Text Processing in the Leningrad Research Group 'Speech Statistics' — Theory, Results, Outlook

R. PIOTROWSKI
Herzen Pedagogical Institute of Leningrad, USSR

Abstract

The intention of this article is to discuss some semiotic informational aspects of language along with their interpretation in terms of computational linguistics. The paper describes and indicates the resolution of some semiotic and linguistic paradoxes, which create, at present, a rejecting barrier between natural language and the computer.

Twenty years separate us from the advent of the first publications of the Soviet research group on Speech Statistics (Sp St) devoted to the analysis of MT problems (that is, the problem of text translation with the use of a computer). The work on MT in the Sp St group, initiated 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] As a result, the majority of pioneers in MT have abandoned engineering linguistics (EL).[2] The origin of this crisis was the ignoring of the paradoxes which characterise the general problem of artificial reason and one of its kernel sub-problem — that of MT and EL.

Let us consider the principal paradoxes which create the rejecting barrier between the 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 the traditional man-man communicative system, on the one hand, and that of the new man-computer-man system, on the other hand. These distinctions are determined by principal differences which contrast the brain of a human being and 'the electronic brain'. The ability of the human's brain for the unlimited purposeful association of the obtained information together with the heuristic thought possibilities brings us to the point that in the man-man system the natural language functions as an open system, constantly changing along the line of 'form-building' and metaphoric shift of meaning. The man-robot paradox is connected with the well-known second theorem of Gödel. In accordance with this theorem, the non-contradiction of a given formal system can be shown only by the methods of another, still more powerful system, etc. Striving for the 100% formalization of language, we must create an extremely powerful formalization (L) on the basis of our unformalized heuristic knowledge of language and its descriptions. However this formalization 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 unformalized system of heuristic knowledge have been exhausted.

   But not having the possibilities of creating more and more powerful formal models of language which asymptotically bring us toward 100% formalization, we are deprived of the possibility of constructing machine models of language which are in practice close to its 100% formalization.

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

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.[3]

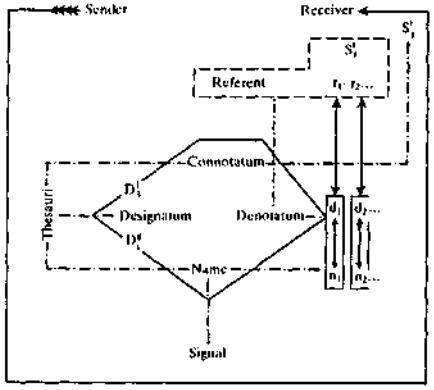
The first aim of this paper is to discuss the semiotical aspects and sources of these paradoxes. As a general framework for the discussion I give here the extended schemes of linguistic sign and a generalization of the classic communication process pattern.

First of all, let us to determine the semiotical concepts of the linguistic sign and its environment presented in Figure 1.

The environment of sign includes:

1. referent r,
2. signal (the information vehicle) i.e. physical state or physical process which serves to mark the object r,
3. paradigmatical system of language (human 'data-base') which contains stylistic (Σ), conceptual

Correspondence: 194021 Leningrad Prospekt M. Toreza 9 kv. 6, R. Piotrowski, USSR.



Environment (situation) — Valencies and senses
Sign borders
Border-line between semantics and form (cf. Figure 2)

Fig. 1

(lexical — 0^1; grammatical — 0^{II}) thesauri ..., a set of means of formal expression (Γ, as well as linguistic competence of sender (Send) and receiver (Rec),

4. a set of communicative situations (S^1_1, S^1_2, ... etc),
5. some texts, which contain meaning elements (d_1, d_2, ... etc) and formal entities (n_1, n_2, ... etc),
6. pragmatical intentions of sender and receiver.

In the strict sense of the word, a sign is a physical entity which includes:

- a denotatum Dn, i.e. a mental image of referent in the mind of human being,
- a designatum with its lexical (D^1_s) and grammatical (D^{II}_s) aspects, or that part (an intersection) in the structural pattern of social practice which corresponds to the denotatum (mental image of referent),
- a connotatum, which summarizes all supplementary semantical shades and emotional and evaluative associations contained in the meaning of the sign,
- a name N, i.e. an internal mental image of the sign.

The signal is formed as a result of semiosis described by Ch. Morris as a 'mediated-taking-account of' ('significance').[4] In other words semiosis is a sign-formation process which can be determined as a five-term relation R_s (d→N, r→Dn→Ds→Cn, Send, Rec). But in speech communication process the real sign structure becomes more complicated due to the metaphorisation or so-called resignificance (connotative or secondary semiosis).[5] To see this, compare the current metaphorical use of Russian летучая мышь, Italian ***pipistrello*** 'bat' in the sense of 'kimono sleeve'. Here the primary sign летучая мышь, ***pipistrello*** becomes a signal (respectively, a name of sign) of a referent (respectively denotatum) 'kimono sleeve' (Figure 2).

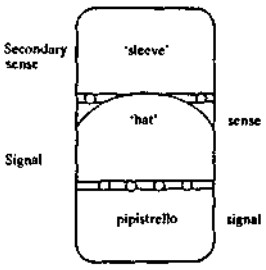


Fig. 2

Now let us turn our attention to the semiotical mechanism of message generation and its decoding in the course of natural language communication. Our communication scheme presented in Figure 3 was originally designed for discussing the whole field of phenomena involved when the information is communicated between human beings. But this scheme is also relevant to the discussion of all other information processes, including the man-computer interaction.

The most plausible line of explanation of message generation and encoding is the following dynamic model:

1. the sender's consciousness receives communicative stimulus from his environment;
2. the stimulus excites the sender's thesauri, his mechanism of goal-setting, planning, choice of priority strategies etc which create a denotative image and designative plan of message;
3. on the base of thesauri and linguistic competence some mysterious mechanism of actualisation realizes the verbalisation and linearisation of image and plan, thus forming the message itself.

In Figure 4 a scheme of the hypothetical semiotic

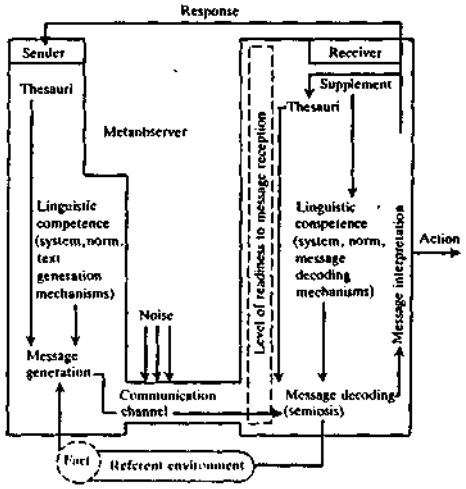


Fig. 3

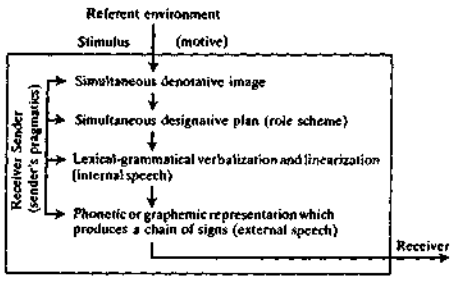


Fig. 4

model is given. This scheme shows how a sender (speaker or writer) passes through a number levels in message formation from his environment to its expression in external speech (or writing).

Although the message is generated by means of a sender's individual semiosis as a bilateral (meaning-form) entity, it enters the channel as a succession of signals to be perceived and decoded by a receiver. The receiver performs decoding by means of certain, no less mysterious, mechanisms using thesauri, linguistic competence, estimation of nonlinguistic environment, subconscious goal-setting, priority strategies etc. This decoding is none other than a new semiosis realized by the receiver himself.

Thus, the natural language communication process implies two semiotic stages: the first stage takes place while generating a message, and the other one while receiving and decoding a message. The results of these two stages are not the same, especially when the message sender and receiver use different thesauri or when the message generator's presupposition while perceiving the environment is not the same as the recipient's one. Environment perceiving occurs primarily over other channels of man's communication with the external world, rather, than through language. One should also bear in mind that widely accepted and normalized lexical and grammatical language resources are not able to express and convey rhematic novelty of all situations.

So it is clear why practically any meaningful sentence implies significance, i.e. metaphoric sense-shift of lexical and grammatical units previously included in language system.

A good illustration of connotative semiosis may be shown by such simple and trite in its primary sens cliche sentence as *Cats adore fish* in T. Winograd's description of computer grammar.[6] This sentence narrates here neither of cats or fish, nor that someone adores something. The word *cats* expresses here an indefinite subject, the noun *fish* is a metaphorical expression for an indefinite object and the token adore symbolizes a verb-copula. Thanks to this resignificance as well as to linguistic-cybernetic presupposition of text environment and to common subconscious goal-setting of author and reader, the above-mentioned sentence expresses a distinct rhematic novelty.

Now we turn to man-computer linguistic interaction. While analysing our semiotic communications map (Figure 3) in the light of this interaction three following 'hot points' are revealed:

1. estimating the non-linguistic environment with the computer,
2. account of the sender's subconscious goal-setting and priority strategies by computer,
3. recognising the connotative semiosis and understanding the meaning of resignified signs by computer.

These problems, now practically irresolvable, form the nucleus of man and robot antinomy — the main paradox of computational linguistics. Before we discuss the above-mentioned questions of man-robot antinomy, we must first determine what kind of information can be transmitted and processed in man-computer communication.

As is well known, there are five types of information:

1. potential information (preformation) which measures statistical constraints as found in the relative frequencies of signal occurrence in message,
2. syntactic information, which considers qualitative relations between information vehicles in natural-language text (its scope is determined by the totality of syntactic constraints of a natural language),
3. sigmatic information, which studies and measures relations between sign denotata and their referents, ignoring both the sender and the receiver of the message,
4. semantic information, which evaluates relations between designata and referents, ignoring both the sender and the recipient of the message,
5. pragmatic information, which may be briefly defined as that reducing the uncertainty in goal-directed behaviour of the message receiver and sometimes also of its sender.[7]

At this time there are well-known formal procedures for measuring potential and syntactic information, as well as information on meaning (the latter being a generalisation of sigmatic, semantic and pragmatic information). The measurement of potential information by computer is quite realistic with a sufficiently representative text sample provided. Syntactic and meaning information, the possibilities of the computer are much more limited. Presently the problem is solved only by means of some indirect man-computer methods. The experiment based on a native speaker's guessing the letters of an unknown text (with subsequent computer processing of the results) is one of these methods.[8] Informational and semiotical analysis of these procedures and their comparison with the possibilities of linguistic automaton (LA)[9] show that the computer, being the recipient of information, can receive and process not only the pre-formation (potential information) and the syntactic information, which are determined by statistic and combinatorial proprieties of signals (names). Having entered a sufficiently complete linguistic database into computer[10] we may simulate the

reception and processing of meaning (sigmatic, semantic and pragmatic) information by LA.

The sigmatic information can be received and processed by an LA on condition that its database is constructed after the denotatum principle, i.e. this base includes the denotata analogs of this lexical and grammatical signs. At the same time the semantic information becomes accessible to LA on condition that the linguistic data are arranged in the LA database as a semantic net, which represents a system of semantic relations (cf. Saussure's 'valeurs') of signs and its meaning combinations.[11]

Theoretically, the goal-setting and priority strategies of some group of receivers can be also simulated by an LA which in this way obtains possibility to perceive and process pragmatic information.[12] But how can one explicate the mysterious mechanism for resignifying the linguistic sign meaning by the sender and for decoding this secondary semiosis by the receiver who employs his own previous experience and information extracted from nonlinguistic and referential environment? These problems comprise the central philosophical question of MT theory and that of artificial intelligence.

The second purpose of this paper is to answer two questions closely associated with informational and semiotical paradoxes described above.

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 and tolerant sets' (see above), 'system of language — norm (idiomaticity of text)' it became clear that the MT problems cannot be treated exclusively by the set-procedure and generative grammar. On the contrary, we had to prefer the text linguistics procedures for the MT analysis and synthesis.

The next question we had to answer was: what kind of technology would correspond to MT? There exist two current approaches to the problem of speech production. The first one suggests that this process is a unit-by-unit sequencing (cf. Markovian process) and in its more recent history had led mainly to the ineffectual, near-linear, pure-statistically oriented theories (Skinner). The second hypothesis (Luria, Chomsky) claims that text is internally organized and planned.

The psycholinguistical studies and informational measuring of speech[13] indicate to us a realistic compromise solution: the text production appears to be not a simple Markovian process, but the one in which fairly regular periods of planning and organization govern the final text output for short periods ahead. Hence the combination of deterministic and stochastic procedure.[14]

The last problem we had to solve was the question of the kernel technological idea, organizing the MT investigation. This idea was realized in the concept of a linguistic automaton (see above).

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

Based on the theoretical, technological and organisation criteria, outlined in the previous part, the strategy of the Sp St group investigations in the field of MT could be briefly formulated as follows:

1. All the computer programs must be designed on the basis of the informational and statistical investigation of different language levels (strata), — cf. Figure 4, with the purpose of determining the weight of syntactic, semantic and pragmatic information concerning each level and its linguistic units. In this way we make an attempt to resolve the antinomy 'classical set of computer language — fuzzy and tolerant set of natural language' (see above).

2. The MT system of Sp St group is being designated as a modular assembly. One should bear in mind that its interacted program-modules (PM) are characterized by the following features:
   - every autonomous PM corresponds (in some defined way) to a certain language level,
   - all the PM, including a linguistic data base 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.

So far as the vocabulary and phraseology carry the greater part of the semantic information in text,[15] the primary kernel program model of our MT system is the automatic dictionary (AD) which is included into a linguistic database, where the information about the relationships between various linguistic and encylopedic objects is stored in the form of a semantic network: the objects are the nodes of the network and the relationships are indicated by direct labelled links between the nodes. Being an autonomous module of the MT system the AD is used now for the word-by-word and unit-by-unit translation of English and Japanese patent texts.[16]

The next step of our MT system activity is progressing in the multidirectional course of further syntactic and semantic development:

- toward the elimination of polysemanticity of lexical and grammatical units of a text based on an analysis of their contextual environment and the thesaural reading,[17]
- toward the performing the syntactic analyses of a sentence based upon Tesnière's conception and a frame technique,
- toward the semantic pattern recognition.[18]

There is no doubt that the word-by-word processing and then unit-by-unit translation coupled with gramma-

tical analysis, rearrangement, and context-dependent restrictions prove inadequate for achieving high-quality translation. The vital feature which the present translating LA does not possess is the ability of a human translater to understand the text in a given language and to express its content in another one, simultaneously adapting it to the interest and knowledge of his interlocuter. Thus the last step of developing our MT system consists in designing such a program-module that is capable of 'understanding' the input text and realizing its semantic processing adapted to pragmatics and world description of a human user. As a prototype of an 'understanding' and 'adaptive' LA we can indicate the computer question-answering program-module TAND. This PM designed and programmed at the Sp St group can correctly answer in Russian a wide variety of simple questions about information contained in French articles on agriculture, oncology and technology of painting.[19]

Finally, it is worth saying that the conception of step increments and that of operating our MT system is controversial. On the one hand, its designing is developed from down to up — from syntactic and semantic program-models to pragmatic PM. 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 PM of the higher precedence. How can one make a linguistical automaton to solve this problem? That is a question.

Notes

1. *Cf. Language and Machines.* National Academic of Sciences. Washington, 1966, p. 29.
2. About the notion 'engineering linguistics', which is the advanced and general form of MT, text automatic processing and computational linguistics, see in: K. B. Bektaev, S. K. Kenesbaev and R. G. Piotrowski. 'Engineering Linguistics'. *Linguistics 200.* The Hague — Paris: Mouton, 1977, pp. 43–52.
3. R. Piotrowski. Text-Computer-Mensch. *Quantitative Linguistics.* Vol. 22. Bochum: Studienverlag Dr N. Brockmeyer, 1984, S.42–43, 47–53.
4. Ch. W. Morris. 'Signe and Act'. — In: Ch. Morris, *Writings on the General Theory of Signs. Approaches to Semiotics,* ed. by Th. A. Sebeok, 16. The Hague-Paris: Mouton, 1971, pp. 401–414.
5. L. Hjelmslev, 'Prolegomena to a Theory of Language'. Supplement to *International Journal of American Linguistics,* vol. 19, Indiana University Publications in Antropology and Linguistics. Indianapolis: Indiana University Press, 1953, §22; G. P. Melnikov, *Systemology and Linguistic Aspects of Cybernetics.* Moscow: Soviet Radio, 1978, pp. 226–290 (in Russian).
6. T. Winograd, *Understanding Natural Language.* Cambridge, Mass.: Massachusetts Institute of Technology. 1972, §3.3, example N 97.
7. D. Nauta, 'The meaning of Information'. In: *Approaches to Semiotics* ed. by Th. A. Sebeok, 20. The Hague-Paris: Mouton, 1972, pp. 39–41, 203–205, 214–225; H. G. Georgiev and R. Piotrowski, 'A new method of measuring meaning', *Language and Speech.* Vol. 19, p. 1, 1976, pp. 41–45.
8. N. Petrova, R. Piotrowski, R. Giraud. 'Caractéristiques informationelles du mot français'. *Bulletin de la Société de linguistique de Paris,* t.LXV, f.1, 1970, pp. 14–28; R. Piotrowski. 'Meaning Information and the Measures'. *Soviet Studies in Language Behaviour.* Amsterdam: North-Holland Publishing Company, 1976, pp. 137–138.
9. The concept of a linguistic automaton (LA) refers to the combination of digital computer hardware and the operating computer program for text information processing (software and lingware), cf. R. Piotrowski. Text-Computer-Mensch ... op cit., pp. 25–34.
10. L. Beliaeva, R. Piotrowski, 'The Stratificational Approach to Modelling of Unitary Linguistics Data Base'. *Computers in Literary and Linguistic Research. Eleventh International Conference.* 3–6 April 1984. Université Catholique de Louvain (Louvain-la-Neuve). Louvain, 1984, pp. 26–27.
11. L. Vaina, 'Towards a Computational Theory of Semantic Memory'. *Cognitive Constraints on Communication. Processes and Representations.* (L. Vaina, J. Hintikka ed). Synthese Language Library 18. Dodrecht: D. Reidel Publishing Company, 1983.
12. J. C. Brown, R. R. Burton, J. de Kleer, 'Pedagogical, natural language and techniques en SOPHIE I, II and III'. *Intelligent tutoring systems.* (Ed. by D. Sleeman and J. S. Brown). London, New York, etc.: Academic Press, 1982.
13. N. Petrova, R. Piotrowski, R. Giraud. *Caractéristiques informationelles* ... op. cit., pp. 14–28.
14. R. Piotrowski. 'The Linguistics of a Text and Machine Translation'. *American Journal of Computational Linguistics,* 1975, no. 3, p. 55.
15. H. Bogusławskaja, T. Korzeniec, R. Piotrowski, 'Language and Information'. *Biuletyn Fonograficzny,* XIII, Poznań, 1972, S.3–32.
16. K. B. Bektaev et al., 'Experiences in machine translation of English and Japanese scientific texts'. *Scientific and Technical Information,* series 2, 1981, no. 5, pp. 26–31 (in Russian).
17. Cf. F. E. Knowles, 'Recent Soviet work on computer techniques for representing natural language meaning'. *Informatics 5.* 26–28 March 1979. The Queen's College Oxford, London: Aslib, 1979, pp. 70–73.
18. The definition of this notion is given in: R. Piotrowski, Text-Computer-Mensch ... op. cit., S.255–278.
19. For a detailed description of the TAND program, see: R. G. Piotrowski, L. N. Beliaeva, A. N. Popesku, E. A. Shingariova. 'Bilingual indexing and abstracting'. *Informatika.* Tom 7. Moscow: VINITI, 1983, pp. 165–245 (in Russian).