# Machine Translation in the Research Group on Speech Statistics: Theory, Results and Outlook

**RAIMOND G. PIOTROVSKY**

A. I. Herzen State
Pedagogical Institute,
Leningrad, USSR

*Describes the strategy of research into machine translation chosen by the Soviet group on speech statistics. Discusses linguistic paradoxes and problems solved by the group in the course of an MT system development. Presently the MT system is applied to word-for-word and unit-for-unit translation of English and Japanese patent documents. Future tasks include the development of syntactical and semantic procedures within the system and the design of a dispatching program module for the purposes of determining hierarchical relationships between other modules and controlling the system's operation. This structural pattern should provide for the system's capability to understand the input text and adopt to user needs.*

Fifteen years separate us from the advent of the first publications by the Soviet research group on speech statistics (SpSt) devoted to the analysis of MT problems. The work on MT in the SpSt group, founded in 1957, was started in 1964. MT is considered not only as a problem worthy of solution in its own right, but also as a sub-problem of the general theoretical problem of artificial intelligence.

There is no other branch of science which has undergone such dramatic upheavals as MT study. It is enough to mention the powerful reverberation of machine translation ideas among experienced linguists of the middle and older generations as well as among the young at the end of the fifties—a wave that turned into deep disenchantment in the mid-sixties [1, 2], As a result, the majority of pioneers in MT have abandoned engineering linguistics *.

What was the motive of this methodological crisis in MT, the highest point of which coincided with the initiating of the SpSt group? The origin of this crisis was the ignoring of the internal paradoxes which characterise the general problem of artificial intelligence and one of its kernel sub-problems—that of MT and engineering linguistics.

Let us consider the principal paradoxes which create the rejecting barrier between natural language and the computer.

1. The main paradox (which is usually called 'the paradox of man and robot') consists in the contradiction existing between the natural language function in traditional man — man communications, on the one hand, and in the new man—computer—man system, on the other hand. These distinctions are determined by the principal differences between the brain of a human being and 'the electronic brain'. The ability of the human's brain for the unlimited purposeful association of information received together with heuristic thought possibilities brings us to the conclusion that in the man—man system natural language functions as an open system, constantly changing along the line of form-building and metaphoric shifts in meaning. The man—robot paradox is connected with the well-known second theorem of Gödel. In accordance with this theorem, the noncontradiction of a given formal system can be shown only by the methods of another, still more powerful system, etc. Striving for the 100% formalisation of language, we must create an extremely powerful formalisation *L* on the basis of our unformalised heuristic knowledge of language and its descriptions. However, this formalisation must contain the expression *F*, which is insoluble in system *L*. Moreover, we are not able to construct a more powerful description of language, in which the expression *F* would be soluble, because the possibilities of our unformalised system of heuristic knowledge have been exhausted.

But not having the possibility of creating more and more powerful formal models of language which asymptotically bring us toward 100% formalisation, we are deprived of the possibility of constructing machine models of language which are in practice close to its 100% formalisation. The incompleteness of machine formalisation is especially evident in machine semiosis. An artificial sign formed in the computer, consisting of a signifier of a natural language and a signified which includes formalised meanings, is always poorer that the natural language sign.

2. The second linguistic paradox of Achilles and the tortoise, reflecting the Saussurian antinomy of synchrony and diachrony, consists of the following: the closed formalised description of language, oriented to the synchronic section which coincides with the beginning of the processing of this formalisation, as a result of the diachronic processes operation in the open system of a natural language, turns out to be somewhat obsolete by the moment of the realisation of this description on the computer. This paradox serves as still another obstacle in the construction of a 100% formalised machine description of language.

* About the notion 'engineering linguistics', which is the most advanced and general form of MT, automatic text processing and computational linguistics, see in [3].

3. The third paradox consists in the antinomy of idiolect (that is, individual knowledge of language) and collective language — the antinomy, formulated by W. von Humboldt [4, 5].

These paradoxes are closely associated with contradictions existing between the classical equivalent sets of computer language and the fuzzy tolerant sets of natural language [6, 7].

The SpSt group understood these methodological difficulties just at the moment of its formation. Therefore, some of the major problems to be solved in designing the MT system were the following:

The first question to be answered was: What sort of linguistics would be suitable for MT? Starting from the paradoxes 'language—idiolect', 'classical sets — fuzzy tolerant sets', 'system of language—norm (idiomaticity of text)', it became clear that MT problems cannot be treated exclusively on the basis of the set theory [8] and generative grammar. On the contrary, we preferred the procedures of speech linguistics when building algorithms for MT analysis and synthesis.

The next question we had to answer was: What kind of technique would best suit the needs of MT? Currently there exist two approaches to the problem of text generation. The first one suggests that this process is a unit-by-unit sequencing (cf. Markovian process) and in its more recent history had mainly led to ineffective, near-linear, purely statistically oriented theories [9].

The second hypothesis (Luria, Chomsky) claims that text is internally organised and planned. Psycholinguistical studies and informational measuring of speech [10—12] indicate a compromise solution: Text generation appears not to be a simple Markovian process, but one in which fairly regular periods of planning and organisation govern the final text output short periods ahead [13]. Hence, the combination of deterministic and stochastic procedures seemed to us more suitable for MT procedure.

The last problem we had to solve was the choice of the kernel technological idea, organising our MT investigation. This idea was realised in the concept of a linguistic automaton. The concept refers to the combination of digital computer hardware and operating programs for textual information processing (software and linguistic support means).

An ideal linguistic automaton must be constructed as a multilevel system capable of realising such kinds of automatic language processing as MT, indexing, storage of information, semantic pattern recognition, man—machine dialogue, proof of logical and linguistic theorems, and statistical and information investigations [14].

A real linguistic automaton capable of overcoming the rejecting barrier of conceptual difficulties described above has not yet been developed today and, I think, will not be constructed tomorrow. But the partial lowering of this barrier is quite a realistic and solvable problem. Its solution will require not only the implementation of new ideas, but the use of nontrivial heuristic programs as well. These programs must minimize the losses of information arising from the confrontation of the open, dynamic, and fuzzy system of natural language with the closed, static system of computer language, based on classical sets.

Proceeding from the theoretical, technological and organisational criteria, outlined in the previous part, the strategy of the SpSt group investigations in the field of MT could be briefly formulated as follows:

1. All computer programs must be designed on the basis of the informational and statistical investigation of different language levels (strata) with the purpose of determining the weight of syntactic, semantic and pragmatic information for each level and its constituent linguistic units. In this way we attempt to resolve the antinomy of classical sets of computer language and fuzzy tolerant sets of natural language.

2. The MT system of SpSt group is being designed as a modular assembly. Its interactive program modules are characterised by the following features:

— every autonomous program module corresponds (in a certain way) to some language level,

— all the modules, including a linguistic data bank, must be extendable without reprogramming the whole system (in this way we attempt to loosen the paradox of Achilles and the tortoise and that of language and idiolect).

3. In accordance with the level hierarchy of the language, our MT system is being developed on the basis of step increments. In brief, the perspective of this development could be defined as follows.

Insofar as vocabulary and phraseology carry the greater part of the semantic information in a text [15, 16], the primary kernel program module of our MT system is the automatic dictionary which is included into a linguistic data bank, where information about the relationships between various linguistic and encyclopedic objects is stored in the form of a semantic network: objects are the nodes of the network and relationships are indicated by labelled verges between the nodes [17].

Being an autonomous module of the MT system, the automatic dictionary is used now for the word-by-word and unit-by-unit * translation of English and Japanese patent texts [18].

The next step in our MT system activities is the parallel development of syntactic and semantic procedures aimed at:

— the elimination of polysemy of lexical and grammatical units of a text based on an analysis of their contextual environment and the thesaurus reading [19];

— the syntactic analyses of a sentence based upon Tesnière's conception [20] and a frame technique [21];

— semantic pattern recognition [22].

There is no doubt that the word-by-word processing and then unit-by-unit translation coupled with grammatical analysis and rearrangement, taking into account context-dependent restrictions, prove inadequate for achieving high-quality translation. The vital feature which the present translating automaton does not possess is the ability of a human translator to understand the text in a given language and to express its contents in another one, simultaneously adapting it to the interests and knowledge of his counterpart. Thus, the last step in developing our MT system consists in designing a program module that is capable of 'understanding' **

* A unit (a machine idiom) is a sequence of text words that must be translated as a group, not one-by-one.

** The machine 'understanding' can be demonstrated by means of a dialogue, which requires the participants to indicate an awareness of the matter under discussion. Therefore, a linguistic automaton is considered to be able to 'understand' if it can converse intelligently, i. e. if it can remember what it is told and respond to questions in such a way that its replies could be considered reasonable by a human interlocutor.

the input text and realising its semantic processing adapted to pragmatics and the perception of a human user.

As a prototype of an 'understanding' and 'adaptive' linguistic automaton we can indicate the automated question-answer program module TAND, intended for the thesaurus-aided annotating of scientific and technical documents [23]. This module designed and implemented at the SpSt group can correctly answer in Russian a wide variety of simple questions about information contained in French articles on oncology and painting technology.

Finally, it is worth mentioning that the conception of step-by-step increments and that of operating our MT system are internally controversial. On the one hand, the system is designed in an upward direction — from syntactic and semantic program modules to pragmatic modules. But the foreign text 'understanding' by a MT system must operate in the opposite direction: the program module of a lower level should make its decisions on the basis of instructions obtained from the module of a higher level. How can one make a linguistic automaton solve this problem? To attain; the needed solution, it is necessary, first, to constantly accommodate the already finished modules of lower levels to newly created modules of higher levels. Secondly, an automatic dispatcher must be created which can determine the hierarchy of the MT system's modules and control the operation of the system in the downward direction — from higher modules to lower ones [24].

## REFERENCES

1. *Language and machines. Computers in translation and linguistics.* Washington: National Academy of Sciences — National Research Council. (Publ. 1416). 1966, p. 29.
2. Bar-Hillel, Y. The present status of automatic translation of languages. — In: *Advances in Computers, Vol. 1.* New York —London: Academic Press, 1960, p. 94.
3. Bektaev, K. B.; Kenesbaev, S. K.; Piotrowski, R. G. Engineering linguistics. — In: *Linguistics 200.* The Hague—Paris, 1977, p. 43—52.
4. Piotrowski, R. The antinomies of linguistics and automatic interpretation of the text. — In: *3rd International Congress of Applied Linguistics.* Copenhagen, 1972.
5. Dreyfus, H. L. *What computer can't do. A critique of artificial reason.* New York, 1972.
6. Zadeh, L. A. *The concept of a linguistic variable and its application to approximate reasoning.* New York, 1973, p.3—l0.
7. Zeeman, E.; Buneman, O. Tolerance spaces and the brain. — In: *Towards a theoretical biology. An IUBS symposium. I. Prolegomena.* Ed. by C. H. Waddington. Chicago : Aldine, 1968.
8. Marcus, S. *Algebraic linguistics: analytical models.* New York — London: Academic Press, 1967.
9. Skinner, B. F. *Verbal behavior.* London: Methuen, 1957.
10. Piotrowski, R. The place of information carrying elements in a word. — In: *Abstracts of the Conference on Mathematical Linguistics, April 15—21, 1959.* Soviet Developments ….., IPRS-893D. Washington, August 31, 1959.
11. Piotrowski, R. Entropy and redundancy in four European languages. *Statistical Methods in Linguistics* (Stockholm), 1969, No. *5.*
12. Georgiev, H. G.; Piotrowski, R. G. A new method of measuring meaning. *Language and Speech,* 1976, **19,** No. 1, 41—45.
13. Piotrowski, R. The linguistics of a text and machine translation. *American Journal of Computational Linguistics,* 1975, **12,** No. 3, 55.
14. Komlev, L. P.; Belyaeva, L. N.; Chajkovskaya, I. I. A stratificational approach to machine translation and modelling of an integral information base. *The International Seminar on Machine Translation, Moscow, 26—30 November, 1979. Short presentations of reports.* Moscow, 1979, 24—25 (in Russian).
15. Boguslawskaja, H.; Korzeniec, T.; Piotrowski, R. Language and information. *Biuletyn Fonograficzny,* 1972, XIII, 3—32.
16. Piotrowski, R. Meaning information and its measures. In: *Soviet Studies in Language and Language Behavior.* The Hague, 1976, 137—142.
17. Piotrowski, R. G. *Engineering linguistics and language theory.* Leningrad: Nauka Publishers, 1979, 67—74 (in Russian).
18. Bektaev, K. B.; Sadchikova, P. V. A pilot industrial system of lexical machine translation. *The International Seminar on Machine Translation, Moscow, 26—30 November 1979. Short presentations of reports.* Moscow, 1979, 18—19 (in Russian).
19. Piotrowski, R. G.; Anikina, N. V.; Apollonskaya, T. A.; Bilan, V. N.; Borkun, M. N.; Lesokhin, M. M.; Shingareva, E. A. Formal recognition of text meaning. — In: *Speech statistics and automatic text analysis.* Leningrad: Nauka Publishers, 1980. p. 10—61 (in Russian).
20. Ivanova, N. M. An algorithm of syntactical analysis of sentences for machine translation. *School-and-seminar in applied and engineering linguistics, Makhachkala, 3—14 July 1978. Short presentations of reports.* Makhachkala, 1978, p. 26 (in Russian).
21. Shingareva, E. A, Experiences in frames application to semantic pattern recognition of texts. *Scientific and Technical Information, series 2,* 1978, No. 10, 20—128 (in Russian).
22. Piotrowski, R.; Palibina, I. Automatic pattern recognition applied to semantic problems. — In: *Computational and mathematical linguistics. Proceedings of the International Conference on Computational Linguistics, Pisa, 1973.* Firenze: Olschki, 1977, 19-20.
23. Arzikulov, K. A.; Piotrowski, R. G.; Popescu, A. N.; Khazhinskaya, M. S. An automatic system for thesaurus-aided annotating of scientific and technical documents. *Scientific and Technical Information, series 2,* 1978, No. 12, 12—20 (in Russian).
24. Arzikulov, Kh. A.; Leonova, E. M.; Lesokhin, M. M.; Piotrowski, R. G.; Popescu, A. N., Khazhinskaya, M. S. An automatic system for thesaurus-aided annotating of scientific and technical documents. *Problems of Cybernetics, No. 41. Speech statistics and automatic text analysis.* Moscow — Leningrad, 1978, 20—30 (in Russian).