[*Currrent Trends in Linguistics*, vol.1 (The Hague: Mouton, 1963)]

MACHINE TRANSLATION

KENNETH E. HARPER

1. INTRODUCTION

Machine translation (MT) has been an active area of research in the Soviet Union for the past decade. (In other East European countries these studies are only now beginning.) This activity may be divided into two periods (1). The first period was characterized by intensive efforts to construct sets of rules (algorithms) by which electronic computers could effect a translation between given language pairs. Translation algorithms were drawn up in varying degrees of depth and complexity for at least twenty language pairs (Russian-English, English-Russian, Hungarian-Russian, etc.). These programs were based on school-grammars of the languages in question; ambiguities in syntax and meaning-transfer were solved by *ad hoc* rules primarily derived from the examination of small text samples. It was assumed that additional ambiguities in new texts would be solved by additional rules. The chief goal appears to have been the demonstration of the feasibility of MT by example; the main shortcomings devolved from an overestimation of the powers of computers (the fascination with a new toy), equally, from an overestimation of the state of linguistic knowledge.

The second period, beginning roughly in 1959, and still continuing, is characterized by a far greater emphasis on linguistic research as a prerequisite to MT. The inadequacy of the algorithm approach was dearly indicated in 1959 by three leading MT researchers (2). This paper, and a later paper by V. V. Ivanov, set forth the following strategic principles: (i) too much attention has been given to detailization of isolated language facts, and no effort has been made to relate these facts to broader principles -- "One cannot see the forest for the trees." (2); (ii) "The practical tasks of machine translation can only be solved in the future, after the preliminary study of languages is completed." (3); (iii) "The facts of language must be collected by the machine itself so that linguists will be able then to process them for inclusion in a general system." (2); (iv) "The importance of machine translation is now determined by its stimulating part in the development of linguistics." (3).

The necessity of fitting together fact and theory has been an important motivation for the recent interest in structural linguistics. In this connection, we may note the formation in 1960 of a Sector of Structural and Applied Linguistics in the Institute of Linguistics, AN SSSR. In describing the orientation of this Sector, A. A. Reformatskij emphasized the importance of theory to applied linguistics, and set forth the case for structural linguistics as follows: "The structural aspect presupposes the examination of a language as a whole and each level of its structure as an interconnected system of levels of significance given in a hierarchical gradation of symbols and their combinations, organized in contrast with each other, paradigmatically connected, and linearly distributed in speech." (4) It is not at all clear that this "examination" of a language has yet yielded important results; it is clear, however, that the current emphasis is upon organization of language data, with a view towards generalization. The day of the pair-wise translation algorithm, with its limited objectives, is now past. (It may be added that the reasons for the abandonment of these schemes were not all theoretical: the Soviets did not possess, or did not want to spare, the machines, time, and talent necessary to test and develop these programs.) In summary, recent MT research in the Soviet Union is proceeding along two main paths: a more detailed analysis of specific language phenomena and the construction of linguistic theory. On the first point, it is interesting to note the large number of studies devoted to a description of the Russian language. (In earlier years, Russian was taken for granted; the challenge lay in transforming Russian into exotic languages, say, Burmese.) Generally, these studies have been based on the examination of small text samples, and apparently without the aid of data processing equipment. The shortage of equipment may, in fact, be taken as the chief reason for the heavy investment in theoretical work. Here, one may note the introduction of concepts from mathematics and symbolic logic, and the lively interest in interlanguage models (the Intermediary Language). In a word, although the computer has not yet been a partner in linguistic research, it has been the stimulus for critical re-examination of linguistic theory.

2. RESEARCH CENTERS IN MT

Soviet researchers in MT number in the hundreds. The 1958 MT Conference was attended by 340 representatives of 79 institutions. During the past ten years, however, the most important work has been done at four institutions: the Institute of Precise Mechanics and Computer Technique (ITMVT), the Electromodeling Laboratory of the Institute of Scientific Information, Leningrad State University, and the Steklov Institute of Mathematics. Two additional groups have been active in recent years: the First Moscow State Pedagogical Institute of Foreign Languages, and the Institute of Linguistics. These groups have contributed a major portion of MT literature. Research work at other institutions has been on a smaller scale, and often appears to lack continuity (for example, the studies by individuals at the state universities of Gorky, Kiev, Kharkov, Erevan, Tbilisi, and Petrozavodsk, and at a number of scientific-research institutes).

3. PUBLICATIONS IN MT

The sources of publication for MT studies are exceedingly diverse. No regular avenues of publication exist. Because of the relative newness and "newsworthiness" of the subject, many purely promotional and popular articles have been printed, in a wide variety of periodicals (the popular press, semipopular journals, and in scholarly journals ranging from philosophy to computer technology). The recent National Bureau of Standards *Bibliography* (5) cites some 69 different sources for these publications. The more serious papers have also appeared in a number of different source documents, sporadically and in a rather haphazard fashion; papers frequently appear, perhaps in revised form, in more than one periodical. When individual researchers or groups publish their own papers, the quality of editing is often low, and the distribution is limited. Copies are generally available outside the Soviet Union only on an individual exchange basis.

By content, three types of MT publications may be distinguished: *generally informative* (promotional articles, surveys, and articles introducing such concepts as transformational grammar, information theory), *theoretical* (efforts to develop speculatively a given linguistic concept), and *substantive* (results of specific grammatical or lexical studies, routines, programs). All these publications clearly reflect the newness of the subject. In exploratory or projective papers, the line of investigation is rarely pursued to satisfactory lengths. When specific results are reported, the effect is usually unconvincing because of the smallness of the data base. Work completed in a given year may be completely ignored two years later, either because of reassignment of personnel or because of the overall deficiency in planning. Papers written in the first few years of MT research are by now quite dated. (These characteristics, it should be added, are not peculiar to Soviet MT literature.)

To the knowledge of the writer, the only good bibliography of Soviet MT publications is that issued by the National Bureau of Standards (5). Actually, this is a bibliography of translations made by the U.S. Joint Publications Research Service; the translations themselves are inferior, but the coverage of Soviet literature is excellent. A total of 519 items (including abstracts) are contained in the author index of this report. In many instances, the JPRS translation is the only version of the original paper available in this country.

Three main avenues of publication are open to Soviet MT researchers: (i) Scholarly meetings. The abstracts or complete texts of papers presented at large MT conferences are, with one exception, available. To date, four such conferences have been held: the 1958 All-Union Conference on Machine Translation (Moscow), the 1959 All-Union Conference on Mathematical Linguistics (Leningrad), the 1960 Inter-VUZ Conference on Applied Linguistics (Chernovtsy), and the 1961 Conference on Information Processing, Machine Translation, and Automatic Text Reading (Moscow). Translations of papers given at three of these meetings have been made (6, 7, 8); titles and brief summaries of papers presented at the Chernovtsy conference are available (9,10). In addition, there have been several smaller meetings and seminars, for which papers are apparently not printed, *e.g.,* the 1961 inter-VUZ conference on the application of structural and statistical methods in studying the vocabulary of a language, and the 1961 meeting on structural linguistics (transformational method) held by the Sector on Structural Linguistics, Russian Language Institute, AN SSSR (11).

(ii) Collections *(sborniki).* Two of these collections are of a semiserial nature. *Masinnyj perevod i prikladnaja lingvistika* (12), once entitled the "Bulletin" of the Society *(Ob"edinenie)* for Machine Translation, has published substantive papers by workers of a number of different groups since 1959, as has *Problemy kibernetiki* (13), since 1958. In addition, various research groups have irregularly issued collected papers by their staff members. Three such collections have been issued by the Leningrad University group (14,15,16), two by the Institute of Scientific Information (17,18), and two by the Institute of Precise Mechanics and Computer Technique (19, 20). Moscow State University has issued the collection, *Exact Methods in Linguistic Research* (21), and the Institute of Linguistics has published separate studies (22,23). I. S. Muxin is the author of a book surveying MT problems (24).

(iii) Scholarly journals. Papers appearing in these journals tend to be addressed to non-specialists, and rarely deal with problems in a detailed way. *Voprosy lingvistiki* has carried the greatest number of these general articles. Most of the papers referred to below belong to the first two categories mentioned.

4. A SAMPLING OF MT STUDIES

The NBS *Bibliography* (5) contains a twelve-page subject index. For present purposes, it seems appropriate to cite representative papers in the traditional areas of linguistic research. Most of the following are concerned with the grammar of Russian.

The *generally informative* papers include I. A. Mel'čuk's objective and rather complete survey of MT studies in the U.S. and Western Europe (25, 26), and the two surveys and critiques of the Soviet MT effort cited above (2, 3). Several papers have introduced MT researchers to non-linguistic concepts and techniques: information theory (27), statistical methods (28), probability theory (29), information processing techniques as applied, for example, in automatic abstracting (30), etc. The importation of linguistic theory from the West is well known (see, *e.g.,* item 31). The most interesting applications of extra-linguistic ideas to linguistic theory have been suggested in the area of mathematical linguistics, rather than in the area of MT research *per se.*

MT studies in *morphology* have centered on problems of automatic recognition, *i.e.,* decomposition of text forms into constituents that can be used in dictionary lookup and in syntactic analysis (of the input language) and synthesis (of the output language). Except as they illustrate the enormously complex mechanism of language, these studies have no great theoretical interest. A number of schemes for achieving machine recognition of grammatical morphemes have been devised. To the knowledge of the writer, none are founded on a rigorous definition of the morpheme or on a systematic processing of text or word lists. Most MT programs are built on the assumption that the computer dictionary will be composed of stems or roots of words, rather than canonical forms or paradigmatic forms. (The purpose here is conserve storage space in the computer). How shall the machine be programmed so as to detach derivational and inflectional affixes from forms encountered in text, so that text stems can be matched automatically with dictionary stems? The most elaborate algorithm for this purpose is that proposed by Mel'čuk, (32); this method is Russian-oriented, but is intended for application to other languages. The problems of stem-homography arising from automatic segmentation are dealt with in several papers, *e.g.,* for Russian (33, 34), for English (35, 36), and for Swedish (37). Other programs describe the means of utilizing the morphological information obtained from these segmentation routines, both in analysis and in synthesis, but in particular with Russian as the output language, (38). One program (French-Russian) has been tested on a computer, and may be considered operational (39).

Distributional characteristics of Russian inflectional affixes have been studied (although not by automatic procedures), as an aid in morphological coding. Thus the frequency of case forms in nouns has been counted in scientific prose (40, 41). Certain coding schemes have taken advantage of the redundancy in Russian declension patterns; *e.g.,* the dative and prepositional cases are coalesced (22,42).

In *syntax,* as in morphology, MT research has contributed little to theoretical understanding. The chief concern of the algorithm-builders was to solve individual problems as they arose at a given stage in the translation process. The solutions were usually effected in terms of "how to translate" a given construction; there was a minimum of interest in explaining or typifying the construction. An example is the complex set of rules in one scheme for resolving homography in French-Russian MT (39). The rules, as in all Russian algorithms of sentence analysis, are embedded in a kind of flow chart that, for each ambiguous word, asks yes/no questions about the presence or absence of specific words or word-classes in context. (The operational limitations of the flow chart seem not to have been understood by Soviet MT workers.) Absent is any motivation to describe the various syntactic functions of the French homographic words except as they can be fitted into the Russian syntactic and lexical pattern. In effect, this procedure is geared to the solution of isolated problems, rather than to the description of larger syntactic units (the clause or the sentence), in which the isolated problem words or constructions may fit unambiguously.

Recently, the desirability of automatic parsing as a part of the MT process has aroused a certain interest in sentence structure theory. MT researchers have been busy with routines designed to establish in Russian the syntactic connections between pairs of text occurrences. These "governor-dependent" pairs, or "configurations", are of course the building blocks for complete sentence structure description. Two of the most detailed programs are a routine for testing adjective-noun agreement (22), and a routine for testing verb complementation (23). The latter, derived from syntactic information in Daum and Schenk, *Die Russischen Verben,* presents as a first model more than 130 patterns of verb complementation of the type: čto; čto/čemu; čto/čem (čerez čto); čemu/na čto/čem. Criteria of equivalence/non-equivalence and compatibility/incompatibility are employed in the classification. Another system of classification for Russian words (not only verbs), according to their governing capabilities, is given in (43). Other studies have been made of these grammatical configurations in Russian (44), and in English (45). Routines for determining the syntactic governors of prepositional phrases have been written, for Russian (46), and for English (47). Studies have been made on the syntactic role of formulas in Russian mathematical texts (48) and on the function of punctuation marks in Russian (49).

Only in the past year or two have Soviet MT workers come to realize the enormous difficulties of describing syntactic behavior with the required degree of specificity. A native command of the language does not suffice, the best traditional treatments of syntax are notoriously inadequate, and the "brute force" attack on isolated problems through analysis of the microcontext has not proved satisfactory. The alternative source of information is written text, and it is to this source that students are now turning. In this connection, one of the most significant developments in Soviet MT work is a recent paper on the use of machine aids for the collection of syntactic information (50). Here is described a program for automatic parsing of text that will make the computer a full-fledged partner in research: the program is designed to provide the researcher with facts about parsing, or configuration-building, which he originally had not known, or which he had been unable to encode in the grammar. Such a program is indeed a powerful tool, leading to a more complete understanding of the syntactic function and the meaning of word combinations *(slovosočetanija).* The implications of this development to grammar and lexicography are tremendous, so that the achievements of the past decade, as set forth in the present literature, will seem to represent a first stumbling step unaccountably long in the taking.

SUMMARY

Within the past four years, Soviet linguists have perforce been introduced to a striking variety of new concepts. There has been a veritable onslaught of ideas, from the West, from the symbolic logicians, from engineers, and particularly from mathematicians. The impact of mathematics has perhaps been less dismaying to Soviet linguists than to their confreres in the West. In this initial period, research is almost certain to be suggestive, but non-productive. Linguists flourish mathematical weapons that they are ill-equipped to handle; mathematicians and engineers attack language problems with a depressing degree of self-confidence and a surprising lack of finesse. The literature is often difficult to understand, and even more difficult to evaluate. Nonetheless, it is evident that many of the theoretical constructs presented in MT literature are of doubtful validity, and that some are essentially trivial. This is not to underestimate the potential of cross-fertilization. Here, it would appear that Russian linguists have shown greater responsiveness to this challenge than have linguists in the West. The analogy with computer sciences is perhaps instructive: the best Soviet mathematicians are deeply involved in problems of computer design and programming, and it can well be argued that they benefit from this involvement in both "pure" and "applied" science. The older, established Soviet linguists have made no great contribution to applied linguistics, but they exhibit a remarkable willingness to "be shown". They are likely to foresee and to encourage the rapid change in linguistics that will become evident in the next generation.

REFERENCES

With few exceptions, the papers listed below are to be found in the National Bureau of Standards *Bibliography* (5). Since a great majority of these papers are available only in the English translation distributed by the U.S. Joint Publications Research Service, the titles have been given as translated by the JPRS. Russian titles are given for the few items not contained in the NBS *Bibliography.*

1. K. E. Harper, "Soviet Research in Machine Translation", *Proceedings of the National Symposium on Machine Translation.* Ed. H.P. Edmundson (Prentice-Hall, Inc., 1961).
2. N. D. Andreev, V. V. Ivanov, and I. A. Mel'čuk, "Some Remarks and Suggestions Relative to Work on Machine Translation in the USSR", JPRS:8026, pp. 1-14; *Mašinnyj Perevod i Prikladnaja Lingvistika* 4.3-24 (1960).
3. V. V. Ivanov, "Some Problems of Machine Translation in the USSR", JPRS: 13439, 49 pp.; *Doklady na Konferencii po Obrabotke Informacii, Mašinnomu Perevodu i Avtomatičeskomu Čteniju Teksta* 10.1-29 (1961).
4. A. A. Reformatskij, "In Place of a Preface" (see Ref. 22, below).
5. J. L. Walkowicz, "A Bibliography of Foreign Developments in Machine Translation and Information Processing", *National Bureau of Standards Report* 7721, Sept. 1, 1962.
6. *Abstracts of the Conference on Machine Translation (May 15-21, 1958),* Ministry of Higher Education, USSR, First Moscow State Pedagogical Institute of Foreign Languages (Moscow, 1958). (JPRS :DC-241).
7. *Tezisy Soveščanija po matematičeskoj Lingvistike,* Ministry of Higher Education, USSR (Leningrad, 1959). (JPRS :893-D).
8. *Doklady na Konferencii po Obrabotke Informacii, Mašinnomu Perevodu, i Avto-*

maticeskomu Čteniju Teksta, Institute of Scientific Information, AN SSSR (Moscow, 1961).

9. O. S. Širokov, "Conference on Structural and Mathematical Linguistics", JPRS:8132, pp. 1-8; *VJa* 10/1.155-159 (1961).
10. D. M. Segal, "Intervuz Scientific Conference on Applied Linguistics", JPRS: 13761, pp. 118-123; *Mašinnyj Perevod i Prikladnaja Lingvistika* 5.93-99 (1961).
11. S. K. Šaumjan, "Urgent Problems of Structural Linguistics", JPRS:14252, 15 pp.; *IzvAN* 21/2.103-111 (1962).
12. *Mašinnyj Perevod i Prikladnaya Lingvistika,* Association for Machine Translation, First Moscow State Pedagogical Institute of Foreign Languages (Moscow).
13. *Problemy Kibernetiki,* State Publishing House of Physico-Mathematical Literature (Moscow).
14. (See Ref. 7, above).
15. *Materialy po Mašinnomu Perevodu, Sbornik I* (Leningrad, 1958). (JPRS:2150-N).
16. *Voprosy Statistiki Reči,* Edited by L. R. Zinder (Leningrad, 1958), 148 pp. (JPRS:6543).
17. *Soobščenija Laboratorii Elektromodelirovanija,* Institute of Scientific Information of the Academy of Sciences USSR, Moscow, 1.1-250(1960).
18. *Lingvističeskie Issledovanija po Mašinnomu Perevodu,* All-Union Institute of Scientific and Technical Information Publishing House, Issue No. 2 (Moscow, 1961) (JPRS:13173).
19. *Sbornik Statej po Mašinnomu Perevodu,* Institute of Precise Mechanics and Computer Technique AN SSSR (Moscow, 1958). (JPRS:925-D).
20. *Trudy Instituta Točnoj Mexaniki i Vyčislitel'noj Texniki Akademii Nauk SSSR,* No. 2 (Moscow, 1961). (JPRS:13543).
21. O. S. Axmanova, I. A. Mel'čuk, E. V. Padučeva, and R. M. Frumkina, *O Točnyx Metodax Issledovanija Jazyka,* Izdatel'stvo Moskovskogo Universiteta (Moscow, 1961).
22. I. A. Mel'čuk, "Two Operators for Establishing Correspondence (for Automatic Syntactical Analysis)", JPRS: 13444, 71 pp. Preliminary Publications of the Sector of Structural and Applied Linguistics, Institute of Linguistics, AN SSSR, 1961, pp. 1-38.
23. L. N. Iordanskaja, "Two Operators for Processing Word Combinations with 'Strong Government' (for Automatic Syntactic Analysis)", JPRS:12441, 41 pp. Preliminary Publications of the Sector of Structural and Applied Linguistics, Institute of Linguistics, AN SSSR (Moscow, 1961), pp. 3-33.
24. D. Ju. Panov, *Avtomatičeskij perevod,* Izdatel'stvo AN SSSR (Moscow, 1958).
25. I. A. Mel'čuk, "Some Problems of Machine Translation Abroad", JPRS:13135, 75 pp. (see Ref. 8, above).
26. I. A. Mel'čuk, "Mašinnyj perevod i lingvistika" (see Ref. 21, above).
27. E. V. Padučeva, "Vozmožnost' izučenija jazyka metodami teorii informacii"

(see Ref. 21, above).

28. R. M. Frumkina, "Primenenie statističeskix metodov v izučenii jazykov" (see Ref. 21, above).
29. R. M. Frumkina, and V. M. Zolatarev, "Toward a Probability Model of a Sentence" JPRS:893-D, p. 27 (see Ref. 7, above).
30. V. A. Purto, "Automatic Abstracting Based on a Statistical Analysis of the Text", JPRS:13196,15 pp. (see Ref. 8, above).
31. T. M. Nikolaeva, "What is Transform Analysis?' JPRS:3796, pp. 32-41; *VJa* 9/1.111-115(1960).
32. I. A. Mel'čuk, "Morphological Analysis in Machine Translation", JPRS-:13514, pp. 129-302; *Voprosy Kibernetiki* 6.207-276 (1961).
33. L. N. Iordanskaja, "The Morphological Types of Stems in the Russian Language (For Distinction of Homonymy of Morphemes During Analysis in Machine Translation)", JPRS:13514, pp. 313-329; *Voprosy Kibernetiki* 6.281-287 (1961).
34. L. N. Zasorina, N. B. Karačan, S. N. Med'vedeva, and G. S. Cejtin, "A Project of Programs for Morphological Analysis of the Russian Language in Machine Translation", JPRS:2150-N, pp. 99-148 (see Ref. 15, above).
35. M. M. Langleben and E. V. Padučeva, "Elimination of Morphological and Syntactic Homonymy in Analyzing English Texts", JPRS:DC-241, pp. 69-70 (see Ref. 6, above).
36. T. N. Mološnaja, "Problems in Distinguishing Homonyms in Machine Translation from English into Russian", JPRS:646-D, pp. 19-27; *Problemy Kibernetiki* 1.216-221 (1958).
37. S. S. Belokrinickaja and T. N. Mološnaja, "On an Algorithm for Independent Morphological Analysis of the Swedish Language", JPRS:13543, pp. 338-354 (see Ref. 20, above).
38. T. M. Nikolaeva, "Synthesis of Forms of Russian Words During Machine Translation into Russian", JPRS: 12047,19pp.; *Problemy Kibernetiki* 5.263-269 (1961).
39. O. S. Kulagina, "French-to-Russian Machine Translation. French-to-Russian Translation Algorithm", JPRS:6494, pp. 14-86; *Problemy Kibernetiki* 4.207-257 (1960).
40. V. A. Nikonov, "Statistics on Russian Cases", JPRS:3758, pp. 31-51; *Mašinnyj Perevod i Prikladnaja Lingvistika* 3(10).45-65 (1959).
41. Z. M. Volockaja, L N. Šelimova, and A. L. Šumilina, "Some Numerical Data Pertaining to Forms of Nouns and Verbs of the Russian Language", JPRS:13173, pp. 339-347 (see Ref. 18, above).
42. E. V. Padučeva, "Description of the Case System of the Russian Noun (Certain Problems of Homonyms in Machine Translation)", JPRS:6588, pp. 1-13; *VJa* 9/5.104-111 (1960).
43. E. V. Padučeva and A. L. Šumilina, "Syntagmas of the Russian Language", JPRS:13173. pp. 120-150 (see Ref. 18, above).

44. B. M. Lejkina, "Program for the Analysis of Phraseological Complexes", JPRS: 13134, 11 pp. (see Ref. 8, above).
45. T. N. Mološnaja, "Statistical Investigation of Grammatical Configurations in English Mathematical Text", JPRS:8026, pp. 4050; *Mašinnyj Perevod i Prikladnaya Lingvistika 4.64-Sl* (1960).
46. I. N. Šelimova, "Establishment of Syntactic Cues for Prepositional Phrases", JPRS.-DC-241, pp. 80-82 (see Ref. 6, above).
47. M. M. Langleben, "Syntactic Analysis of Prepositional Groups in the English language", JPRS:13173, pp. 314-324 (see Ref. 18, above).
48. M. M. Langleben, "Determination of Syntactic Connections for Formulas in Russian Mathematical Texts", JPRS:DC-241, pp. 68-69 (see Ref. 6, above).
49. T. M. Nikolaeva, "Analysis of Punctuation Marks During Machine Translation from Russian", JPRS:DC-241, pp. 73-75 (see Ref. 6, above).
50. O. S. Kulagjna, "Ob ispol'zovanii mašiny pri sostavlenii algoritmov analiza teksta", *Problemy Kibernetiki* 7.209-223 (1962).