RESEARCH IN MACHINE TRANSLATION

RUSSIAN-->ENGLISH

TEN-YEAR SUMMARY REPORT

1958-1968

MAY 1968

DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
RESEARCH AND STUDIES PROGRAM
PROJECT NO, NONR-2562(00)

PRINCIPAL INVESTIGATOR:
HARRY H. JOSSELSON

I. INTRODUCTION

Research in computer-aided Russian-English machine translation at Wayne State University commenced in 1958. The Wayne State University machine translation team has sought, as its primary objective, to develop an experimental system of computer-aided translation through the combined efforts of linguists and programmers, and through the interaction of man and machine. In pursuing this objective, the research staff has taken an approach based firmly on both theoretical and traditional linguistic considerations. This approach was deemed preferable to certain simpler approaches taken by some other research groups because it was felt that the achievement of high quality machine translation would be more difficult and require more time than had been originally expected.

Like most efforts in machine translation research, ours has been specifically directed toward the development of procedures for analyzing and translating Russian scientific and technical text. Hence, at the outset a corpus of fifteen Russian mathematical articles was selected to provide the raw data for experimentation. The experimental procedures adopted by the Wayne State University group fall systematically into three general areas:

- the compilation of a machine translation dictionary, and grammar coding of each of its entries;
- the creation of computer routines for performing syntactic analysis of the source text;
- the development of an algorithm embodying semantic procedures for English synthesis.

A fourth area of research has evolved from the recognition on the part of serious researchers that an enormous quantity of distributional information about word usage, not presently extant, has to be acquired, described, and stored in the computer. These data have been and are continuing to be, accumulated through this

fourth area of research, viz. straightforward linguistic investigation of Russian grammar and lexicon. The results of these investigations are directly applied to the existing system in appropriate areas of need, thereby refining and expanding the system as desired.

The above four areas of research are reviewed and elaborated below. Documentation relevant to work carried out in each area appears in the list of project publications at the end of this report.

II. THE WAYNE STATE UNIVERSITY MACHINE TRANSLATION DICTIONARY

The Wayne State University machine translation dictionary is based on a source text of fifteen Russian mathematical articles (as listed in the Fifth Annual Report) which pertain, for the most part, to the field of partial differential equations. This corpus of articles has provided exclusively the raw data for our experimentation, and theoretically allows that any word(s) or text looked up would be found in the present dictionary. As new text is selected for translation, the dictionary is naturally updated to incorporate new words.

The dictionary was compiled directly from text, and the English equivalents were taken either from existing professional translations or from a Russian-English glossary prepared by the American Mathematical Society. The composition of each dictionary entry includes, in addition to the Russian heads and English equivalents, an encoded grammar profile describing primarily the morphological characteristics of each Russian entry. These grammatical profiles play an important part during syntactic analysis of each text sentence.

The demand for additional information (of a syntactic nature) in the dictionary has increased in proportion to the levels of sophistication attained in syntactic analysis. This information, in the form of complementation patterns (i.e. the specific dependent structures and their combinatorial potential) for all predicatives in the dictionary, was encoded and implemented as an auxiliary dictionary, related to the base dictionary but serving a specialized purpose.

Further additions to the dictionary, both in terms of new entries as well as expanded grammatical profiles, are contemplated as our experimentation progresses into problems of semantic analysis Such analysis must be performed in order to enable vital translation rules (affecting the quality of the output) to be written. It is even anticipated, for purposes of higher quality output, that some grammar coding will be required for English equivalents.

III. SYNTACTIC ANALYSIS

Research in computer-aided translation at Wayne State University has, since its inception, been focused primarily on the problems associated with Russian syntax and their resolution. In our approach we have concentrated on analysis at the level of the sentence, and the translation of a sentence from one language into another is precluded if that sentence has not been syntactically resolved. Therefore, our major objective has been to write computer routines to perform syntactic analysis in order to discover the structure of given Russian sentences.

To date the syntactic computer routines employed to perform automatic sentence analysis are of three types. The first type comprises the blocking routines (nominal, prepositional, governing modifier, predicative, and gerund) which group immediate constituents of a sentence into phrases consisting of a core word and its dependents. The second type comprises the profiling routine which arranges the sentence constituents into columns according to their expected syntactic function(s) in the sentence. The third type (comprising routines called PARSE AND HYPERPARSE), using the sentence predicative as a pivotal unit, determines the actual syntactic roles of many of the sentence constituents (at present all unnested nominal blocks and certain unnested prepositional phrases) on the basis of the predicative's complementation patterns which are stored in the auxiliary dictionary.

IV. SEMANTIC ANALYSIS

There have been two formal attempts at conducting semantic analysis with respect to our corpus. The first of these dealt with the problem of multiple meaning (i.e. Russian source text items having multiple English equivalents); the results appeared in the Second Annual Report in 1960. This preliminary investigation gave rise to a general discussion of multiple meaning in machine translation which was presented in the form of a paper at the First International Conference on Machine Translation of Languages and Applied Language Analysis in Teddington, England, in 1961.

In the next section, some linguistic investigations carried out in conjunction with this research project are discussed. These investigations had direct semantic implications even though no formal semantic analysis was viewed as an objective. The objectives of these studies were to expand the grammar codes of the dictionary to include syntactical-semantic information in order to increase the potential of our output; and to isolate certain complementary relationships among sentence constituents in an effort to describe their functions and distributions in terms other than those which formerly were limited to morphosyntactic criteria only.

At present, we are undertaking to incorporate into the existing system the last major component -- viz. a general procedure for synthesizing the English output. Many semantic considerations will of necessity be made.

Eventually, the semantic characteristics of the main word categories (substantives, predicatives, and modifiers) will have to be classified, coded, and entered in the dictionary. These data will be crucial in the establishment of complementation patterns for the members of these categories as well as for determining their translation equivalents.

V. LINGUISTIC INVESTIGATIONS

The trend in machine translation appears to have shifted several years ago from the so-called "95 per cent approach" to the "100 per cent approach". This shift seemed to stem from the realization that machine translation is not an end in itself; it is, rather, a first attempt to adapt computer techniques to natural language analysis and manipulation. The "95 per cent approach" at the outset of machine translation research was sufficient because of the quantity of technical questions which had to be resolved. But as goals shifted from high quality machine translation output to automation of language problems in general, purely linguistic investigations of natural language were recognized as not only an essential direction to follow, but one which indeed was mandatory. The purpose of these investigations was the accumulation of information the lack of which precluded effective automatic analysis.

Hence, the adoption of the "100 per cent approach" was deemed desirable. This approach held the promise of great benefit to machine translation research, heretofore rigorously confined to scientific texts. It also made possible the general formulation of rules which could be applied to the languages of both scientific and literary character.

Such specialized investigations were first undertaken at Wayne State University in 1961 and 1962, with a study of the instrumental and dative cases of Russian. Since that time, linguistic studies of Russian (see the list of project publications) have included: the structure of Russian clause nuclei; syntactic-semantic analysis of the Russian verbs of motion; and observations on the complementation of certain Russian predicatives by specifically introduced subordinate clauses. The results of these investigations are incorporated, wherever possible, into our machine translation system.

We hope to make further improvements in our system by conducting similar investigations of vital Russian linguistic areas not presently illuminated in existing grammatical and linguistic literature. We hope, moreover, to be able to extend our studies to ever broader texts.

VI. SYSTEMS PROGRAMMING

The various subroutines of our existing software system (as outlined in the Sixth Annual Report) are designed either (1) to manipulate text, or (2) to operate the dictionary, or (3) to perform syntactic analysis. Until last year, the programs for (1) and (2) were written for the IBM 1401 and 7070 systems. All processing procedures for (3) were written for, and executed on, the IBM 7090 system.

The Wayne State University Computing and Data Processing Center has acquired within this past year an IBM Systems/360, and consequently a complete conversion of our processing system is required. Toward this end, the dictionary has already been re-formated and appropriate arrangements are presently being worked out for adaptation of the other components of our system to the new hardware. It is expected that any necessary debugging and checking out of the newly converted system will be completed within the near future.

Meanwhile we are taking advantage of the changeover to the new computer system by incorporating certain changes into our analysis algorithm based on insights and inferences drawn from interim linguistic investigations. The subjects of these investigations of the structure of Russian have encompassed both semantic and syntactic problems.

VII. Project Reports, Publications, and Presentations

1. Betz, Robert and Hoffman, Walter. "The Use of a Random Access Device for Dictionary Lookup". Research in Machine Translation, Fifth Annual Report. (Wayne State University, Detroit, Michigan, August 31, 1963).

2. Hoffman, Walter and Janiotis, Amelia and Simon, Sidney. "Decision Tables in Syntactic Analysis". Research in Machine Translation, Fifth Annual Report. (Wayne State University, Detroit, Michigan, August 31, 1963).

3. Janiotis, Amelia and Josselson, Harry H. "Multiple Meaning in Machine Translation". Proceedings of the First International Conference on Machine Translation of Languages and Applied Language Analysis. (Teddington, England, 1961).

4. Josselson, Harry H. "Research in Machine Translation". Proceedings of the National Symposium on Machine Translation, H.P. Edmundson (ed.). (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1961).

5. _____ . A Report on MT: Goals and Results. A Lecture at the University of Michigan Summer Program in Linguistics, Ann Arbor, Michigan, July 27, 1961.

6. _____ . Research in Machine Translation. A presentation before the USAF Scientific Advisory Board Ad Hoc Committee on Mechanical Language Translation, October 10-11, 1962.

7. _____ . "Computer-Aided Language Translation". A lecture delivered at the Conference on Quantitative Linguistics in Stockholm, May 11-15, 1963.

8. _____. "Linguistic Basis of Mechanical Translation : Contributions of Standard Linguistic Theory". Presented at the Seminar on Mechanical Translation, held under the auspices of the U.S.-Japan Committee on Scientific Cooperation at Tokyo, Japan, April 20-28, 1964.

9. _____ . "Machine Translation Dictionary: A Linguist's View". Prepared for the conference on Machine Translation at Yerevan State University, USSR, April 18, 1967.

10. _____ . "Machine Translation in Review". Naval Research Review. (Office of Naval Research, Washington, D.C. , March 1968).

11. Matejka, Ladislav. Prerequisites to_ Syntactic Recognition in Russian Clause Nuclei. Machine Translation Research Project Report 7. (Wayne State University, Detroit, Michigan, May 1965).

12. Proceedings of the Conference on Computer-Related Semantic Analysis.Machine Translation Research Project. (Wayne State University, Detroit, Michigan, December 1965).

13. Rakušanova, Jaromira. Complementation Characteristics of the Russian Verbs of Motion. Machine Translation Research Project Report 9. (Wayne State University, Detroit, Michigan, 1967).

14. Research in Machine Translation, Russian to English, First Annual Report. Machine Translation Research Project. (Wayne State University, Detroit, Michigan 1959).

15. Research in Machine Translation, Russian to English, Second Annual Report. Machine Translation Research Project. (Wayne State University, Detroit, Michigan, 1960).

16 . Research in Machine Translation, Russian to English, Third Annual Report. Machine Translation Research Project. Machine Translation Research Project. (Wayne State University, Detroit, Michigan, 1961).

17. Research in Machine Translation, Russian to English, Fourth Annual Report, Vols. I and II. Machine Translation Research Project. (Wayne State University, Detroit, Michigan, 1962).

18. Research in Machine Translation, Russian to English, Fifth Annual Report. Machine Translation Research Project. (Wayne State University, Detroit, Michigan, 1963)

19. Research in Machine Translation, Russian to English, Sixth Annual Report, Machine Translation Research Project. (Wayne State University, Detroit, Michigan, 1964).

20. Steiger, Amelia J. Parsing By Matrix. Machine Translation Research Report 8.(Wayne State University, Detroit, Michigan, 1966).

21. Steiger, Amelia J. and Simon, Sidney. Observations on the Complementation of Some -o Forms by_ CHTO/CHTOBY Clauses. Machine Translation Research Report 10. (Wayne State University, Detroit, Michigan, 1968).

22. Summary of the Proceedings of the Wayne State University Conference of Federally Sponsored Machine Translation Workers. (Wayne State University, Detroit, Michigan, July 1960).

23. Summary of Proceedings of the Russian to English Grammar Coding Conference. (Wayne State University, Detroit, Michigan, April, 1961).

24. Summary of the Proceedings of the Conference of Federally Sponsored Machine Translation Groups on MT-Oriented Syntactic Analysis. (Wayne State University, Detroit, Michigan, June 1962).

Annotations to Project Publications

1. This article, included in the Fifth Annual Report, documents the dictionary lookup procedure adopted by this research project. It was presented in the form of a paper read at the first annual meeting of the Association for Machine Translation and Computational Linguistics in Denver, August 1963.

2. This article, included in the Fifth Annual Report, describes an attempt to adapt a computational procedure involving decision tables to syntactic analysis. It was presented at the first annual meeting of the Association for Machine Translation and Computational Linguistics in Denver, August 1963.

3. This paper represents preliminary efforts at conducting semantic analysis in connection with Machine Translation research at Wayne State University.

4. This paper, presented at the National Symposium on Machine Translation, describes the organizational and procedural developments during the initial stages of Machine Translation research at Wayne State University.

5. This lecture, delivered at the University of Michigan, covered the international status of Machine Translation research as of 1961, and illustrated applications at Wayne State University with special emphasis on dictionary compilation, grammar coding, and syntactic analysis.

6. This presentation before the USAF Scientific Advisory Board Ad Hoc Committee on Mechanical Language Translation included a statement of approaches to and objectives of Machine Translation at Wayne State University as well as a detailed description of the then existing system.

7. This lecture, delivered at Stockholm, described the basic methodology of the Wayne State University Machine Translation Research Project, including a discussion of the existing programming system and certain applications such as dictionary lookup and syntactic analysis.

8. In this monograph the author compares and contrasts the dictionary procedures, grammar coding schemes, and syntactic analysis routines of various Machine Translation research projects with those employed at Wayne State University.

9. This paper is devoted to a discussion of the ideal format and content of a machine translation dictionary, and a description of the Wayne State University Machine Translation Dictionary.

10. This article reviews the historical highlights of Machine Translation and includes statements of the current status of Machine Translation at Wayne State University and internationally.

11. The aim of this monograph was to provide a more formal linguistic basis for existing syntactic recognition routines, to extend the analysis to the clause level, and to express the subject-predicate relationships of sentence blocks whose complements and/or adjuncts could already be recognized as well.

*12. This is a mimeographed report of the proceedings of the Las Vegas meeting. It includes all papers presented at the conference as well as a summary of discussions which followed each presentation.

13. This monograph grew out of an investigation the aim of which was an attempt to establish government properties of Russian verbs of motion, and to classify the verbal complements as "obligatory" or "optional".

14. First Annual Report (September 1959):
- General discussion of linguistic problems associated with Machine Translation
- Linguistic models for analysis
- Computational considerations
- Analysis of multiple meaning in mathematics text
- Preparation of text for processing (creation of the corpus)
- Preliminary flowcharts to syntactic analysis

15. Second Annual Report (August 1960):
- Grammar coding (coding instructions and formats)
- Syntactic analysis (blocking routines: descriptions and flowcharts)
- Programming: systems and flowcharts
- Semantic analysis: preliminary investigation of multiple meaning

16. Third Annual Report (August 1961):
- Grammar coding extended
- Syntactic analysis: Theoretical study of the Russian instrumental case
- Programming: the Interpretive System
- Multiple meaning in machine translation
- Dictionary lookup procedures: description and flowcharts

17. Fourth Annual Report, Vols. I & II (August 1962):
V.I - General programming
- Output from syntactic analysis routines
- Predicative blocking routine
- Linguistic investigations: Study of the Russian dative case

V.II - Revised instructions for Coding Russian words and idioms
- The Wayne State University Machine Translation Dictionary - printout

18. Fifth Annual Report (August 1963):
- Programming development: Random device for dictionary lookup
- Output from predicative blocking routine
- Linguistic investigations: Decision tables and syntactic analysis
- Additions to the Wayne State University Machine Translation Dictionary

19. Sixth Annual Report (August 1964):
- Programming development
- Syntactic routines updated
- profiling routine (preliminary to Parsing the corpus)
- Blocking and profiling of Article V of the corpus - Output

20. This monograph is based on an investigation the aim of which was to systematize the identification of the roles of Russian sentence components; identification is based on the information coded into the sentence fulcrum, i.e. the predicative.

21. This monograph contains detailed observations of the complementation of a set of -o forms by CHTO/CHTOBY clauses. The observations are based on data supplied by a native informant and discussions between the informant and the authors. A list of the -o forms investigated, as well as the complements established by the native, are included in an Appendix.

*22. These documents record the summarized proceedings of the
*23. meetings convoked under the auspices of the Machine
*24. Translation Research Project of Wayne State University. They reveal respectively the dictionary techniques, grammar coding schemes, and syntactic analysis routines employed by federally sponsored Machine Translation Projects.

*The conferences for which these proceedings were prepared were co-sponsored by the Office of Naval Research, the U.S. Air Force, and the National Science Foundation.