder, Reformatskiy, Rozentsveyg, Olderoge, etc., most people at the conference were young workers. Professor A. A. Lyapunov, the outstanding Soviet specialist in computers and cybernetics, was unfortunately unable to attend the conference.

The conference was lively and produced many new ideas. It was concentrated mainly on these subjects: mathematical models of language, the application of statistical methods in the investigation of language, intermediate and information languages, setting up programs, and technical equipment capable of processing language. The papers delivered at

the conference went beyond these subjects. A characteristic of the entire conference was the effort to engage the linguist fully in developing automatic reading equipment, automatic abstracting of scientific and technical literature, the use of information languages, the vocal control of machinery, and particularly the practical use of machine translation. This requires the linguist to work vigorously. The conference also evaluated past results in the structural examination of language, and in its resolution asked that the structural view of language be connected with the use of mathematical methods. It emphasized the importance of training personnel for machine translation and decided to recommend to the Ministry of Higher Education that at all institutes of higher education teaching linguistics lectures also be given in mathematics with linguistic implications. The resolution also asked that machine-translation centers be set up in Gor'kiy, Tbilisi, and Yerevan, and that mathematical linguistics and machine translation be introduced as specialties for students. The conference showed clearly that mathematical linguistics is a mature discipline in the Soviet Union. Machine translation and related problems are attracting large numbers of young workers, whose efforts will surely lead to important results.

The first work on machine translation published in the USSR was of a general and informative nature. It was a necessary basis for further systematic work on the one hand and for propagating machine translation on the other. Articles published in the journal Voprosy yazykoznaniya /Problems of Linguistics/ and in the first numbers of BMB were very important (cf. V. P. Berkov and B. A. Yershov, On Attempts at Machine Translation, VYaz, 1955, No 6, p 145; I. S. Mukhin, Experience in Automatic Translation on the BESM Electronic Computer, Moscow, 1956; On Some Current Problems of Modern Linguistics, VYaz, 1956, No 4, p 3; P. S. Kuznetsov, A. A. Lyapunov, and A. A. Reformatskiy, Basic Problems of Machine Translation, VYaz, 1956, No 5, p 107; L. I. Zhirkov, The Limits of Applicability of Machine Translation, VYaz, 1956, No 5, p 121; T. N. Moloshnaya, V. A. Purto, I. I. Revzin, V. Yu. Rozentsveyg, Some Linguistic Problems of Machine Translation, VYaz, 1957, No 1, p 107; L. S. Barchudarov, G. V. Kolshanskiy, Problems of the Feasibility of Machine Translation, VYaz, 1958, No 1, p 129; I. K. Bel'skaya, Some General Problems of Machine Translation, MP, p 3; M. I. Steblin-Kamenskiy, The Importance of Machine Translation for Linguistics, MMP I, p 3; A. A. Reformatskiy, The Problems of Machine Translation of Oral Speech, BMP, 6 p 31; I. I. Revzin, Some Aspects of Modern Theoretical Investigations in Machine Translation, BMP, 7, p 13; V. Yu. Rozentsveyg, Work on Machine Translation from Foreign Languages into Russian, and from Russian into Foreign Languages in the USSR, IV Mezhd. s"yezd slavistov, Moscow, 1958; I. A. Mel'chuk, Work on Machine Translation in the USSR, Vestnik Akademii nauk SSSR, 1959, 2 p 43; T. L. Gavrilova, O. S. Kulagina, T. N. Moloshnaya, G. V. Chekova, Translation from One Language to Another by Machine, Materialy konferentsii po teorii informatsii, Moscow, 1958. The most basic information on machine translation

is found in the brochure of D. Yu. Panov, Avtomaticheskiy perevod /Automatic Translation/, Moscow, 1956, which came out in a second, enlarged edition in 1958. For an idea of electronic computers the brochure by S. A. Lebedev, Elektronnye vychislitel'nyye mashiny /Electronic Computers/ is very informative). In addition to popular studies and articles on the general problems of machine translation a large number of articles has now been published specializing in certain sectors of machine translation.

In developing the theory of machine translation it was noticed that binary translation from one language to another (e.g., from Czech to English) is possible by means of special algorithms, i.e., the algorithm for English-Czech translation would be different from the algorithm for Czech-English translation. Considering the number of languages which should be translated, one can see the large number of such binary translations. This has led to the effort to form an intermediate language permitting translation from any language to any other language.

This problem has been attacked in the Soviet Union by N. D. Andreyev and I. A. Mel'chuk. Andreyev published an article (Machine Translation and the Problem of the Intermediate Language, VYaz, 1957, No 5, p 117) in which he discusses the semantic problem of the intermediate language, and particularly the problem of polysemy, which plays an important role in translating from a large number of languages. N. D. Andreyev believes that different world languages have different semantics; on the basis of this difference he suggests forming semantic groups subject to certain rules. These rules would permit changing the semantics of one language into the semantics of the other. These semantic groups would be established for a large number of languages. In addition to semantic groups the dictionary would contain frequency data. In the frequency system there would be a separation of words which have special uses, words used in technical text, and words in general use. In translating the machine would obtain information on the type of text being translated, and on this basis would choose first the technical words and then the words in general use. This type of progressive translation would, in N. D. Andreyev's view, largely solve the problem of polysemy. Andreyev also proposes using the semantics of standard and high-frequency suffixes. Comparison of various suffixes in various languages will make it possible to determine the semantic categories to which the individual suffixes belong. Andreyev also spoke on the intermediate language at the Leningrad conference (cf. N. D. Andreyev and S. Ya. Fitialov, An Intermediate Language for Machine Translation and the Principles of Its Construction, Tez II, p 53.) Here he clearly expressed his views of the intermediate language as a concrete language with its own morphology, lexicon, and syntax. A symbolically written intermediate language would, in Andreyev's opinion, be useful also as an information language.

Andreyev's concept differs from that of I. A. Mel'chuk, who discussed the problem of an intermediate language at last year's Moscow conference on machine translation and also at this year's conference in Leningrad (cf. Mel'chuk, A Model of an Intermediate Language for Machine Translation, Tez I, p 20; and Mel'chuk, The Problem of Grammar in an Intermediate Language, Tez II, p 60). Mel'chuk's approach to the problem is characteristic of the group around the Linguistic Institute of the Academy of Sciences USSR. Mel'chuk believes that an intermediate language is a system of mutual relations among individual languages. These relations apply on three levels: lexical, morphological, and syntactic. In an intermediate language, as Mel'chuk sees it, only the lexicon (with the addition of word-forming units) and syntax apply. This means that the intermediate language will express differences between individual languages. Morphological categories, insofar as they are needed (time, number) will be words in the intermediate language. (cf. also article, below, by Mel'chuk.)

When discussion started concerning an intermediate language the question arose whether it would be possible to use some natural or artificial language as an intermediate language. This question was discussed at the first conference on machine translation by Ye. A. Bokarev (cf. Bokarev, The Intermediate Language and Artificial International Languages, Tez I, p 4). He poses the problem soberly, unlike some fervent proponents of artificial languages. He believes that it should be possible to investigate an intermediate language parallel with artificial languages, and to transfer from one language to another the experience gained in studying both systems. The group around the ITM i VT wants to work with Russian since this is the language into which most translation is done. So far it is difficult to say which of these procedures is best, but each of them produces new information on an intermediate language.

Intensive work is also being done on the problem of the "languages" of various scientific disciplines, including linguistics, which is closely connected with the problems of an intermediate language. Since machine translation will be used principally in translating technical text, it would be desirable to standardize the languages of individual branches of science, to avoid ambiguous interpretations, etc. This simultaneously means a contribution to each scientific discipline, since it will be necessary to formulate views of scientific problems clearly and unambiguously (cf. V. V. Ivanov, Linguistic Problems of Creating Machine Translation for an Information Machine, MMP I, p 10; N. D. Andreyev, A Meta-Language for Machine Translation and Its Application, MMP I, p 40). The question of an intermediate language is associated with that of an information language, i.e., a language permitting the storage of information in the computer. At the Leningrad conference Berkov discussed the difference between an information and an intermediate language (cf. V. P. Berkov, Grammatical Information and an Information Language, Tez II, p 68). He believes the difference to

lie in the fact that an intermediate language must be maximally suitable for translating from many languages and must have a minimum bulk. A characteristic of an information language is that it contains a minimum of information, unlike natural languages, whose grammatical structure always contains a large percentage of so-called redundancy. In this connection work is being done on other problems connected with the efficient use of automatic computers, particularly automatic abstracting (cf. V. A. Agrayev, and V. V. Borodin, The Problem of Automatic Abstracting and Methods of Solving It, Tez II, p 88). In this case the goal is to make the machine itself able to abstract any scientific article, to produce a short summary sufficient for those interested. Work is being done, too, on the recognition of written and spoken text, etc. (cf. N. D. Andreyev, Principles of Building Electronic Reading Equipment, MMP I, p 223).

A large number of problems is connected with the dictionary, particularly its system in machine translation. The problem has been discussed at length, principally by I. L. Bratchikov, S. Ya. Fitialov, and G. S. Tseytin (cf. their The Structure of a Dictionary and Information Coding for Machine Translation, MMP I, p 61). In their article they examine various methods of compiling a dictionary and seek the best. A dictionary for machine translation must contain the stems of all words. But many words have stem variants, which must then be entered several times (palatalization in the paradigm, epenthetic vowels, etc.). Furthermore endings must also be included in the dictionary. The main difficulties arise in comparing shades of meaning of individual words; for some, entire phrases or idioms must be included. Organized the usual way such a dictionary would take too much space in machine memory. Therefore methods are being sought to handle the dictionary economically. So-called compression has been found useful; i.e., in one machine-dictionary entry a certain code contains information on several entries. Each compressed entry carries an index number which permits simple searching in the dictionary. The approach to this problem is demonstrated in a program for the Leningrad computer, the "Strela." The problems of dictionary structure were also discussed by G. M. Strelkovskiy at the Moscow conference in 1958 (cf. Strelkovskiy, Some General Principles of Compiling Dictionaries for Machine Translation, Tez I, p 56; also I. K. Belskaya, The Principles of Constructing a Dictionary for Machine Translation, VYaz, 1959, No 3, p 89; a more specific problem is discussed by A. V. Superanskaya in The Process of Transcribing Personal Names and the Possibility of Automating It, BMP, 6, p 44). In addition to problems similar to those discussed by the authors of the above article he considers the possibilities of supplementing the dictionary with new words. A very important dictionary problem is the question of homonymy, which is serious particularly in English, especially with regard to the conversion of word classes (cf. T. N. Moloshnaya, Problems of Differentiating Homonyms in Machine Translation from English to Russian, PK I, p 215). The solution will come only through formalization of the particular relations. Some cases of conversion are, for the time being, formally

insoluble, particularly if several follow one upon another. Still more complicated is homonymy among words of the same class. Here it will be necessary to examine various contextual influences on the use of equivalents. A very interesting article on a machine-translation dictionary was published by S. S. Belokrinitskaya, in which she shows what a German algorithm should look like. Her approach is circumscribed by the limitations of the BESM computer (cf. Belokrinitskaya, The Principle of Compiling a German-Russian Dictionary of Multiple-Meaning Words for Machine Translation, MP, p 89). She divides the dictionary into single-meaning and multiple-meaning words. In determining the meaning of a multiple-meaning she proceeds thus: 1. She makes a semantic and structural analysis of the environment; 2. She determines the meaning on the basis of formal signs.

The most important thing in machine translation is syntax. Syntactic research is important not only for word-order, but also for solving homonymy of word classes, the dictionary, morphology, etc. Syntactic relations are extremely complicated, and thus difficult to describe. In one of her first articles T. N. Moloshnaya discussed syntactic problems in English. She connected the problem of syntactic analysis with the problem of word classes and, following C. Fries' example, attempted to find formally determined word classes (cf. Moloshnaya, Some Problems of Syntax in Connection with Machine Translation from English to Russian, VYaz, 1957, No 4, p 92; see also similar articles by I. I. Revzin, Some Problems of the Formalization of Syntax, BMP I, p 5; T. M. Nikolayeva, Analysis of the Russian Sentence, ITM i VI, Moscow, 1958; and C. Fries, The Structure of English, New York, 1952). On the basis of this differentiation she sought syntactic relations among individual formations. Her method is very similar to Bar-Hillel's description and analysis of language (cf. Yehoshua Bar-Hillel, A Quasi-Arithmetical Notation for Syntactic Description, Language 29, 1953, p 47). Using these relations it is possible, in the author's opinion, to determine individual language types. B. M. Leykina's article is a substantial contribution to research on English syntax (cf. Leykina, The First Stage in the Independent Analysis of the Structure of a Simple Sentence in English, MMP I, p 216). Leykina approaches the analysis of the English sentence from the standpoint of independent analysis of English, i.e., she assumes translation through an intermediate language, and not binary translation from one language to another (cf. L. N. Zasorina, N. B. Karachan, S. N. Medvedeva, G. S. Tseytin, Programs for the Morphological Analysis of Russian in Machine Translation, MMP I, p 136). The author limited her discussion to the simple English sentence. She considered the main purpose of her work to establish a working hypothesis whereby it should be possible to find individual sentence terms. The possibility of formal handling of sentence punctuation was discussed by T. M. Nikolayeva (cf. Nikolayeva, Analysis of Punctuation Marks in Machine Translation from Russian, MP, p 33). This problem is very important, particularly for coordinating sentence terms with the sentence. A similar topic was discussed by D. Dzhanaridze, I. A. Mel'chuk, and T. N. Moloshnaya at

the Leningrad conference in 1959 (cf. Dzhanaridze, Mel'chuk, and Moloshnaya, The Treatment of Coordinating Bonds in Going from Input Language to Intermediate Language, Tez II, p 63). They were interested in coordinating bonds, their expression in the intermediate language, and how to translate from the input language to the intermediate language. General syntactic problems connected with machine translation were solved at the 1958 Moscow conference. L. I. Iliya discussed the relationship between the two main schools of syntactic research: "immediate constituents" (Pike, Bloomfield) and sentence terms (Bazell, Kurylowicz, Diderichsen), and believed that the two methods could be complementary to one another (cf. Iliya, Methods of Breaking Down a Syntactic Whole, Tez I, p 43).

All work is directed toward developing algorithms for individual languages. Initially most of this work was done on English. The problem of honomymy, for instance, was discussed by T. N. Moloshnaya in an article mentioned above (cf. PK I, p 215). The same author also worked on problems of English syntax. I. K. Bel'skaya, working in the same group with Moloshnaya, devoted one of her articles to problems of the English dictionary (cf. Bel'skaya, The Basic Characteristics of a Dictionary and Grammatical Schemes for Machine Translation from English to Russian, MP, p 47; also see T. N. Moloshnaya, Generalization of Grammatical Rules for Machine Translation from English, BMP 2, p 1). It should be noted that a great deal of work has been done in the USSR for a long time on the English algorithm, and by now it has been thoroughly worked out. This was helped by the fact that, along with the Russian algorithm, it was the first to be worked on (cf. the 1954 New York experiment). Work soon began on a French algorithm. I. A. Mel'chuk contributed greatly to the analysis of French (see his article in this journal, infra), and he formulated his conclusions in general form (cf. O. S. Kulagina and I. A. Mel'chuk, French-Russian Machine Translation, VYaz, 1956, No 5, p 111; also O. S. Kulagina, Machine Translation from French, Izvestiya VUZov, Matematika /News of the Higher Educational Institutions, Mathematics/, 1958, 5(6), p 46; and I. A. Mel'chuk, Statistics and the Dependence of the Gender of French Nouns on Their Endings, VSR, p 112; Mel'chuk published a similar article with the same title in BMP 7, p 13). Statistical research on the gender of nouns and their endings and a comparison of these results with the situation in other languages shows how close or remote languages are typologically. Intensive work is being done to develop a German algorithm. Along with Belokrinitskaya (cf. above, MP, p 89), V. V. Parshin is interested in a typical phenomenon of German -- compounds -- and their translation into Russian (cf. V. V. Parshin, The Translation of Compound Nouns from German to Russian in Machine Translation, MP, p 81). The procedure is that the words are divided into individual components and, on the basis of the Russian equivalent, various translation possibilities are sought. The problem of compounds should also be solved from the standpoint of the independent analysis of German, and not only through binary research.

A characteristic of machine-translation work in the USSR, unlike that in the USA, is that work is being done on a large number of algorithms from a wide variety of languages. Thus I. A. Mel'chuk has essentially worked out a Hungarian algorithm (see article below; cf. also Mel'chuk, Some Problems of Machine Translation from Hungarian to Russian, BMP 4, p 1; Mel'chuk, Some General Conclusions in Connection with Machine Translation from Hungarian, BMP 6, p 44; Mel'chuk, Machine Translation from Hungarian to Russian, PK 1, p 222). Work is also continuing on algorithms for Chinese and Japanese (cf. I. V. Sofronov, General Principles of Machine Translation From Chinese, VYaz, 1958, No 2, p 116; V. A. Voronin, Grammatical Analysis in Chinese-Russian Machine Translation, MP, p 101; M. B. Yefimov, Some problems of Japanese-Russian Machine Translation, MP, p 114; A. A. Babintsev and Yu. P. Semenishchev, Japanese-Russian Machine Translation, MMP I, p 209). In Leningrad, where the Laboratory for Machine Translation is working under the direction of N. D. Andreyev, algorithms are being developed for European languages (including Norwegian, cf. V. P. Berkov and M. P. Cherkasova, Work on a Norwegian-Russian Machine-Translation Algorithm, MMP I, p 98) and oriental languages, including Hindi, Vietnamese, Burmese, and Arabic (cf. T. Ye. Katenina, Work on a Hindustani (Hindi)-Russian Machine-Translation Algorithm, MMP I, p 191; N. D. Andreyev, D. S. Batova, V. S. Panfilov, and V. M. Petrova, Elements of an Independent Analysis in a Vietnamese-Russian Machine-Translation Algorithm, MMP I, p 199; Y. K. Lekomtsev, The Structure of the Vietnamese Verb Syntagma, VSR, p 131; N. D. Andreyev, E. A. Zapadova, and O. A. Timofeyeva, Some Problems of Constructing a Burmese-Russian Machine-Translation Algorithm, MMP I, p 126; O. B. Frolova and V. I. Strelkova, The Initial Stage of Work on an Arabic-Russian Machine-Translation Algorithm, MMP I, p 112). In the autumn of 1959 work will begin on an independent algorithm for Czech electrical-engineering text.

Machine translation has an important influence on linguistic theory, particularly on linguistic methods. Recently several publications have appeared attempting to apply set theory to linguistic research; cf. J. H. Greenberg, Essays in Linguistics, New York, 1957. In the Soviet Union O. S. Kulagina has been interested in these applications, and has already achieved considerable success (cf. O. S. Kulagina, A Method of Determining Grammatical Concepts on the Basis of Set Theory, PK I, p 203; A Method of Determining Linguistic Concepts, BMP 3, p 1; Kulagina's view was originally presented in her candidate's theses, cf: Certain Theoretical Problems of Machine Translation, Otdeleniye prikladnoy matematiki Matematicheskogo instituta im. Steklova /Department of Applied Mathematics of the Mathematics Institute imeni Steklov/, Moscow, 1958; V. A. Uspenskiy, Determining Part of Speech in a Set-Theory Linguistic System, BMP 5). Her very original view was in essence adopted at the Leningrad conference by all centers in the USSR. Intensive work is also being done on linguistic statistics, in which the most modern techniques of mathematical statistics are being used (cf. Ye. V. Paducheva, Statistical Investigations of Syllable Structure,

VSR, p 100; Z. M. Volotskaya, I. A. Mel'chuk, T. N. Moloshnaya, and I. N. Shelimova, The Russian Dictionary of Frequency, VSR, p 93; R. G. Piotrovskiy, Some Problems of the Statistical Investigation of Lexical Groups, VSR, p 85; V. A. Artemov, The Use of Statistical Methods in Experimental-Phonetic and Psychological Studies of Speech, VSR, p 73; I. I. Revzin, The Relationship of Structural and Statistical Methods in Modern Linguistics, VSR, p 45; A. L. Shumilina and Z. M. Volotskaya, Some Statistical Data on Russian Nouns (Based on Mathematics Texts), BMP 7, p 41). Probability theory and the theory of information and communication are being applied (cf. L. R. Zinder, Linguistic Probability, VYaz, 1958, No 2, p 121; the author's article by the same name in VSR, p 58; V. V. Ivanov, Probability Determinations of Linguistic Time, VSR, p 62; A. Suprun, The Cause of the Redundancy of Linguistic Information, Tezisy dokladov i soobshcheniy na 7-oy konferentsii professorsko-prepodavatel'skogo sostava /Contents of Papers and Reports at the Seventh Conference of Professors and Lecturers/, Frunze, 1959, p 7).

Applied linguistics in the USSR has a broad theoretical base; the relationship to traditional linguistics is considered, and previous results of linguistics are reevaluated in the light of new findings (cf. V. V. Ivanov, Linguistics and Mathematics, BMP 5; p 5; R. L. Dobrushin, An Elementary Grammatical Category, BMP 5, p 19; P. S. Kuznetsov, Basic Aspects of Phonology, BMP 5, p 27; V. V. Ivanov, Code and Communication, BMP 5; V. V. Ivanov, The n-Dimensional Space of Language, BMP 5; V. V. Ivanov, The Concept of Neutralization in Morphology and Lexicon, BMP 5; S. K. Shaumyan, The Concept of the Phoneme in the Light of Symbolic Logic, BMP 5, p 58; Ibid, Logical Analysis of the Phoneme Concept, in Logicheskiye issledovaniya /Logic Research/, p 158, Moscow, 1959). It is interesting that many authors try to apply mathematical methods to comparative linguistics (cf. V. V. Ivanov, Some Concepts of Comparative-Historical Linguistics, BMP 5; G. A. Klimov, The Glottochronology Method of Dating the Breakup of a Proto-Language, VYaz, 1959, No 2, p 119). Soviet linguistics have been helped considerably by the understanding of logicians, mathematicians, and engineers (cf. L. R. Zinder, An Experience Shared by Phoneticians and Communications Engineers, VYaz, 1957, No 5, p 111). Linguistics can work on these new problems only together with them. The advantage of this collaboration lies in the fact that it is important not only for linguistics, but also for mathematics, logic, etc., not to speak of the great future importance of the practical application of machine translation and other related disciplines of applied linguistics.

PROBLEMS OF MACHINE TRANSLATION IN MOSCOW

Slovo a slovesnost, No 4,
1959, pages 285-289

I. A. Mel'chuk (Moscow)

This article concerns the content, goals, and outlook for one of the Soviet groups interested in problems of machine translation. This group formed around A. A. Lyapunov and the mathematicians in his group at the Mathematics Institute imeni Steklov in Moscow.

This group began work in 1955, when the mathematicians O. S. Kulagina, T. Venttsel', and the author of this article undertook to develop an algorithm for machine translation of mathematics text from French to Russian. Mathematics text seems to be most suitable for the first attempt because of the standardized terminology, small lexicon, and simple, largely standardized syntax.

With the help of students from the Philology Faculty of Moscow University a dictionary abstraction was made of French mathematics texts. About 30,000 text words were examined. Among these 30,000 words more than 1,000 different words were found. Without statistical investigation the most important auxiliary words and a number of important terms were added, although they did not appear in the examined texts. This made a dictionary of about 2,000 terms. In 1955-1956 a special translation algorithm was developed, a partial description of which was published as Machine Translation from French to Russian, by O. S. Kulagina and I. A. Mel'chuk, VYaz, No 5, 1956, pp 111-121. A complete description of this algorithm will be published by O. S. Kulagina, An Algorithm for Translation from French to Russian, in Problemy kibernetiki /Problems of Cybernetics/.

In 1957 O. S. Kulagina and G. V. Chekova finished programming the French-Russian algorithm and test translation was begun on the "Strela." Several dozen sentences have been published, with no pre-editing of the input text or post-editing of the output. We present examples of machine-translated sentences:

Indiquons une autre méthode pour établir l'existence des intégrales des équations différentielles ordinaires.	Ukazhem drugoy metod, chtoby ustanovit' sushchestvovaniye integralov obychnykh differentsial'nykh uravneniy.
Si le point (a, b) est sur une ligne E, la valeur cherchée de l'intégrale sera nulle.	Yesli tochka (a, b) yest' na linii E, naydennoye znacheniye integrala budet ravnym nulyu.
Nous pouvons par suite énoncer pour l'equation (α) le théorème suivant.	My mozhem sledovatel'no sformulirovat' dlya uravneniya (α) sleduyushchuyu teoremu.

Le théorème qui vient d'être établi subsiste dans ces nouvelles conditions.	Teorema kotoraya tol'ko chto byla ustanovlena sushchestvuyet v etikh novykh usloviyakh.

Many translated sentences contain errors caused by mistakes in the program or technical difficulties, and sometimes by shortcomings in the algorithm. For example:

Pour effectuer l'intégration, nous chercherons d'abord l'intégrale générale, qui sera une certaine fonction de y, dans laquelle figurera x.	Chtoby osushchestvit' integrirovaniye my naydennem snachala obshchiy integral kotoraya budet nekotoroy funktsiyey y v kotoroy figurirovayet x.
Or si le nombre K augmente indéfiniment, on obtient (formule) c'est-a-dire que tous les coefficients deviennent égaux à l'unité.	Itak, yesli chislo K vozrastet neogranicenno my poluchim (formule), to-yest', chto tous koeffitsiyenty stanovyatsya ravnymi yedinitse.

Here tous was not translated for purely technical reasons.

During the testing the translation programs were improved and perfected; the actual algorithm was also improved upon. Recently (summer 1959) O. S. Kulagina and G. V. Chekova performed the machine translation of connected text fragments with quite satisfactory results.

The main features of the French-Russian algorithm are:

1. The translation process is divided into two parts: analysis and synthesis. In analysis the program determines for each French word what form the corresponding Russian word shall have. Synthesis is the formation of these Russian forms. Translation is thus word-for-word.

2. Analysis is done principally on the basis of morphology. Morphological differences in the written French are expressed quite rigorously, which makes it easy to analyze endings and auxiliary words (prepositions, auxiliary verbs, etc.). In numerous cases, of course, one must use a larger context; but these cases of syntactic analysis are isolated, do not form a system, and play only an auxiliary role vis-a-vis morphology.

3. Since word order in French and Russian is similar it was found that in analysis of the French text it is unnecessary to examine the position or relations of words. Word order is retained in translation with the exception of individual cases of simple rearrangement (e.g., the rearrangement of following adjectives: équation différentielle -- differentsial'noye uravneniye).

The French-Russian algorithm is set up especially for binary translation; this means that it is suited for the given language pair and for translation in only one direction, only from French to Russian.

It is unuseable, or at least very impractical, for translation into any other language, since the analysis of the French text was approached from the standpoint of translation into Russian.

After the French-Russian algorithm T. N. Moloshnaya of the Mathematics Institute developed an English-Russian algorithm, also for translating mathematics text. Although this algorithm is also binary (since it was intended for translation from English to Russian) it is constructed quite differently than the French-Russian algorithm. Because of the sparse morphology and widespread conversion (a word may belong to various word classes: work may be verb, noun, or adjective) the most important thing in English is syntactic analysis, which requires examination of a broad context. T. N. Moloshnaya worked out a special system of methods for such analysis, based on determination of typical elementary word connections (usually of two words) in English and in Russian. The principles on which she developed her English-Russian algorithm have been described in Voprosy Yazykoznaniya, 1957, No 4, p 92, under the title Some Problems of Syntax in Connection with English-Russian Machine Translation. Here we shall mention her main points.

1. English sentences are broken down into elementary word groupings of two to four words; these are not groups of words themselves, but structures composed of types of words located in a certain sequence and having a certain grammatical form. For example adjective + noun agreeing in gender, number, and case /sic/ ("deep snow"), etc. (cf. p 92 of article cited).

2. Analogous Russian groupings are assigned to these simplest English word groupings. In most cases the assignment is unambiguous; special rules apply to cases of multiple meaning. As she writes: "The English sentence is subjected to analysis (determination of the simplest free groupings of words) and then the Russian sentence is synthesized out of the Russian groupings corresponding to the English ones" (p 92).

3. In analysis of English text so-called compression is used; i.e., simplification of word groupings such that a whole grouping is replaced by its main component. For example the grouping adjective + noun is syntactically equivalent to a noun and under certain conditions can be replaced by a noun. Compression is used, as the author says, "so that relations can be progressively determined among all words in the sentence, not only among those which are immediate neighbors but those between which there is a real connection" (p 94).

The English-Russian algorithm, based on comparison of elementary structures (configurations) of two languages, goes beyond the French-Russian algorithm, in which comparison was limited to individual words.

After the English-Russian algorithm, in 1956-1957 the author developed a Hungarian-Russian algorithm, intended principally for translation of a special linguistic text (cf. Byulleten' ob"yedineniya po mashinnomu perevodu pri MGPIIYa, 1957, No 4; see also the author's

article On Machine Translation from Hungarian to Russian, Problemy kibernetiki, 1, 1958, pp 222-264). The Hungarian was treated so as to reveal the peculiarities introduced into the algorithm by the unusual structure of that language, which differs so from the structure of French and English. It was found that, although the morphology of Hungarian is very rich and grammatical meanings are expressed primarily through word forms, it was impossible to develop a Hungarian-Russian algorithm solely on a morphological basis, analogous to the French-Russian.

The fact is that word order in Hungarian differs so from that of Russian that a properly translated sentence which retains the initial word order can be quite incomprehensible, or comprehensible only with difficulty, in Russian. Therefore numerous and quite complicated word inversions must be made in the translated text, requiring a knowledge of the mutual relations and dependencies of words. This shifts the burden to syntactic analysis; here it is more convenient to procede slightly differently from what was done in the English-Russian algorithm. In syntactic analysis of Hungarian text it is usually necessary to determine the relationships among words remote from one another. This operation requires special apparatus, not developed for English or French.

Experience from working on these three algorithms permits certain conclusions, on which all subsequent work has been based:

(a) The rejection of purely binary algorithms, closely bound to a given language pair, and changing instead to the attempt to create general algorithms permitting translation from the given language to any other language, and conversely.

(b) Translation primarily based on syntactic analysis, in which the underlying factor is comparison of the simplest elementary word groupings of one language with analogous word groupings in the other language.

(c) The need for developing a system of analogies among languages to include lexical analogy; morphological analogy -- including such extrasyntactic categories as number in nouns, tense and mood in verbs, etc.; and syntactic analogy among elementary word groupings of various languages.

This system of analogies, forming a certain calculus, can be considered an intermediate language (Russian: yazyk-posrednik) for machine translation. (This is one of many approaches to the problem of intermediate language. Cf. for instance N. D. Andreyev, Machine Translation and the Problem of an Intermediate Language, VYaz, 1957, No 5, p 117, and the foregoing article)

At present the greatest efforts are concentrated on developing this intermediate language.

Translation through an intermediate language is conceived thus: For each translated language a special algorithm is developed made up of two groups of rules: rules of analysis and rules of synthesis.

The rules of analysis permit passing from text in the input language to an agreed-upon numerical code in which each word in a given form and syntactic function is unambiguously assigned to a certain numerical series called the information about the word. The information is the totality of ciphers, denoting all possible characteristics of the word.

The information is broken down into typically two-termed groupings, or configurations. One term of a configuration is the main one, the other is the dependent. Each configuration is assigned a number for the syntactic relationship existing between the words forming the configuration; and this completes the analysis.

The rules of synthesis are used to perform operations opposite to analysis; knowing the numbers of the syntactic relations which were expressed one can find the corresponding configurations and proceed from a sequence of configurations to real text in the output language, since a word in a certain form is unambiguously assigned to each numerical series, each unit of information.

Thus analysis and synthesis are performed independently of the translation and remain unchanged for a given language when translating from that language to any other, and conversely. Therefore the algorithms for analysis and synthesis are developed separately for each language.

Furthermore there is the intermediate language, i.e., special tables and sets of rules whereby one proceeds from the numerical code of the input language to the numerical code of the output language, such that to a certain number of units of information in one language certain groupings corresponding to information about the second language are assigned. The relationships are often ambiguous; in such cases rules for selection are used, based on context analysis.

Work is now being done on an intermediate language in these institutes: Institute of Linguistics, Acad Sci USSR; Mathematics Institute imeni Steklov (linguist T. N. Moloshnaya, mathematician O. S. Kulagina, and G. V. Chekova); and in the Laboratory of Electromodeling, Acad Sci USSR (linguists Ye. V. Paducheva, M. Langlebena, A. L. Shumilina, I. N. Shelimova, and Z. Volotskaya). Work is being done on a model of the intermediate language. The following algorithms are being developed for it: Russian, English, French, Hungarian (principally analysis rules). All algorithms are based on short text, of about 150 words; not all grammatical phenomena are included in such text, but the most important ones, those most characteristic of each language, are taken into account in the algorithms of the model. The goal of the work at this stage is not an exhaustive description of the languages under investigation, but rather the development of a basic, but concretely precise, scheme of such a description. If this goal is achieved satisfactorily, i.e., if the scheme is successful, it will not be too difficult to expand the algorithms and no basic changes will be necessary in the scheme itself.

In the work a great deal of purely linguistic research had to be done in order to solve problems which are usually not included in grammars. For example, A. L. Shumilina compiled formal rules for so-called "reconstruction of the pronoun," i.e., for determining which noun is represented by a given pronoun; I. N. Shelima investigated the question of the word or words to which a prepositional structure (preposition + noun) is related; etc.

The creation of a well-developed and universal system of analogies among languages will be useful not only to practical machine translation but also to linguistic theory. Perhaps on the basis of an intermediate language -- the abstract calculus -- with sufficient logical treatment it will be possible to create a "meta-language" of linguistics, and to develop methods of formal description, which linguistics needs so badly.

The plans for further work are these:

1. In work on an intermediate language to use other languages, particularly German, the Scandinavian languages, Slavic languages (except Russian), Vietnamese, Arabic, Indonesian, etc., to record, in a common system, the maximum number of different linguistic features, and thus achieve a more general system.

2. Expand and supplement existing algorithms:

(a) Increase the stem dictionaries up to several thousand entries;

(b) Make up exhaustive lists of the elementary word groupings possible in any given language;

(c) Treat phraseology more fully and carefully;

(d) Record and describe the maximum number of individual properties of individual languages (specific constructions, word usages, etc.).

3. In collaboration with specialists in logic and logical semantics, to improve methods and schemes of formal description. Perhaps later, on the basis of an intermediate language arrived at empirically as a generalized system of analogies among a number of concrete languages, it will be possible to create a logically treated, strictly formalized machine-translation language, a type of logical calculus.

4. To develop statistical research on language so that mathematicians and linguists can solve problems of economy and convenience of particular methods of description on the basis of frequency characteristics of the phenomenon under investigation.

5. To describe the methods used in developing machine-translation algorithms so that in the future, when the problem of the actual process of developing algorithms is solved, this work can be done by machines (using prepared dictionaries, parallel texts in several languages, etc.).

6. In theory, the basic task is to make maximum use in linguistics of the collaboration of linguists and mathematicians, logicians, and engineers. The methods of the exact sciences must be used in linguistics, proper rigor of formal descriptions must be achieved in various linguistic research, etc.

7. In conclusion it should be added that because no special electronic machine exists for translation all work has so far been done on general-purpose computers. Although translation by general-purpose computer is possible, it is not efficient or economical. Therefore experience gained must be collected and present procedures analyzed in terms of the design and construction of an electronic machine especially for translation. The construction of such a machine will encourage further work on machine translation. (For progress in this type of work, see the foregoing article.)

NEW PROBLEMS OF MATHEMATICAL METHODS IN LINGUISTICS

Slovo a slovesnost, 1959,
No 1, pages 44-55

Petr Sgall

We should like to give here a survey of the most recent work in various sectors of so-called mathematical linguistics, whose methods are still relatively little known in Czechoslovakia. We shall concern ourselves particularly with the relationship between the mathematical theory of information and linguistics. The purpose of the article is to increase the interest of our specialists in these new methods, with of course a critical review of the somewhat uneven results which have been achieved hitherto.

Mathematical methods are today quite widespread in linguistics in western countries, particularly the USA, and in the USSR some aspects have undergone considerable development. Harvard University has held permanent courses in mathematical linguistics for several years; cf. J. Whatmough, Reports for the Eighth International Congress of Linguists I, Oslo, 1957, p 211. At Moscow University this branch of study will be introduced this year; a seminar has existed since 1956 whose participants have been applying mathematical methods in linguistics on the basis of work by A. N. Kolmogorov, and A. A. Lyapunov (cf. V. Yu. Rozentsveyg, Work on Machine Translation, Moscow, 1958, page 9, paper for the Fourth International Congress of Slavists). And even in Czechoslovak linguistics the use of statistical procedures in certain fields -- phonology, lexicology -- is quite common. Thus we shall concentrate our attention here on the use of logical calculi (connected with set theory) and the application of the methods of information theory.

Statistical methods in linguistics have been the interest of Professor B. Trnka, who is the author of the first international bibliography of quantitative linguistics and of several special articles dealing mainly with problems of phonology (cf. B. Trnka, A Tentative Bibliography, Utrecht, 1950; Pokus o vedeckou teorii a praktickou reformu tesnopisu /Attempt at a Scientific Theory and a Practical Reform of Shorthand/, Collection of Papers and Discussions of the Philosophical Faculty, 20, Prague, 1937; A Phonological Analysis of Present Day Standard English, Prace z vedeckych ustavu /Work of Scientific Institutes/ 37, Prague, 1935; On the Development of Phonological Statistics, Slovo a slovesnost, 1948, p 59-64. A second and more extensive bibliography of quantitative linguistics was compiled by P. Guiraud, Bibliographie critique de la statistique linguistique, Utrecht, 1954). Trnka's report on quantitative linguistics (CMF 34, 1951, p 66-74) acquainted Czechoslovak specialists with the most important results achieved up to that time in the precise analysis of the quantitative aspect of linguistic fact. In this article, in criticizing Zipf's work, Trnka showed the error in drawing quick generalizations from statistical computations.

(In Slovo a slovesnost Zipf's methods were criticized by N. S. Trubetzkoy (2, 1936, p 252); later Trnka criticized them in the above-mentioned article. Thus we can today generally ignore Zipf's work, although it remains valuable as a pioneer effort. Zipf's results have been sharpened recently by A. Kotsoudas in an article in which he computes the size of dictionary necessary for a text of a certain length (Language 33, 1957, No 4, p 545-552.)).

The use of statistics in problems of phonology is well-known in Czechoslovakia from the work of V. Mathesius, B. Trnka, J. Vachek, and others. (Cf. V. Mathesius, Cestina a obecny jazykozpyt /Czech and General Linguistics/, Prague, 1947, p 48ff, on the use of individual phonemes; p 62ff on quantitative relations among vowels and consonants, etc.; p 87ff on the relation between the rarity and the expressivity of phonemes). J. Vachek is interested in the significance of phoneme frequency for historical phonology; his most recent publication is in the Zeitschrift f. Angl. u. Amerikanistik 1, 1957, pp 5-28. Important contributions to the application of mathematical methods in another area are contained in his Jazykovedna problematika zkousek srozumitelnosti reci /Linguistic Problems of Tests of Language Comprehensibility/, Prague, 1956. Trnka's work is referred to by J. Kramsky in On the Quantitative Phonemic Analysis of English Mono- and Dissyllables, Philologica 8, 1956, pp 45-59; etc. In the western countries, particularly when phonological analysis emphasizes the criterion of distribution, statistical treatment of phonological problems is common.

As has been the case often with new methods at earlier periods of linguistics, statistical methods have found their first application in phonology, where the systematic relations among linguistic elements are most obvious. (As long ago as the middle of the last century E. Forstemann, in several articles which Professor Trnka has kindly pointed out to me, showed that statistical data on the occurrence of sounds help to show the relationships among related languages; cf. Kuhns Zeitschrift, Vol 1, 1852, pp 163ff, and Vol 2, pp 35ff). In the articles to which we refer principal emphasis is placed on the linguistic aspect of the problem (which, even with its relative transparency, is quite complicated). A rigorous distinction is made between the occurrence of phonemes in the system and in text, the numerical occurrence of phonemes and the number of different positions in which any given phoneme can appear; and the basic concepts used (phoneme, variant, relevant feature, etc.) are quite clearly delimited. This is true also in the application of statistical methods to problems of lexicon, particularly in working on frequency dictionaries. (On the problem of frequency dictionaries several articles have recently appeared in Czechoslovakia; cf. V. Vey, Apropos de la statistique du vocabulaire tcheque, Slavia 27, 1958, pp 396-409; V. Fried, Semanticka frekvence anglickych slov a nektere moznosti jejiho vyuziti /The Semantic Frequency of Englich Words and Some Possibilities of Utilizing It/, CMF 37, 1955, pp 129-142; and The Numerical Determination of the Lexicon, Cizi jazyky ve skole, 1, 1957, No 4, pp 145ff).

In other areas the situation is not so good. The statistical treatment of linguistic phenomena has frequently been approached by mathematicians with little linguistic preparation. Of course these authors use verified statistical procedures, in which they can include a large quantity of linguistic facts, but they often fail to devote sufficient attention to linguistic problems. This can be seen in some basic work, as for instance Yule's computation of the constants characteristic of the style of a particular author, based on the use the author makes of his lexicon (cf. G. U. Yule, The Statistical Study of Literary Vocabulary, Cambridge, 1944). Yule approaches this computation on the basis of material selected without an unambiguous theoretical foundation, and examines only the occurrence of all nouns in the texts. He states, with justification, that a similar examination of verbs and adjectives would yield essentially the same results, whereas auxiliary words are not suitable for this purpose, since their occurrence is stable, regardless of the author. But classification according to word classes is not the only possibility: within word classes considerable differences can be found on this basis. Furthermore not all auxiliary words have stable occurrence for all authors, while some frequently-used nouns are less significant for the author's style.

Another and much more clear-cut example is W. Fucks' article on a "mathematical theory of word-formation." Not only is the author's view of the word as made up of syllables somewhat imprecise linguistically -- his point of view certainly has its justifications. But the author bases his computations on the statement that "every word has at least one syllable" (pp 221, 228, 242, 244, etc.), while one of the nine languages he examines is Russian and the author has nothing to say about his treatment of such prepositions as k, v, etc. Nor does he mention the possibility of nonsyllabic words in other languages. Did he consider the Russian nonsyllabic prepositions syllabic, or did he not consider them words at all? The reader has no chance to check the author's procedure, and the results which Fucks reached lose some of their impact. This is true of the work of other mathematicians who have been interested in problems of the occurrence of words with various numbers of syllables; e.g., Herdan's book, which we shall discuss below, or an older article by the Soviet mathematician S. G. Chebanov (Doklady AN SSSR 55, 1957, No 2, pp 103-106).

Even outside linguistics it is known what danger attends the use of statistical methods, particularly in investigating social phenomena, wherein not only quantity but also quality of phenomena is important, and where the complexity of the problem leaves the rules of mathematical statistics somewhat short of the required precision. In discussions held by Soviet statisticians a few years ago the conclusion was drawn that statistical analysis should always be combined with analysis of the nature of the phenomena studied. (This agrees with the present stand of the Soviet linguists who emphasize the need for analysis of linguistic structure and a precise determination of the units which

are to be treated statistically. Cf. a paper of V. A. Uspenskiy at the conference on the statistics of speech, Voprosy yazykoznaniya, 1958, No 1, pp 170-173, particularly page 172, on I. I. Revzin's paper).

Only those computations can be fully valid linguistically whose basis was facts properly determined and classified linguistically. If we can overcome this danger there is no need to reject mathematical methods in linguistics; it cannot be said that we can get along without counting in linguistics, that the facts are not homogeneous, or suitable for statistical treatment. It is well-known that the determination of quantitative relations is important for examining the hierarchy of the linguistic system (although this is not the only decisive criterion) and thus has a heuristic value; it is also important for teaching foreign languages, etc. (cf. the Trnka article cited above, and also V. Fried's article in CMF 33, 1950, pp 157-162).

The methods of quantitative linguistics are applied in various ways in the investigation of problems of historical linguistics: in determining the chronology of texts, the degree of relatedness of languages, and other problems. Here, too, of course, mathematical accuracy does not guarantee the correctness of conclusions arrived at from inaccurate or uncertain assumptions.

For example, in determining the relative chronology of old monuments efforts -- often unprofessional from the mathematical standpoint -- have long been made to count various archaisms and innovations: phonological, morphological, lexical, metric, etc. In examining the age of individual parts of the Rigveda the British Indic scholar E. V. Arnold used this method with success (cf. Arnold, Vedic Metre in its Historical Development, Cambridge, 1905). The German scholar W. Wust, in a later work on this problem (Stylistic History and Chronology of the Rigveda, Abh. f. die Kunde des Morgenlandes 17, No 4, Leipzig, 1928) concentrated his attention mainly on lexical criteria, which are less reliable). P. Poucha, in Czechoslovakia, has attempted the same thing (cf. Poucha, Layering of the Rigveda, Archiv orientalni 13, 1942, pp 103-141, and 225-269; AO 15, 1944, pp 65-86). A large number of archaisms will indicate quite reliably the relative age of a given song, and a large number of archaic songs in a particular book of the Rigveda can indicate its relative chronology. But it cannot be stated that we can thus determine the oldest or youngest song in the Rigveda. The criteria for determining archaism are often unclear, and archaisms can be used long after the introduction of innovations, so that we have no precise criteria. The fact that we do not consider such results precise does not, of course, deny their value to linguistics; other methods of solving these problems are usually even less precise.

In determining degrees of relatedness of languages statistical methods were used as early as the 1920's by the Polish anthropologist and ethnographer J. Czekanowski. He published statistics on phonological and morphological phenomena important from the standpoing of comparative linguistics in his book Wstep do historii Slowian /Introduction

to the History of the Slavs/, published first in Lwow in 1927, now in its second edition, Poznan, 1957. Cf. particularly pages 5ff on general problems; 158-166 on the classification of Indo-European language groups; 179-201 and tables on the relations among the Slavic languages. See also Czekanowski, Na marginesie recenzji p. K. Moszynskiego o ksiazce: Wstep do historii Slowian /Answer to K. Moszynski's Critique of the Book: Introduction to the History of the Slavs/, Lwow, 1928, pages 6-16. Czekanowski was followed by some American investigators (cf. particularly A. L. Kroeber and C. D. Chretien, Quantitative Classification of Indo-European Languages, Language 13, 1937, pages 83-103). In the USA the method of so-called lexicostatistics, or glottochronology, is widely used today. This method was worked out in recent years mainly by M. Swadesh and R. B. Lees. It operates with a list of selected basic words which denote the most common objects and relations (hand, give, big) and thus change very slowly, and are minimally influenced by cultural development and neighboring languages. (Cf. M. Swadesh, Int. Journ. of Amer. Ling 16, 1950, pages 157-167, with a list of 225 words (for English); in IJAL 21, 1955, No 2, pages 121-137, he eliminates from this list some words which are not sufficiently stable or are not semantically differentiated from others, so that the list is left with only 100 words.) By comparing these basic words in two languages, or in two developmental periods of the same language, two dialects, etc., the percentage of words (of the list) is determined whose form permits assuming a common origin for the word in both languages. From the number of words which are not cognate in the two languages one then determines the relative time depth from the moment when the languages separated.

This method is used primarily in classifying American languages whose monuments do not reach far back into the past. Glottochronology provides no improvement in accuracy for languages which have been well-investigated by the historical-comparative method. The situation is rather the opposite: the American linguists use computations based on the development of the Indo-European languages to improve their statistical procedure. Thus R. B. Lees, in Language 29, pages 113-127, improves the foundations of lexicostatistical investigation by control computation of languages previously handled by the historical-comparative method. This procedure does not provide a reliable absolute chronology, and its results are not fully satisfactory in other ways as well. A critique of lexicostatistics was published by H. Hoijer in Language 32, 1956, No 1, pages 49-60; and J. H. Greenberg, Essays in Linguistics, New York 1957, pages 40ff and 54, opposes the overestimation of mathematical methods in solving questions of language relatedness. But for classifying those languages which have not been handled by the historical-comparative method glottochronology is a definite help, even if it evaluates the development of languages in terms of only one component (selected basic words), and even within that component overlooks many factors in linguistic development.

The possibility of mathematical treatment of language typology is indicated by J. H. Greenberg (Method and Perspective in Anthropology, Papers in Honor of Wilson D. Wallis, Minneapolis, 1954, pages 192-220), based on Sapir's typology (Language, New York, 1921). He is interested primarily in linguistic analysis of the question, and only suggests its mathematical treatment. For each language considered Greenberg computes ten indexes, such as the relationship between the number of morphemes and the number of words in the text, etc. Such a procedure is surely more convenient than the older view of typology as the classification of entire languages, in which the existence of different types of elements in the same language was overlooked. But the inclusion of different facts in one index has its disadvantages in comparison with linguistic analysis of the typological make-up of a language, in which one analyzes problems of the interrelationship of features, particularly in comparison with the procedure familiar in Czechoslovakia from Skalicka's work. (Cf. most recently V. Skalicka, The Present State of Typology, Slovo a slovesnost, 19, 1958, pages 224-232).

Some linguists object to the emphasis on statistical methods with justice, saying that the occurrence of a certain phenomenon in a text or in the system of the language per se is not decisive, but is only one manifestation of a hierarchy within the language system, in conjunction with other factors. From this standpoint it is necessary to evaluate the work which statistical research connects with the application of the logical calculus and similar procedures closely allied to the methods of set theory. These methods, too, are probably now applied most widely in phonology. The American investigators F. Harrary and H. H. Paper have attempted a description of Japanese phonology using the methods and results of the theory of relations (Language 33, 1957, No 2, pages 143-169). L. R. Zinder, the Soviet phonetician, treats some similar problems more succinctly (problems of the probability of occurrence of phonemes in a given position, cf. Voprosy yazykoznaniya 1958, No 2, pages 121ff). The frequency of various consonant groups has been investigated by a somewhat different method by S. Saporta, who concludes that the most frequent combinations are of those consonants which do not differ either maximally or minimally (i.e., groups such as sk, pt). Cf. Language 31, 1955, pages 25-30. Saporta bases his work on the methods of so-called psycholinguistics, which will be discussed below. The fact that phonemes differentiated by only minimal contrast do not occur in immediate juxtaposition within the same morpheme was established earlier by B. Trnka, TCLP VI, 1936, pages 57ff.

O. S. Kulagina is investigating a system of grammatical categories on the basis of set theory (cf. the symposium Problemy kibernetiki I, edited by A. A. Lyapunov, Moscow, 1958, pages 203-214. A general survey of Soviet work in this field was given by V. Yu. Rozentsveyg in a paper to the Fourth International Congress of Slavists,

cited in Raboty po mash. perevodu, Moscow, 1958, pages 9ff). Some American and west European linguists use similar methods in treating general problems of the linguistic sign (cf. J. H. Greenberg, op. cit., footnote 18, page 4; H. J. Uldall, Outline of Glossematics I, Copenhagen, 1957. Cf. also E. Koschmieder, Forschungen und Fortschritte 30, 1956, pages 210ff).

Some related questions have been solved principally in connection with work on machine translation. We have no intention of analyzing here the problem of machine translation; but we cannot overlook the importance of these problems for theoretical linguistics, which finds here a broad new field for practical work. The most important thing for theoretical linguistics is not the practical use of machine translation, but rather the demands which this new discipline places on theory, and the ideas which flow from it. There can be no doubt of the limitations on the practical use of machine translation, which have already been pointed out (cf. K. Horalek, Will It Be Possible to Translate from Russian to Czech by Machine?, Cs. rusistika 2, 1958, Nos 2-3, pages 85-88). Scarcely any specialist thinks seriously of automatic translation of belles lettres. The intention is not to replace creative artistic work by machine, but rather to use machines for mechanical tasks with which people need not be burdened. The decisive fact today, however, is that the foundations for machine translation must be laid by the work of linguists. This will cause linguistics to sharpen its tools, to make a detailed and penetrating treatment of sectors which have not yet been fully mastered (the classification of syntactic relations, of word meanings, etc.) which will make it possible to formulate unambiguous rules wherever linguistic reality permits. Today we are only beginning to outline new goals and fields of interest, whose importance will surely grow. For example, the treatment of the problem of the so-called intermediate language, i.e., an artificial system of relations, which is to serve as the intermediate level in translation from any language to any language, will surely be very important in the further development of so-called analytic comparison of languages, for comparison of their syntactic and lexical content, and for the general theory of grammar. The very fact that we are beginning to have available machines into which such a system can be "inserted" seems to open up new horizons and experimental opportunities in linguistics. And the theory of machine translation is also a contribution to the theory of translation in general (V. Yu. Rozentsveyg and I. I. Revzin have shown that the theory of machine translation has for the first time made it possible to make a strict differentiation between linguistic and literary problems in the general theory of translation (Tezisy konferentsii po mashinnomu perevodu /Papers from the Conference on Machine Translation/, Moscow, 1958, pp 26ff). M. I. Steblin-Kamenskiy (ibid, p 23) shows that work on machine translation is important for linguistics as a critique of traditional grammatical concepts based on practice, and thus more objective and effective than purely theoretical criticism. Cf. also his article in the symposium Material po mash. perevodu I, Leningrad, 1958, pp 3ff).

For machine translation a system of instructions must be worked out in advance for grammatical analysis permitting an unambiguous analysis of various sentences in the input language in terms of the written form of individual words and their position in the context; and the synthesis of corresponding sentences in the output language. In translating from a language with a sparse morphology, in which the function of the word in the sentence is frequently determined only by its position with respect to other words, the matter is much more difficult. The theoretical treatment of syntax from this standpoint is only beginning, whether so-called operational syntax, which Bar-Hillel is working out (cf. Language 29, 1953, No 1, pp 47-58, among his works available in Czechoslovakia) or other work based on the mathematical theory of Markov chains (cf. for instance V. H. Yngve, in Russian translations of two symposia: Mashinnyy perevod /Machine Translation/, Moscow, 1957, pp 271ff, and Teoriya peredachi soobshcheniy /The Theory of Information Transmission/, Moscow, 1957, pp 255ff. It is of interest to the linguist to note that A. A. Markov developed his statistical analysis of chains using, among others, linguistic material: he computed the probability of occurrence of letters in the text of Pushkin' Onegin; cf. Izvestiya Imper. Akad. Nauk, series 6, volume 7, St. Petersburg 1913, pages 153-162). F. W. Harwood (Language 31, 1955, pages 409-413) is attempting to develop a general mathematical delimitation of the syntactic system which will include all possibilities of morpheme combinations.

In addition to the foregoing words on phonology, lexicon, and syntax individual articles must be mentioned in which mathematical methods are applied to the analysis of other components of language. An article by Z. S. Harris (Language 31, 1955, pages 190-222) is important for morphology; in it the author counts phonemes which may appear after a given phoneme (or series of phonemes) in a text. The curve peaks obtained show quite accurately the limits of correspondence between word and morpheme. An example of the mathematical treatment of semantic problems is the work of R. Wells (Cahiers F. de Saussure 15, 1957, pp 117ff) who uses factor analysis of the relations among word meanings to show the possibility of a certain classification of word meanings.

The methods which we have mentioned permit the overall statistical treatment of a large number of linguistic phenomena, and also simplify the generalization of scientific procedures with a simple logical approach leading to higher abstraction. These methods of course assume a careful linguistic handling of the phenomena in the study of which they are to be applied, and per se make no change in basic linguistic concepts.

Some scientists expect still greater results from a new branch of mathematics -- information theory.

The problems of information theory and its basic concepts are analyzed in the following article by J. Kramsky; here we shall limit ourselves to the statement that the term "information" cannot be taken in

its usual sense. Some scholars distinguish several concepts of information: S. Goldman (a Russian translation of his work is Teoriya informatsii, Moscow 1957, pp 42ff) speaks of "semantic information" (i.e., information in the usual sense of the word) and "linguistic information" (which can be measured only by the methods of information theory; this theory does not take into account the meaning of the response, which is not ergodic, but only the features of the code in which the report is carried and the length of the message. The term ergodic refers to any process whose individual elements do not affect the probability of occurrence of other elements in the environment, or whose elements influence one another only over a distance expressed by a finite number of elements.) General problems of the formal treatment of semantic information are investigated by Y. Bar-Hillel and R. Carnap in the symposium Communication Theory, published by W. Jackson 1953, pages 503-512. The symposium includes the papers given at the international conference on communication theory in London, 1952. P. Neidhardt (cf. Informationstheorie, Berlin 1957, page 49) distinguishes 1. "semantic information" (the absolute quantity of new knowledge); 2. "structural information" (the relative quantity of new knowledge to a certain message recipient); 3. "idealized information" (the quantity of new knowledge according to a certain rule of selection, e.g., from an artistic standpoint); and 4. "selective information," in which there is no question of the correctness of the message, its significance to the recipient, or its esthetic value, but only of the signs of a certain code. Of these four concepts information theory, as P. Neidhardt points out, works only with the last. Selective information is a quantity indicating the degree of freedom of selection in compiling a given message; thus it is not a direct measure of the content of a message. A. A. Kharkevich (Ocherki obshchey teorii svyazi /Outlines of the General Theory of Communication/, Moscow 1955, p 34) notes that the term "information" here refers rather to the information which the message might contain than to that which it actually does contain.

The authors of work in information theory usually mention only briefly the possibility of applying their results to research on natural languages. Frequently they stop at the analysis of relationships among the letters of a given language (cf., in addition to the aforementioned articles, L. Brillouin, Science and Information Theory, 1956, pp 23ff. A closer connection to linguistics is seen in the creator of cybernetics and one of the founders of information theory, N. Wiener; cf. particularly his book The Human Use of Human Beings, translated into Russian as Kibernetika i obshchestvo, Moscow, 1958, pp 93ff). Among linguists H. A. Gleason has studied in detail the significance of information theory in his book (An Introduction to Descriptive Linguistics, New York 1955, pp 266ff), where he analyzes particularly problems of so-called redundancy. Redundancy is the difference between the theoretical capacity of a given code (computed on the assumption that all units of the code can combine in a message without limitation, i.e., that all

possible groupings of signs or their components can be used) and the average quantity of information contained in real messages. Most codes have various limitations, various nonpermissible combinations of individual components (e.g., in the Morse alphabet code several pauses cannot follow one another; in natural language various such limitations are relatively frequent). Gleason (pp 275ff) points to certain principal sources of redundancy in spoken English, such as the difference in phoneme frequency; limitations on phoneme groups; the nonexistence of more than half the morphemes which, from the standpoint of phonological structure, could exist; differences in morpheme frequency; limitations on morpheme series; semantic limitations (the greater or lesser probability of the content of a response). The concept of redundancy in language is in part close to the concept of neutralization, particularly as B. Trnka formulated it (cf. most recently Voprosy yazykoznaniya 1957, No 3, p 48 on morphological neutralization). V. Skalicka proposed a similar problem from a different standpoint (CMF 26, 1940, p 24-29), in which he used the term "retardation" concerning groupings such as "they fought a fight" ("reduplication"), or "bird of legs" ("subsummation"). Here Skalicka is interested in the relations among semes, the elements of linguistic meaning, and not the redundant elements of the means of expression. The redundancy of language reduces its theoretical effectiveness as a "code," but it must be associated with linguistic structure and is very important for the comprehensibility of speech under difficult transmission conditions, with noise, etc.

So-called psycholinguistics has appeared in the USA as a new branch of linguistics, uniting linguistics with information theory and with the psychological theory of learning (cf. particularly the symposium Psycholinguistics, Supplement to the International Journal of American Linguistics 20, No 4, Baltimore 1954). The adherents of this school are using the new point of view to treat problems of the units of the linguistic system and linguistic changes, the learning of the mother tongue and foreign languages, etc. At present it is difficult to say what significance this new method will have for linguistics; there is no doubt, however, that the collaboration of linguists with psychologists, mathematicians, and communications engineers is very important for the further development of linguistics. But the problems involved here should be examined broadly and from various angles; their limitation to so-called pure linguistics is now a thing of the past. (Cf. O. S. Akhmanova, O psikholingvistike /On Psycholinguistics/, Moscow 1957, page 4).

The first person to attempt an overall application of mathematical methods -- statistics and information theory -- to linguistic problems was the English scientist G. Herdan, in the book Language as Choice and Chance (Groningen, 1956). Here the author assembles the results of previous detailed work. He sets himself no small task: he wants to "make a science of linguistics" by mathematical treatment, and to show that statistical relations (based on chance) are more basic than causal relations (based, as he says, on choice). Let us consider the problems in his work.

In investigating individual style in Chapter I of his book he bases his work on the numerical relations of the occurrences of words in the speech of a particular author. Here he is looking for a characteristic quantity independent of the length of the text under investigation: he is not interested in a basic statistical count based on the number of occurrences of particular words in the given text. The author perfects Yule's older method, which we mentioned before, and improves on his formulas. Here we cannot evaluate his work mathematically. But it must be noted that what the author takes as the basis of computation, as something dependent on selection by the speaker (i.e., the numbers of words with various frequencies) is seen in Chapter II of the book to be a matter of chance. Where he examines the norm of the language and the relationship of speakers to the norm it is understood mathematically that their differences do not exceed the limits of chance deviations. This shows that chance and "selection" are not contradictory, mutually exclusive attributes of various phenomena, but only various aspects: each thing has its causes which from one point of view appear as necessary, and from another as chance. But the author does not dwell on this aspect of the matter, and in this particular case actually leaves undecided whether his value v_m is characteristic for the individual (as we should assume from Chapter I) or for the given language (as is stated on pages 100 and 110).

In Chapter II the author proves that the element of chance in language is decisive, that this is the basis, for instance, of the stability of occurrence of phonemes (cf. pp 66ff; of course he does not include poetry), the stability of occurrence of grammatical forms and groupings, and, above all, the stability of the relationship between the length of a word and its occurrence. (According to Herdan, as other language statisticians believe, these quantities are inversely proportional, which is of course far from being a precise statement). The linguistic system (langue) in his opinion is not only a set of linguistic elements, but also contains the probability of occurrence (page 79). We believe that this thesis is not completely new, since linguists have never believed that different grammatical and lexical elements are on the same level, while frequency here is connected, albeit not directly, with the hierarchy within the language system. The mathematical formulation of that fact is certainly very important, and when developed further can yield new results. But the probability of occurrence here is somewhat organically advanced as a third relation, alongside de Saussure's syntagmatic and associative relations. The probability of occurrence of linguistic units depends not only on the langue, but principally on the relationship between langue and parole (if we understand these terms to refer to the linguistic system and to its use in the formation and reception of utterances, respectively). Herdan's view ultimately, it seems to us, again helps to demonstrate that parole is not a purely individual phenomenon (as opposed to the collective linguistic system), as F. de Saussure held (cf. his Cours de linguistique generale, second edition, Paris 1922, page 30).

In discussing information theory and its application to linguistic problems in Chapter III, Herdan frequently emphasizes that the breakdown of the linguistic system into binary relations is not only convenient for various operational procedures, but is basic to the linguistic system; but no convincing arguments are given. The author then seeks in speech properties permitting it to be considered as an ergodic process, capable of treatment by the methods of information theory. This is connected with problems of chance as a stabilizing factor in language, problems which were analyzed in Chapter II. The answer to this is that language surely has these features, if we are referring to those aspects of language to which the attention of the speaker and the listener are not directed: these, aside from artistic utterances and certain special cases, are the occurrence of phonemes and their combinations, the occurrence of words of various lengths. But as regards the meaningful aspect of speech, it is often observed that the methods of information theory cannot be used en bloc (cf. the views of Goldman and Neidhardt, mentioned above, and Kramsky's article, to follow). Thus the meaningful aspect of the response is not ergodic, as the expressed reality is not ergodic, since the individual facts here are interdependent. On the other hand the occurrence of phonemes in ordinary speech is also limited by adjacent and nearby phonemes, but not by remote phonemes, so that the phonological aspect of speech can be treated directly, one can take into account chance occurrence and its laws. He is thus working mathematically with data on language and not with data on reality, on the meaning of the response (with "linguistic" and not "semantic" information, in Goldman's terminology). This is one of the basic difficulties causing a great deal of misunderstanding in the use of information theory in linguistics. Herdan does not deal with these problems at all, nor does he define clearly the limits of use of information theory. He promises that he will show the possibility of treating word meanings, but he does not do so fully (cf. a critique of Herdan by R. W. Brown in Language 33, 1957, particularly pp 178ff. Other errors in Herdan's book have been pointed out by J. Kramsky, Philologica Pragensia 1, 1958 p 89; and M. Halle, Kratylos 3, 1958, p 20ff).

The fourth chapter of the book contains various discussions of problems of duality in language. Here he analyzes the relations between system and occurrence, denoter and denoted, grammar and lexicon, spoken and written language, etc.

At several points in the book Herdan, like other mathematicians, points out the precision of mathematical calculations, which is not to replace the intuitive guessing customary with linguists. This point of view is partly justified; but there is a danger here that in using mathematical procedures we will lose sight of the fine distinctions of which linguists are well aware but which are not easy to express mathematically; for example, in dictionary statistics it is very difficult to determine proper criteria for counting auxiliary words, etc. Mathematical results, even though they may be accurately computed, lose their reliability if

they are based on inaccurate data and thus conceal important differences among the phenomena being treated. And some of Herdan's conclusions are quite unreliable for precisely this and similar reasons. For example he frequently mentions the inaccuracy of his starting data, but nevertheless considers the results of his computations to be completely useful. Or in his reasoning he warns that he is dealing with only one aspect of a particular problem, but in his summary he proclaims the general validity of his results. In comparing linguistic systems for efficiency, for instance, he admits on page 190 that his conclusions concern only the relationship between length and frequency of words (he is interested here in the gradation of words of various syllable numbers) but only five lines later he formulates a conclusion which is quite general, whereby different languages display different degrees of efficiency in coding. English, in which the occurrence of one-, two-, and more-syllable words declines regularly, is, in his opinion, more efficient than German, followed by Russian with a small difference between the occurrences of one- and two-syllable words.

Nor is it entirely clear whether Herdan, in comparing English, Russian, and German, is including all words, even auxiliary words. This would give a distorted picture, since English auxiliary words very often correspond in function to Russian endings, and one cannot conclude that English is more efficient because (in connected text) it has many more one-syllable words, whereas both groups appear with about the same frequency in Russian. This shows how important it is to make a careful linguistic analysis, in this case particularly a typological analysis, before applying mathematical treatment. In this connection the author adopts Jespersen's theory of the tendency to analytic development (which he formulates quite inaccurately as a tendency "to isolate syllables into independent words," page 187) and is unaware that if this tendency operated for many thousands of years, and if development were otherwise governed by chance, all languages would already be fully analytic, if opposing factors were not equally strong, and thus as necessary as analytic tendencies (cf. Skalicka's discussion of a similar overestimation of the tendency to analogy, Sovetska jazykoveda 5, 1955, No 2, pp 85ff).

Herdan overlooks other important criteria which should be sought and developed. We may merely mention several possibilities: It is, for example, an indication of the efficiency of a code if it can transmit a message with a smaller number of signs (or sign components) than other codes. One can thus compare the lengths of semantically similar messages in various languages. To be sure, Herdan mentions this criterion (page 181), but he believes that even in this respect English is more efficient than other languages because it has many short words. Such considerations, however, overlook the fact that English often uses auxiliary words where other languages get along with some form of a basic word (cf. P. N. Savitskiy, who states in his unpublished article Resursy szhatosti russkogo yazyka /Background of the Conciseness of Russian/, that for

these reasons Russian is more succinct than English). In general it should be assumed that inflectional languages, which often include several meanings in one morpheme and have few auxiliary words, are more succinct than other languages. One should also take into account the economy and redundancy of syntactic combinations in various languages, differences in denotative and descriptive names (in this connection even O. Jespersen recognizes the disadvantages of English, cf. his Growth and Structure of the English Language, 9th edition, Oxford 1946, pages 120-139), and finally the problem of homonymy. In an efficient code it is surely assumed that the same sign in the same context has only one meaning. Homonymy interferes with effectiveness, and must be dealt with if we are considering the tendency to become analytic and monosyllabic as basic in the perfection of languages, since when words are reduced to one syllable the number of homonyms necessarily increases.

Linguistics sorely needs mathematical treatment of these and other questions. But the fact that one computation, even though extensive, can be used to draw hasty general conclusions can only discredit the role of mathematics. If we are to achieve precision within our own discipline then we must respect the requirement of precision in other fields, and we cannot be satisfied with mere guesses and inaccuracies once we leave the bounds of our own field. Herdan criticizes Marxism (his concept of Marxist dialectics, which he gives briefly on pages 291ff, and which is concentrated on Hegel's triad, is very naive) for abandoning the principle of the immanent examination of individual fields and for, as he says, introducing duality from language into the physical world without justification. He himself, of course, cannot meet the demand for immanent examination; he does not adhere to "pure mathematics" or linguistics, and dabbles here and there in history. For example the cause of the "decline of German" in developing "from Goethe to Hitler" is, in his opinion, the fact that the "duality of langue and parole" was destroyed (pages 292ff). He seeks the historical cause of these changes in "errors of the intellect and character of the nation." Thus ends the "precise method" of the mathematician once he leaves his own field. But other fields need from mathematics precisely an improvement in the accuracy of their own results, and not accurate mathematical support of such foggy and scientifically rejected guesses.

As we have mentioned, Herdan seeks to show the importance of change and of statistical laws in language, as against the older concept of traditional linguistics, which emphasized more "selection" with its causes, and thus sought causal laws. A positive aspect of this concept, in our opinion, is that it opposes subjectivistic explanations, which overemphasize the influence of the individual on the development of language. But if we consider the well-known Marxist view, according to which necessity and law are manifested in series of chances (cf. Marx and Engels, Collected Works, page 336 of the 1950 Prague edition, in the conclusion of Engels' The Origin of the Family), we see in random processes the opportunity to seek causal laws. Statistical laws

are surely a necessary aid here, but we have no reason to be content with them. Herdan himself -- although it does not correspond to his concept or to the capabilities of his method -- frequently poses the question, either directly or between the lines, of the cause of a particular statistical relationship found by computation. For example, see pages 107 ff, on relationships in the linguistic norm; and page 185 on the "cause" of the fact that word length in Turgenev is distributed similar to that in German, which he says is because Turgenev was influenced more than other Russian writers by western literature (!). And in determining the effect of Naziism on the development of German Herdan seeks historical causes for statistical relations, as we have noted above.

Important for linguistic development is "collective selection," which from the standpoint of individuals and their will is a chance matter just because it is the resultant of many individual selections (even this individual selection is usually not completely conscious). Therefore the development of language does not progress toward a previously defined goal, it is not teleologically directed (cf. Engels' discussion of the development of society in Chapter IV of his L. Feuerbach). Herdan's emphasis on random elements in language can be welcomed from this standpoint, but not as a support of causal laws.

From our survey of Herdan's book one may conclude that mathematical methods are a necessary tool for linguistics, but one which we do not yet know how to use properly. It is perhaps unnecessary to emphasize that it is the linguist's job to master this tool and to collaborate closely with the mathematician. We cannot expect the mathematicians themselves, without collaboration, to master correctly the complex questions of the basic relations among linguistic phenomena. G. Herdan has made numerous good suggestions for the mathematical treatment of language, and it would not be proper simply to reject his book because in this, the first general treatment of these problems, the author was unable to apply his ideas rigorously.

We believe that one cause of confusion in the application of information theory (and communication theory) to language lies in the fact that language is frequently considered simply as one code, placed on the same level as, e.g., Morse, the teletype system, algebraic symbols, etc. (This does not mean that in problems of information theory the specialists have completely overlooked the difference between language and code; e.g., S. Goldman, in the above-mentioned article, distinguishes between the two concepts, on pages 335, 339, 365, etc.). It is overlooked that language (i.e., natural language, not various artificial, logical, symbolic, and other languages) is connected directly with thinking and other aspects of intellectual activity. If necessary we can express ourselves (and perhaps even think) in various codes, but language (i.e., the native language) is the means of communication in which we are not only used to speaking and thinking, but in which out thinking was formed. Messages can be

translated into various codes as the situation demands: individual codes have certain advantages in comparison with language, since messages can be better transmitted over long distances, can be preserved indefinitely, can be kept secret, etc. The concept of code of course includes means of communication which are not concerned with communication between humans. But language is something more basic to our thinking. Therefore translation from one language to another is not the same thing as deciphering a code. (Cf. D. Y. Panov, Avtomaticheskiy perevod /Automatic Translation/, second edition, Moscow, 1958, p 50. On the basic position of language with respect to various systems of signs cf. J. H. Greenberg, op. cit., footnote 18, page 64.)

Furthermore a difference is frequently overlooked which we might call the difference between primary and secondary codes. A secondary code (telegraph, teletype, voice writing, etc.) is basically dependent on language; its various signs or groups of signs represent individual linguistic elements (phonemes, words). For instance the grouping .- in Morse represents the letter a which in the Czech alphabet represents the Czech phoneme /a/. The /a/ is no longer a unit of any such code: we cannot say that, alone or in combination with other phonemes, that it corresponds to the elements of any more basic code. Primary codes, on the other hand, do not correspond so directly to language (e.g., picture writing, systems of lights at crossroads, "the language of flowers," etc.) although they are often subordinate to language (their elements occur as symbols of verbal meaning).

In communications engineering we are usually interested in the codes which we have here called secondary. In various computations, whether for deciphering or for measuring the efficiency of codes (using the concepts of entropy and redundance) we are interested in the relation of systems of signs to the more basic systems on which they are based. The efficiency of a telegraph code can be measured in terms of its relation to the language in which the telegraphing is being done, or, more precisely, to the written form of that language. (Cf. A. A. Kharkevich, Ocherki obshchey teorii svyazi /Outlines of the General Theory of Communications/, pp 64 ff, in which he says that the efficiency of Morse must be measured in terms of the language for which it is being used. The shortest symbol is for the letter e, which is the most frequent letter in English. In Russian, however, o is more frequent, and this is (inefficiently) represented by a longer sign. Thus Morse could be improved for Russian.)

If information theory is to be useful for linguistics as a whole, and not only for its individual elements, we must also consider the relationship between the auditory and semantic aspects of language. Entropy and the efficiency of the linguistic system would then be measured not only in terms of its phonemes and letters, but other elements would come under consideration, such as the comparison of various methods of auditory (or written) representation of the same system of grammatical categories (or grammatical meanings, semes, functional morphological units); or a comparison of various systems of grammatical categories

with the expressed fact and with thinking. After all, every grammatical category itself is a complex fact, fulfilling various functions under various circumstances; individual languages do not agree in grammatical categories, and before we express ourselves concerning their communicative value we should take these relationships into account.

In examining the efficiency of various languages as means of communication it is improbable that languages can be arranged in a series, from imperfect to mature, since in millenia of development they have passed under influences from various factors, have changed fundamentally, and through long use have been adapted to the fulfillment of the tasks confronting them. It is much more probable that various criteria for language efficiency (cf. discussion above) intersect, that each language has its advantages and disadvantages.

In conclusion we can say that mathematical methods introduce into linguistic work not only an improvement in precision, but many new ideas. They contribute to connecting our science with many other fields of inquiry and thus enrich it by providing new points of view. But without close reliance on previous linguistic methods mathematical treatment of language frequently leads to undesirable distortions. Thus it is all the more urgent for linguists themselves, in collaboration with mathematicians, to master new methods and provide proper guidance in their application.

COMMUNICATION THEORY

Slovo a slovesnost, No 1,
1959, pages 55-66.

Jiri Kramsky

I.

The theory of communicative speech is essentially the same as "communication theory." This should be distinguished from so-called "information theory" although, because of the close connection between the two theories, this distinction is usually not maintained. Information theory is a narrower concept, however, than communication theory. In using the term "theory of communicative speech" /teorie sdělné promluvy/ (suggested by Professor B. Trnka) we are emphasizing that we are interested in communication by means of language. Since, however, we wish to explain the principles of a theory which is much broader than speech, we shall generally use the term "communication theory" or "information theory."

As the title of one of the fundamental articles on information theory -- Mathematical Theory of Information, by Claude E. Shannon and Warren Weaver -- indicates, this is a mathematical theory which bears, as we shall see, a very close relation to linguistics. In this study we shall attempt to explain communication and information theory as far as possible in linguistic terms, and will try to specify its importance for linguistics. Therefore we shall limit the mathematical terminology to a minimum.

In broader terms information theory is a part of so-called cybernetics, which is, according to N. Wiener, "the science of relations." (Cf. Cybernetics, or control and communication in the animal and the machine, New York, 1949; and, by the same author, The Human Use of Human Beings, Cybernetic Society, New York, 1956.) As Werner Meyer-Eppler writes (Information Theory, Die Naturwissenschaften 39, 1952, p 341), the first goal of every scientific inquiry into processes is to obtain valid information on process itself. If this is done without human collaboration or if the human is recruited only for experimental purposes, the process is called measurement (Messung). We speak of measurement also when the process derives from a human but occurs independently of the free will of that individual. A process caused consciously, in order to communicate information to another individual, contains more delicate qualities than mere energy; we call them information. Through the sense organs this information may be transferred in various ways. In all higher forms of communication, whose function is not merely appeal or communication, but is intended to represent a material or sensory content, signs (Zeichen) are used which symbolize the meaning of that which is to be represented, or form relations among meanings. These signs comprise a reserve of prearranged material, a repertory, from which information is selected.

Communication by means of speech is surely older than recorded history. But not until the 20th century, the age of communication, was there a systematic scientific examination of the problems of communication. In 1924 and 1928 H. Nyquist published two studies of telegraphic communication, and R. V. L. Hartley published his article Transmission of Information, in the Bell System Technical Journal, 1928, p 535. This work was the basis for Shannon who, together with Norbert Wiener, is generally considered the creator of communication theory in its present form. Shannon himself says that he is indebted to Wiener for the philosophical basis of his work, and Wiener in turn says that Shannon is responsible for the independent development of such basic aspects of information theory as the concept of entropy. Other investigators who have contributed to the development of communication theory, particularly linguistically, include John Lotz, R. M. Fano, Oliver H. Straus, and S. S. Stevens, whose contributions are published in Vol 22 (1950) of The Journal of the Acoustical Society of America; B. Mandelbrot, Structure formelle des textes et communication, Word 10, 1954, pp 1-27; P. Guiraud, Langage et communication, Le substrat informationnel de la semantisation, Bulletin de la Societe de linguistique de Paris 50, 1954, pp 119-133. The journal Die Naturwissenschaften contains the article cited above by Werner Meyer-Eppler on information theory. Information theory is also dealt with in detail by G. Herdan in his book Language as Choice and Chance (Groningen, 1956). Considerable other work, including George A. Miller, Language and Communication; Gunnard Fant, Discussion Symposion on the Applications of Communication Theory (London, 1952); D. Gabor, Lectures on Communication Theory (1951), MIT; etc., are not accessible to us. Therefore the survey which we can present on progress in information theory is necessarily incomplete.

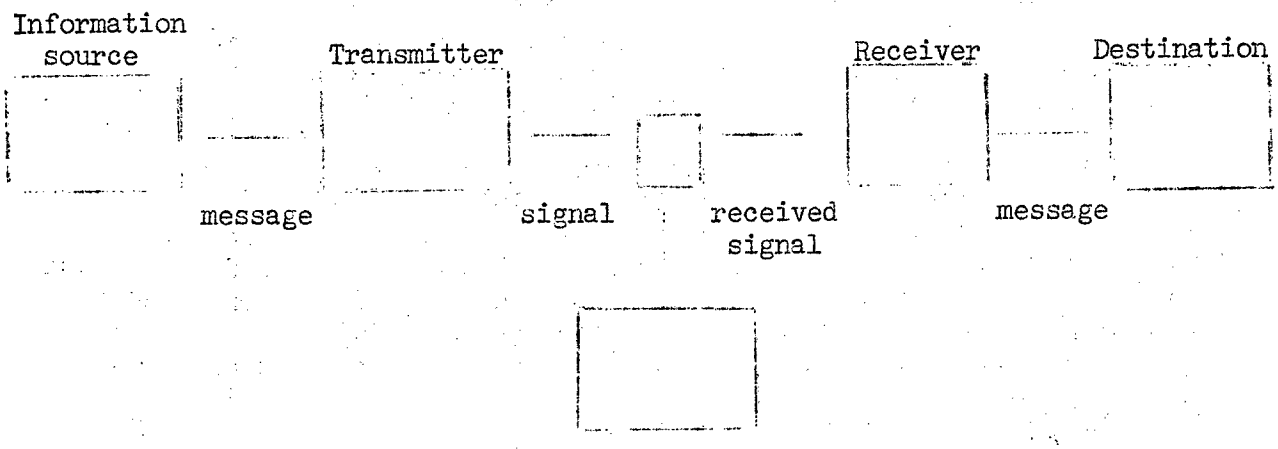
We should add that communication theory has received considerable attention in the Soviet Union, particularly in connection with the problems of machine translation (e.g., Kuznetsov, Lyapunov, Reformatskiy, Zinder, etc.). Attention has also been directed toward quantitative linguistics, as shown by the conference on linguistic statistics held in October 1957 in Leningrad, and reported by V. A. Uspenskiy in Voprosy yazykoznaniya /Problems of Linguistics/, No 1, 1958, pp 170-173.

The fact that there is lively interest in Czechoslovakia in information theory is shown by the conference on information theory, statistical decision functions, and random processes, held in Liblice in November 1956. The Czech scientists, principally those working at the Institute of Radio Engineering and Electronics, together with foreign scientists at the conference (D. Blackwell, B. V. Gnedenko, C. Rajski, H. Hansson, etc.) gave a number of papers on the mathematical problems of information theory. (Cf. Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Prague, 1957). As regards the linguistic approach to information theory we are still becoming acquainted with the problems. The purpose of the present

study is to acquaint the reader with certain problems of information theory from the linguistic standpoint. The fact that we have a view on certain problems in information theory is intended to evoke discussion and arouse interest in Czechoslovakia in independent research on linguistic problems from the standpoint of this theory.

The concept of communication and the communication system. S. S. Stevens defines communication (A Definition of Communication, JASA 22, 1950, pp 689-690) as a discriminatory reaction of the organism to a stimulus. This is a broad definition, embracing not only written and spoken language, but music, theater, ballet, art, and in fact all human behavior.

By a communication system we mean the following system:



It is made up essentially of the following five parts:

1. A source of information which makes up the message which is to be conveyed to the receiver. The information source selects the desired message from a collection of possible messages. The selected message can be written or spoken words, music, pictures, etc.

2. A transmitter, which processes the message to produce a signal so as to make it transmittable. In language this is words, which convey the sequence of thoughts; in written language it is letters or other symbols; in telegraphy it is a sequence of dots, dashes, and pauses.

3. The channel, i.e., the means used for conveying the message.

4. The receiver or recipient, which usually performs the opposite operation to that done by the transmitter: it reconstructs the message out of the signal, e.g., it deciphers the message.

5. Destination, i.e., the person or thing for whom the message is intended.

Let us give concrete examples. In the case of telephony the communications channel is the wire; the signal is the electric current in the wire; the transmitter is the telephone transmitting instrument. In telegraphy the transmitter enciphers the written words into a sequence or interrupted current of various lengths (dots, dashes, pauses). In spoken language information source is the brain, the transmitter is the voice mechanism forming a variable sound pressure (signal) transmitted through the air, which is the channel. The receiver changes the signal back into a message and sends this message on to its destination. In the case of spoken language this receiver is the ear and the associated nerves in the recipient.

Other factors may enter the transmission process and disrupt transmission. This can result in distortion of sound (in a telephone), distortion of an image (in television), etc. Such changes in the transmitted signal are called noise. We shall discuss the significance of noise later.

The basic concepts of communication or information theory which must be clarified are information, entropy, redundancy, and the principle of information transmission.

The concept of information. The word "information" is used in a special sense in this theory, and must not be confused with the concept of "meaning." Information does not refer to the semantic content of a message. For purposes of communication the meaning of a message is irrelevant. In actuality two messages, one meaningful and the other meaningless, can be completely equivalent from the standpoint of information.

The word "information" concerns not so much what we say as what we might say. This means that information is a measure of our freedom in the choice of a message. The concept of information refers not to the individual message (as the concept of meaning does) but rather to the situation as a whole; in a given situation we have a certain freedom of choice in selecting a message, and this is considered the standard or unit of information.

The quantity of information is defined in the simplest cases such that it is measured by the logarithm of the number of possible choices.

Since it is easier to use logarithms to the base 2 than to use ordinary base-10 logs, in the case of two possible choices information is proportional to the logarithm of 2 to the base 2. This unit of information is called a "bit" (the word was first used by J. W. Tukey and is a contraction of "binary digit"). If numbers are expressed in the binary system there are only two digits, 0 and 1, just as there are ten digits, 0 through 9, in the decimal number system based on 10. Zero and unity can be used to represent symbolically any of two choices, so that it is natural to connect the "bit" with a situation in which there are two choices. If, for instance, we have 16 alternative messages among which we can choose, then, since $16 = 2^4$, and $\log_2 16 = 4$, then the situation is characterized by four bits of information.

The foregoing concerns only simple situations. More natural and more important are situations in which the information source chooses from some group of elementary symbols, wherein the sequence chosen makes up the message. Here, however, the role of probability becomes important, since these choices are, at least from the standpoint of the communication system, affected by probabilities, interdependent probabilities, since at each stage the process depends on the preceding choices. For example in English, if the last-chosen symbol is the, then the probability that the next word will be an article or a verb is very small. This force of probability is felt over greater stretches than two words. For instance the probability that the word following the three words in the event will be that is quite high, while the probability that it will be elephant is very low.

The process in which a sequence of symbols is formed according to certain probabilities is called stochastic, and the special case of a stochastic process in which probabilities depend on preceding events is called a Markov process. A special class of Markov processes, which are of prime importance for communication theory, are so-called ergodic processes. An ergodic process is one forming a sequence of symbols in which each symbol tends to be representative of the sequence as a whole.

Entropy and redundancy. The concept of entropy, introduced into physics almost 100 years ago by Clausius and closely connected with the name of Boltzmann and that of Gibbs, who gave it profound significance in his classical work on statistical mechanics, has become a basic concept of physics. In physics entropy is a measure of the degree of mixing or randomness of particles, elements of a physical system. It is actually a disordered state of the system.

That information should be measured by entropy is natural when we consider that information, in communication theory, is connected with the number of choices, or the freedom of choice, which is available to us in the construction of messages. Thus, for a communication source, as for a thermodynamic set of elements, we can say: "This situation is highly organized, it is not characterized by a high degree of randomness or choice. Thus the information content (or entropy) is low."

When we compute entropy (or information, or freedom of choice) of a certain information source we can compare it with the maximum value which that entropy could have assuming that the source will always use the same symbols. The relationship of actual entropy to maximum entropy is called the relative entropy of a source. If the relative entropy of a certain source is, e.g., 0.8, this means roughly that this source in its choice of symbols making up messages is about 80% as free as it might be with the same symbols. Subtracting the relative entropy from unit we obtain redundancy. This is the fraction of the structure of the messages which is determined not by the free choice of the transmitter but rather by the statistical rules governing the use of the symbols. This fraction of a message is redundant in the sense that if it were

lacking the message would still be essentially complete; or at least it could be completed. The redundancy of English is about 50-60%, so that about half the letters of words which we choose in writing or speaking is subject to our will and about half is controlled (although we usually are not aware of this) by the statistical structure of the language.

The value of entropy, which is symbolized H, is maximum when two probabilities are equal (i.e., when our choice is completely free) and tends toward zero when there is no freedom in choice.

When we have more than two cases of freedom then entropy is maximum when the probability of different choices is almost the same. Let us assume, on the contrary, that one choice has a probability close to unity, so that all the other choices will have probabilities close to zero. This is a situation in which a person is very strongly influenced to make a certain choice and thus has little freedom of choice. The entropy, H, in such a case is very small, or, rather, the information content (freedom of choice, uncertainty) is low.

We can see that information content is greater the more nearly equal are the probabilities of various cases of choice. If all cases of choice have the same probability then the greater the number of those cases the greater will be entropy, H. The information content is greater if we select freely from a group of, say, 50 standard messages than if we choose from only 25 messages.

Transmission of information. For information to be transmitted a "communications channel" is necessary. As we have already said this is the means used to convey the signal from the transmitter to the receiver. It may consist of several wires, cables, radio-frequency radiations, light rays, etc.

Let us consider first how information is conveyed. In the case of telephony an audible voice signal is changed into something which is clearly different but clearly equivalent (the variable electric current in the telephone wire). But the conveyor can perform a much more complicated operation on the message to create a signal. It may, for instance, use a code and encipher the message, using a sequence of numbers; these numbers are then sent through the channel as a signal. The function of the transmitter is thus to encipher the message, and that of the receiver to decipher the message. The most efficient type of cipher is one which is most suitable for the channel, when the signal entropy equals the channel capacity. That is to say we must take into account the capacity of the channel to deal with various signal situations. In telegraphy, for instance, there must be pauses between dots, dashes, and groups of dots and dashes, since otherwise the dots and dashes would be indistinguishable.

If we introduce noise into the channel the received message will contain a certain amount of distortion, certain errors; it will display an increase of uncertainty. This is a different type of uncertainty, however, than we have encountered hitherto. The uncertainty arising from the free choice in messages is desirable, whereas the uncertainty

resulting from errors or noise is undesirable. This leads to a practical conclusion: Since English is about 50% "redundant," it should be possible to save about half the time consumed in telegraphy by using a proper enciphering process, assuming that the signals were transmitted through a noise-free channel. But if the channel contains noise there is a certain advantage in using an enciphering process which does not eliminate all redundancy, since the very redundancy helps to overcome the effects of noise.

Connected messages. In addition to messages made up of discrete symbols, such as words made up of letters, sentences of words, melodies of notes, messages can also be connected, such as connected speech with constant variations of volume and energy. In this case the theory is no different; it is merely more difficult and mathematically more complicated. Where there are changes they are smaller.

Generality of communication theory. Communication theory is so general that there is no need to state what type of symbols are being used, words or letters, spoken words, notes, symphonic music, or pictures. It is also so profound that the relationships which it reveals can be applied to all these and other forms of communication. The basic relations apply generally, no matter what the special form of a given case.

The theory is actually the basic theory of cryptography, which is of course a form of coding. Communication theory also contributes to translation from one language to another, and is also closely associated with the theory of computers.

We have briefly explained the basic concepts of communication theory which are necessary for an understanding of the theory. This is also the basis of subsequent explanation of the applications of communication theory to linguistic problems.

II

Linguistic interpretation of entropy and redundancy. In this study we of course cannot deal with all the linguistic problems to which communication or information theory can be applied. Therefore we shall concentrate mainly on a linguistic interpretation of entropy and redundancy.

As we have already noted, we speak of entropy when the elements of a physical system are mixed; while if they are separate, whether they are active or at rest, we speak of the "ordered state" of a system.

We may assume a similar distinction for a linguistic system. Here an "ordered state" corresponds to the deliberate selection of words in which each word is selected independently /sic/ of the others. Of course if we write at length of a certain subject we must take into account the words and grammatical relations which we have used, including those which we shall later use, in order to avoid undesirable repetition of expressions. It may be objected that if we choose an expression in order to avoid another expression this is already a determining

act. We can answer that this is a different method of determination. We do not choose an expression because it is the best possible one but because it is as possible or convenient as another which we wish to avoid.

According to Shannon (Prediction and Entropy of Printed English, The Bell System Technical Journal 30, 1951, pp 50-64), entropy is a statistical parameter which measures how much information is created on the average for each letter of text in the language. If the language is transposed to binary digits (i.e., 0 and 1) in the most efficient manner the entropy, H, is the average number of binary digits per letter of the language. On the other hand redundancy measures the quantity of constraint exerted on the text by the structure of the language: e.g., in English, the high frequency of the letter E, the strong tendency of H to follow T, or the tendency of U to follow Q. It has been computed that, if we take into account the statistical effects which are not exerted over more than eight letters, entropy is roughly 2.5 bits per letter and redundancy about 50%.

By combining experimental and theoretical results we can compute the upper and lower limits for entropy and redundancy.

It follows from this analysis that in literary English the statistical effects (up to 100 letters) reduce entropy to about one bit for the letter S, corresponding to a redundancy of about 75%. Redundancy can be greater if we consider structure extending beyond the paragraph, chapter, etc. Of course when the length increases the parameters become more uncertain and are much more critically dependent on the type of text.

In determining redundancy Shannon and other investigators have considered only the limitations resulting from the structure of the language, and have not considered other limitations imposed by the listener and the situation in which the speaker finds himself. Therefore, as F. C. Frick and W. H. Sumby (Control Tower Language, JASA 24, 1952, pp 595-596) write, an informational analysis has been made of what the authors call a "sublanguage," the language used in controlling the flight of aircraft at airfields. If the situational and linguistic contexts are taken into account, the redundancy increases to 96%.

Very instructive experiments have been conducted with redundancy in printed text, as described by Werner Meyer-Eppler (op. cit., pp 341-347). If, in a German text, we replace all the vowels with a dash, we can still understand a sentence without too much difficulty despite the considerable distortion introduced: M-N K-NN D--S-N S-TZ --CH -HN-V-K-L- L-S-N.

The reason for the high redundancy of German text words should be sought in the fact that the freedom with which vowels and consonants make up words of a certain length is quite limited. Thus in the case of M-N only one vowel is possible, i.e., MAN, while the other possible forms (MEN, MIN, MON, etc.) do not occur.

If we present unsystematically distorted texts to a large number of experienced persons we can obtain numerical values for redundancy. It has been found that in German printed text more than half the letters can be guessed, so that the redundancy exceeds 50%.

Text distortion need not involve elimination of an entire symbol. For instance the word NACHRICHTEN has been distorted variously. It was found that when half of each letter (either the upper or lower half) was removed the word remained completely legible; here, too, redundancy equals 50%. Different amounts of text distortion cause different degrees of legibility, up to complete illegibility. Handwriting has much less redundancy; when part of a symbol is missing the text is almost completely illegible.

The problem of compression of printed text is interesting. If we send a telegram in German we can save paying for more than half the letters, since the recipient can guess the missing letters on the basis of redundancy. Shannon has shown how a certain type of coding can eliminate redundancy without loss of information and simultaneously compress a message to minimum size. This process is demonstrated by Meyer-Eppler with the word NACHRICHTENVERBINDUNGEN. In the first step the most frequent letter pairs are replaced by less-common individual letters, while the less-common letters are replaced by the least-used letter pairs. P. Valerio lists as the six most-frequent German digrams: EN, ER, CH, ND, DE, IE.

If we replace these digrams by the least-used individual letters: Y X Q J P V, and replace the only one of these individual letters appearing in the word NACHRICHTENVERBINDUNGEN, i.e., V, by the digram HC, then we produce NAQRIQTYHCXBIJUNGY.

The first step in coding has reduced the word length from 23 to 18 letters, without touching the information contained in the word. The idea of replacing the most common elements in speech and expressing them with simple signs, and using more complicated symbols for less-frequent elements is also found in Morse code. Of course this code is based on the frequency structure of English and uses the shortest symbol, a dot, for the letter E, which is the most frequent letter in English.

Efficiency of a code. A code must be efficient and economical to meet the demands of communication. One of the most important axioms in this business is the requirement that the smallest combination of elements be used for the most frequent linguistic form, and that the length of the combination be in inverse proportion to the frequency of occurrence. It is quite possible for a coding system to be good as regards the use of the combinatorial method, but to be unsatisfactory in terms of labor economy. Morse code, for instance, although it makes full use of the possibilities inherent in the two-letter code, does not systematically apply the above-mentioned economic principle. This principle requires that all the short combinations be used, and that the shortest combinations be used for the most frequent letters. The first

requirement is fulfilled in Morse code, but the second is not, or only partly so. Thus the letter E, which in the main European languages is most common, is represented by a dot; but O and K are both represented by three elements, although O is about three times as frequent as K in English.

Two of the most important features of coding symbols are their duration, on which depends the rate of communication, and the ease of their formation, on which depends the ease of their use. In the case of the spoken word the length of the symbol can be expressed by the number of syllables or sounds, and the ease of use of certain sounds by the number of combinations in which these sounds occur.

A basic feature of coding symbols is their distinctive value, i.e., the fact that they permit the easy and correct distinguishing of different utterances. Whereas the number of syllables is the first and basic feature for distinguishing among words, the phonemes of the language provide for finer distinction among words of the same length.

First we ask whether the number of syllables per word meets the requirements of the dyadic coding system and then whether word length in the given document bears the relationship to the frequency of their occurrence required by the theory of efficient coding and whether different languages characteristically differ in this. In the terminology of information theory this means that we compute the value of entropy, H, for the given distribution of word occurrences in terms of the number of syllables. The greater the value for the given language the more we are entitled to consider the language as a system meeting the requirements of efficient coding. The efficiency of a code is measured by $1-H/H'$, which is a measure of redundancy. The greater this quantity the greater the compression over even distribution. This permits comparing different languages in terms of redundancy, i.e., in terms of the possibility of estimating word length. For English this compression is about 50%, as has been said above, whereas for Russian it is only around 30%, and for German somewhere in between.

Another important problem, investigated by G. Herdan, is the stability of word-length distribution. He condludes that word length in a given language cannot be considered homogeneous. Differences are seen here depending on the stage of linguistic development, dialect, and even individual writers. It has also been found that samples taken from one and the same language differ less than samples from one language to another. Surprisingly it has been found in English text that length increases with time, which apparently contradicts the general tendency to monosyllabicity in English. We can explain this by the fact that the distribution function of the lexicon changes with time. The result of the increasing enrichment of the lexicon is that word occurrences are less concentrated on a few frequent, and therefore shorter, words. This causes an increase in average word length with time.

Herdan also demonstrates the relative stability of word length in a given language at a given time. There is no important difference between the index for the literary language and the colloquial language, and it does not matter whether the material was connected text or isolated words from the same text, whether prose or poetry. The result of investigations is that word length is a decisive factor for word-use frequency: the shorter the word the more frequently it is used.

As regards word length, the history of language is the resultant of two forces: agglutination and isolation. Not including Chinese, which is almost completely isolating, and the Indian languages, which are polysyllabic, most languages occupy a mean position in this respect. Within one and the same language the process toward isolation is characteristic and a powerful factor in its development. Since the languages studied (English, German, Russian) represent different stages of development toward monosyllabicity or isolation, they also represent various degrees of agreement with the requirements for effective coding. Since English has the greatest redundancy, 1-H/H', it can be said that it has the greatest code efficiency; it is followed by German and then by Russian. Herdan believes that the tendency toward monosyllabicity should be sought in the principle of efficient coding. Since coding efficiency must be an important factor in the speed and precision of communicating thoughts, it would appear that the statistical theory of information leads to an unexpected conclusion: that languages differ in this respect. The general view is that each language is more or less perfect in this respect for the people who know how to use it. That the difference in "perfection" of languages should be determined objectively is a new idea of Herdan's. He considers 1-H/H' as a measure of perfection or efficiency of a language for the communication of ideas; according to him English is highest among the three languages investigated, followed by German and Russian. The weaknesses in this part of Herdan's views were discussed above by Petr Sgall.

As regards word length measured by number of phonemes, here too the principle of efficient coding is retained, according to which word length and frequency of occurrence should be in inverse relationship. The redundance of word length measured by number of letters is considerably less than that measured by number of phonemes, which means that estimates of word length expressed in phonemes are more successful than those expressed in number of letters. This was to be expected, since the phoneme and not the letter has the truly distinctive function. The number of phonemes per word must first be taken into consideration, and then certain phonemes. The first approximation leads us to the conclusion that of words with the same number of syllables the most frequently used are those which contain fewer phonemes. The second approximation provides the discovery that not only the number of phonemes (i.e., their number per word) determines the frequency of word-use but also their quality. Among words containing the same number of phonemes some will be used much more frequently than others, and the feature which differentiates these words is different phonemes or phoneme combinations.

Gustav Herdan's most recent article (An Inequality Relation Between Yule's Characteristic K and Shannon's Entropy H, Journal of Applied Mathematics and Physics, Vol IX, No 1/1958, pages 69-73) sharpens the concept of entropy as used in linguistics. The author states first that entropy, H, in linguistics does not measure information, but rather the lack of information, and that negative entropy, or negentropy, should be used for measuring information. Secondly, entropy, H, is a logarithmic quantity, and thus cannot be interpreted in nonlogarithmic terms; so-called bits of information have no true linguistic meaning. The author transforms the traditional expression for entropy into a statistical quantity of general use in linguistics, and calls it the repeat rate. The repeat rate at a minimum equals the antilogarithm of negentropy, or rather the antilog of negentropy is the lower limit of the repeat rate or Yule's characteristic, K.

Using the concept of inequality we can express, for instance, the difference in entropy for the length of English words expressed by number of letters H = 2.628 and word length expressed in number of phonemes H = 2.274, by saying that the repeat rate for a certain word length is at least one in 6.3 words if the word length is measured by number of phonemes. In other words, if we take two words at random from a very long text and determine word length, it is found that word length is repeated much more frequently if it is measured in terms of phonemes.

Translation from one language to another can be considered as the bivariable coding of a message, i.e., as an example of double coding. An identical phenomenon is expressed by two different language codes. It is first coded in the source language, from which the translator must decipher before he recodes in the target language. If entropy is computed for a given quantitative characteristic such as word length separately for each of the two languages, source and target, the values obtained will usually differ from each other and also from the value of the bivariable entropy, i.e., from the entropy computed from the word length of translation equivalents (words representing the same concept in the two languages). More precisely, conditioned entropy, i.e., entropy for word length in translation equivalents of source-language words of a certain length, is less than unconditioned entropy.

In translating from French to English, for instance, the conditioned entropy is 0.923 bit and unconditioned entropy 1.410 bits; from German to English the former is 1.162 bits and the latter 1.648 bits; and from Russian to English the former is 1.153 bits and the latter 2.000 bits.

These relations are not perfectly clear when expressed in terms of entropy in its conventional form. The inequality relation helps us to make the difference between conditioned and unconditioned entropy much more understandable. This relation expresses the differences thus: for translation from French to English the repeat rate of words

in the French original is at least one in 2.6, or about one in 3, and rises in translation equivalents to at least one in 2; for German-English translation the repeat rate of word length in the original is also about one in 3 and in translation equivalents rises again to about one in 2; and for Russian-English translation the repeat rate of word length in the original is about one in 4 and rises in translation equivalents to one in 2 (actually, one in 2.4).

One of the most difficult problems in communication theory is that of machine translation. But this would fill a separate article and we cannot discuss it here. Furthermore we are only beginning to solve the problems of machine translation. (Cf. Petr Sgall's article, above.)

III

The contribution of communication theory to linguistics. Around 1950 the situation was as Oliver H. Straus described it (The Relation of Phonetics and Linguistics to Communication Theory, JASA 22, 1950, p 709): "Today investigators in communication theory are divided into two camps. In one camp are the engineers and phoneticians, who are interested in physical sounds or phones; the other contains the linguists, who are interested in physical sounds only for the purpose of determining, in a language or dialect, small groups of signal units called phonemes, words, and entire utterances." Since that time, however, the two camps have grown much closer together; they have made contact, realizing that they have a good deal to offer one another, and in many cases they have undertaken fruitful collaboration. A large number of experiments have been conducted and theoretical views have been formulated which agree with the views of information theory. Progress in our understanding of the informational aspects of language and speech has contributed in recent years to the emerging science of communications as well as to linguistics.

There is a large number of linguistic problems whose solution has been promoted by information theory, and a number of problems which linguistics has not encountered or which it could not solve with its previous methods.

First there is the concept of "information." As has been pointed out, it is irrelevant to communication theory whether a message has any meaning or not, since the communication engineers are interested only in whether a certain signal gets from the source through the transmitter and the channel to the receiver. They are interested in a purely physical procedure, in the method and fidelity of transmission. On the other hand linguistics is interested in two things: we are interested in whether certain information gets from the information source to the receiver and the destination, which is essentially the same thing that interests the communications engineer. At the same time, though, we are interested in information in the linguistic sense, i.e., in meaning, since linguistic communication is the transfer of ideas

expressed by the spoken or written word, coded in some manner so that it can be transmitted. This is the normal case. It is of course true that a linguistic message need not always carry information in the semantic sense. There are sometimes situations in which linguistic messages contain no semantic information, they have no "meaning" in the usual sense of the word. If two people meet and begin to speak about, say, the weather, this does not necessarily mean that they wish to convey information about the weather, but rather that the conversation is intended to establish contact between two persons. There are many situations in which a conversation has quite a different purpose than to communicate the semantic content of the spoken word. We conclude essentially that the concepts of information are not the same in communication theory and linguistics, and that the relationship of the two must be clarified in both sciences.

The most fruitful concepts of communication theory for linguistics would appear to be the concepts of entropy and redundancy. We are convinced that here we are quite entitled to compare the state of the physical system and of the linguistic system. This is an absolutely basic problem, since it concerns the comprehensibility of linguistic utterance, which has so far been overlooked in linguistics. To be sure functional linguistics has already applied the views of information theory in its basic criterion of relevance or irrelevance of linguistic phenomena and linguistic features, but the concept of redundancy was the first to demonstrate the full significance of this problem. The question remains, however, whether that which is redundant is always the functionally irrelevant, or whether sometimes it is not the functionally relevant. Redundancy is very important for language, since it increases the reliability of linguistic communication and makes communication resistant to many cases of distortion. On the other hand if the conditions for communication are good a large degree of redundance means a waste of time.

Communication by the spoken word is much more complicated than communication via written or coded language. But even this problem has not been overlooked by information theory, even if it has not yet been so thoroughly investigated as other types of communication. Several good articles have appeared, of which these can be noted: R. M. Fano, The Information Theory Point of View in Speech Communication, JASA 22, 1950, pp 691-696; John Lotz, Speech and Language, JASA 22, 1950, pp 712-717; and the short but important article by Norbert Wiener, Speech, Language and Learning, JASA 22, 1950, pp 696-697.

Fano points out correctly that in communication between two persons we must know, in addition to the characteristics of the communication system external to the two people, also all possible characteristics of the two persons at each moment of transmission. Particularly we must know details concerning the activity of their vocal and auditory organs, and the past experience which is stored in their brains. Of course this is only theoretically possible.

In determining the information content of speech Fano sees two possible procedures. One is based on the assumption that the information transmitted to the listener by spoken language is not very much greater than the information conveyed by a corresponding written message. This assumption may of course be criticized since a written message lacks the characteristics of the speaker's voice, and particularly those vocal inflections which can sometimes completely change the meaning of a sentence. There are cases of messages like Yesterday was a fine day! in which the spoken message can have exactly the opposite meaning from the written one. In actuality each message receives a certain amount more information through vocal features which are lacking in the written message. Fano attempts at least a partial justification of this assumption, but it would appear that he underestimates the role of vocal features as regards their effect on information.

The other procedure requires consideration of the process of generation of speech wave. In the case of voiced phones the vocal chords produce a roughly triangular wave of slowly varying amplitude and frequency. This wave is filtered by the vocal tract, which we can consider as a linear network. The transmission properties of this network change with time in approximately the same proportion as the frequency of the triangular wave, so that we can consider the speech wave as a sort of periodic time function. The case of voiceless phones differs in that the vocal chords must be replaced by a noise source. Thus we can assume that the channel capacity required for voiced phones is sufficient also for voiceless phones. Fano puts forth the interesting idea that different speakers use different codes in some sense. These codes are stored in the brain of the listener, who uses the proper code in each instance. We are always learning new codes when we meet new people, particularly people belonging to different linguistic groups. This agrees with the observation that our ability to understand and the effort required for comprehension depend on our acquaintance with the speaker's voice. Furthermore we are often aware of code switching in our brain, particularly when there is a change of language. The identification of the proper code and the learning of a new code are very complicated statistical processes; they require analytic and storage facilities which are available only in the human brain.

Fano's characterization of the communication process through speech could be supplemented by pointing out another important factor affecting comprehension and thus information: the rate of speech. This means not the absolute rate, but the relative rate with respect to the capacity of the listener to react properly to the rate of speech. In different persons and under different conditions this rate differs; more detailed observations could contribute to the general characterization of the communicative process of speech.

It should be noted that the problem of comprehensibility has recently been the object of a number of studies. These include M. J. Pickett, Perception of Vowels Heard in Noises of Various Spectra, JASA

29, 1957, pp 613-620; Davis Howes, On the Relation between the Intelligibility and Frequency of Occurrence of English Words, JASA 29, 1957, pp 296-305; Grant Fairbanks and Frank Kodman, Jr., Word Intelligibility as a Function of Time Compression, JASA 29, 1957, pp 636-641; G. Fairbanks, N. Guttmann, and M. S. Miron, Effects of Time Compression upon the Comprehension of Connected Speech, The Journal of Speech and Hearing Disorders, 22, 1957, pp 10-19; G. A. Miller and W. G. Taylor, The Perception of Repeated Bursts of Noise, JASA 20, 1948, pp 404-411. Czechoslovak writing on speech comprehensibility, principally from the acoustic standpoint, includes the following: J. Vott, J. Vachek, and J. B. Slavik, Akustika hlediste v divadelnim provozu /The Acoustics of the Theater/, Prague, 1943; J. B. Slavik, Akustika kinematografu /The Acoustics of the Cinema/, Prague, 1943; J. Vachek, A Linguistic View of the Soviet Tests of the Comprehensibility of Telephonic Transmission, Sb. praci fak. elektrotech. inzenyrstvi CVUT /Transactions of the Faculty of Electrical Engineering, Czech Higher Technical Education/ 1956/57, pp 459ff. Cf. also Slovo a slovesnost 15, 1954, pp 165ff; 17, 1956, p 40ff, 110ff, 178ff; and 18, 1958, pp 62ff (Hala, Vachek, Borovickova, Romportl).

The information capacity of man has been studied by G. A. Miller, The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, The Psychological Review, 63, 1956, pp 81-97. Cf. also Gorden E. Peterson, Fundamental Problems in Speech Analysis and Synthesis (Reports for the Eighth Intern. Congress of Linguists, Oslo, pp 319ff).

Norbert Wiener is aware of the similarities and differences between the concepts of "information" in engineering and linguistics. In the transmission of linguistic information Wiener recognizes three stages: reception, the phonetic stage, and the semantic stage. Wiener says that we must be very careful with measurement here, since the semantic stage involves the long-term storage apparatus of the brain which we call memory, and we combine statements just made with recollections taken from memory. We believe that Wiener has shown the difference between mechanical and linguistic communication better than any other investigator in communication theory when he writes that what is a clear quantitative statement in mechanical communication becomes a qualitative statement in linguistics. At the same time, however, Wiener believes that there is no basic contradiction between the problems of communications engineers and those of linguists.

John Lotz sees speech as a tool for communicating semantic content, or meaning. Speech can be analyzed into symbol elements which have their own semantic reference; this is so-called semiotic analysis. He proposes the term "semantic spectrum" for it. According to Lotz the following semiotic units are necessary in linguistics: morpheme, word, sentence, and paragraph. These units form a semiotic hierarchy. The auditory resources signaling semiotic units are various; for instance the sustained accent on the word three in the sequence one, two, three

signals incompleteness, while a falling accent indicates the end of the utterance. The difference in stress and rhythm in the morpheme pair black-bird indicates whether one is speaking of a special type of bird or of a colored representative of the whole class. And the type of juncture distinguishes an aim from a name. Such features are called constructive features. Another type of phonological feature is distinctive or phonemic features, which in themselves contain no semantic information, but are used only for constructing morphemes. Unlike these, constructive features contain information on the formation of semiotic units in their hierarchy. Both types of features are either present or absent: no gradation is possible here. In addition to these features, which determine linguistic affinity, Lotz recognizes a third feature, the speech carrier, which is socially determined: the manner of speech. This feature is constant. Other constant features are the properties of the speaker's speech organs. Finally, there is a variable factor here: the expression of the speaker's emotional situation.

Lotz further emphasizes the binary principle of oppositions, although he properly acknowledges the complications arising from the application of the binary principle when dealing with three elements (e.g., i, e, and ae in English) which all contain a common feature. In this case, according to Lotz, we must either introduce a complex middle term, or admit trinary oppositions. The case of the English vowels i, e, and ae does not contradict the binary theory, since the middle term (e) is a member of both binary oppositions: it is noncompact with respect to (ae) and nondiffuse with respect to (i). It should be added that Trubetzkoy used binary oppositions, although he recognized graduated oppositions as well. From the standpoint of communication theory binary oppositions represent the most economic arrangement of an utterance as a means for transmitting information.

In conclusion, summarizing our views on communication theory, we should emphasize two main points:

1. Where communication theory deals with the transmission of written or printed language, its application to linguistics is quite justified, and it must be admitted that research has made contributions in this direction. I consider the concept of redundancy as the most fruitful concept of information theory from the linguistic standpoint. It is of course necessary to examine redundancy on all levels of language, since not only phonemes, but also morphemes and entire words are redundant. The ideal state in language would be no redundancy. But we know that words and sentences can be in an optimal form and in a reduced form. In the case of words English, for example, has the well-known phenomenon of gradation; in the case of the sentence, this is a stylistic form. Redundancy actually performs the function of controlling information.

2. Where communication theory deals with the transmission of spoken language we must be very cautious in applying it to linguistics, since the human factor here enormously complicates the entire procedure of communication, making very problematical the development of any generally applicable theory at the present stage of investigations. It is no wonder, therefore, that no truly broad-based study has been made which might solve the problem of communicative speech as a whole. This does not mean, of course, that we should not strive for such a theory, since past research has shown clearly that language is basically regular and that the individual units of language demonstrate variations within that regularity. If this were not the case, linguistic communication would be impossible.

- END -