The Semiotical Interpretation of Machine Translation*

R. G. PIOTROVSKY

Leningrad A. I. Herzen State Pedagogical Institute, Mojka, 48, 191186, Leningrad, USSR

The author discusses some semiotical-informational aspects of language along with their interpretation in terms of computational linguistics. Linguistic paradoxes which create a rejecting barrier between natural language and the computer are described. It is not possible to develop a true linguistic automaton capable of overcoming this barrier today (cf. fifth-generation computer projects), and it is dubious that it can be done in the near future. But the partial lowering of this barrier is quite realistic and can be done by a semiotical-informational approach to machine translation.

The question of whether one can speak meaningfully of the computer's ability to truly understand and translate natural-language texts is highly controversial. This question must be considered not only as a problem worthy of solution by itself, but also as a part of the general theoretical problem of artificial intelligence.

The paradoxes inherent in text comprehension which create the rejection barrier between natural language and the computer are well known [1, p. 135; 2; 3 p. 49—50; 4], so there is no need to describe them again in detail here.

The main purpose of this paper is to discuss the semiotical aspects and sources of these paradoxes. As a general framework for the discussion, we give here extended schemes of the linguistic sign and a generalisation of the classical communication process pattern.

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

The environment of the sign includes:

(1) referent *r,* i. e. an object of external reality,

(2) signal α (the information vehicle), i. e. a 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 (lexical—θ^l, grammatical — θ^g) thesauri, a set of formal means of expression (Γ), as well as linguistic competence of message sender *(Send)* and receiver *(Rec),*

(4) a set of communication situations $(S_i^t, S_i^t, ..)$,

(5) sign chains in texts, which contain meaning elements $(d_1, d_2..)$ and formal entities $(n_1, n_2,..)$,

(6) pragmatical intentions of sender and receiver.

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

— a denotatum *Dn,* i. e. a mental image of the referent in the mind of a native speaker,

— a designatum with its lexical (Ds^l) and grammatical (Ds^g) aspects, or that part (an intersection) in the structural pattern of social practice which corresponds to the denotatum (mental image) *Dn,*

— a connotatum *(Cn)* which generalises all supplementary semantical shades and emotional and evaluative associations contained in the meaning of the sign,

— a name (form) *N*, i. e. an internal mental image of the signal [5, p. 22—42; 6, p. 20—21].

The sign is formed as a result of semiosis described by Ch. Morris as a 'mediated-taking-account of' (significance) [7]. In other words semiosis is a sign-formation process which can be determined as a five-term relation Rs (α→N, $r \to Dn \to Ds \to Cn$, Send, Rec). But in real speech communication, standard sign structure becomes more complicated due to metaphorisation or so-called re-significance (cf. connotative or secondary semiosis) [8; 9 p. 226—290].

To see this, compare the current metaphorical use of Russian кенгуру *(kangaroo)* in the sense of 'sweater with pockets'. Here the primary sign кенгуру becomes a signal (respectively, a name) of a referent (respectively, denotatum) 'sweater with pockets' (Figure 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 intended for discussing the whole set of phenomena involved in the communication of information between human beings [6, p. 35; 10, p. 39—53; 11]. But the scheme may be also used to discuss other information processes, including man-computer interaction.

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

(1) the sender's consciousness receives a communicative stimulus from the 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;

* Paper presented at the 3rd International Seminar on Machine Transition (Moscow, 5-10 December 1983).

* As can be inferred from the example, the Russian word *kangaroo* has taken on a secondary meaning, viz., 'sweater with pockets'. — *Trans.*

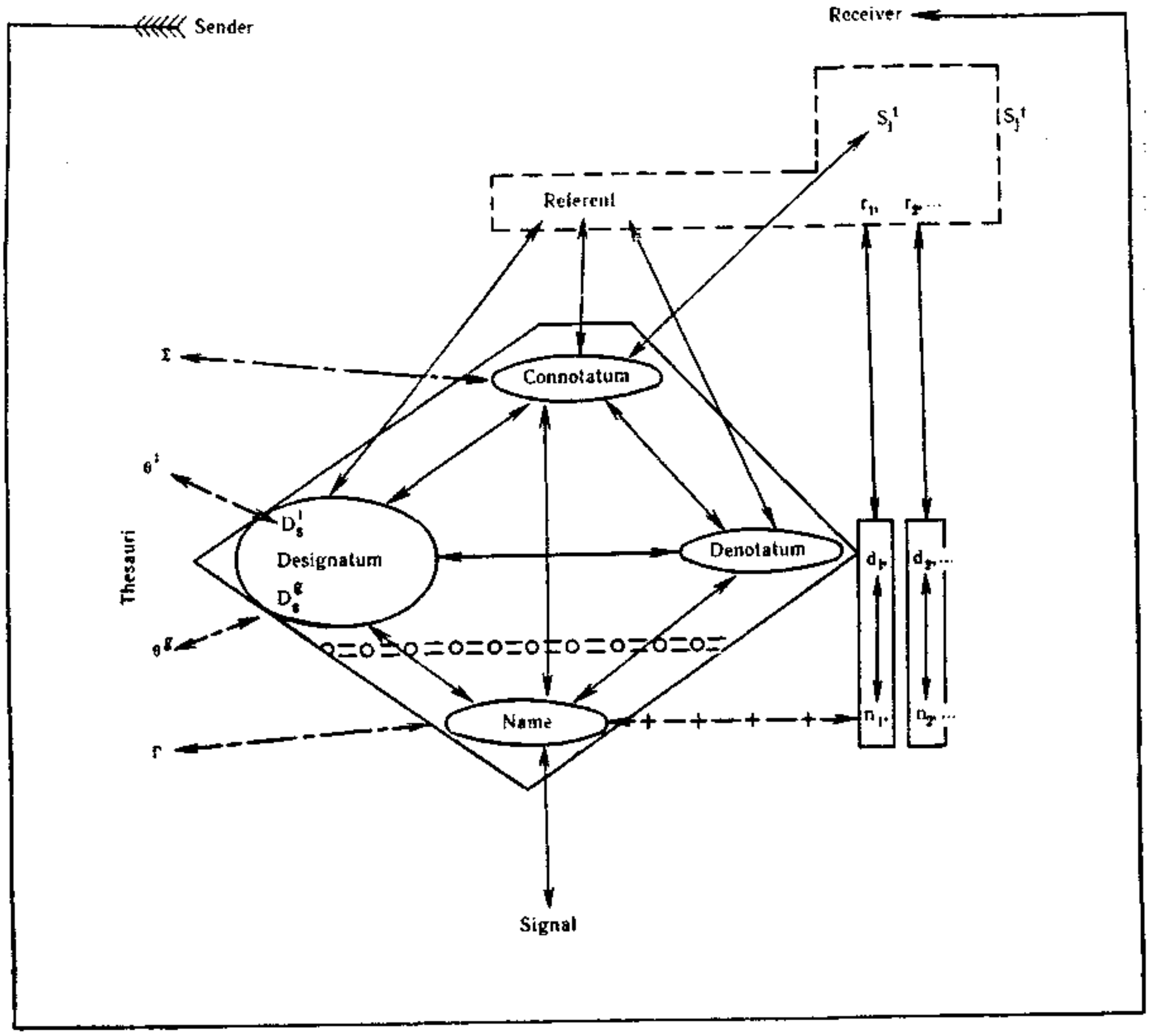
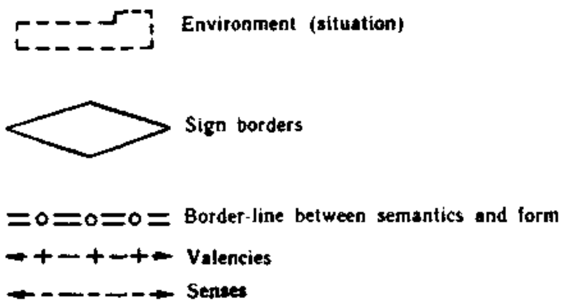


Fig. 1. Linguistic sign and its environment



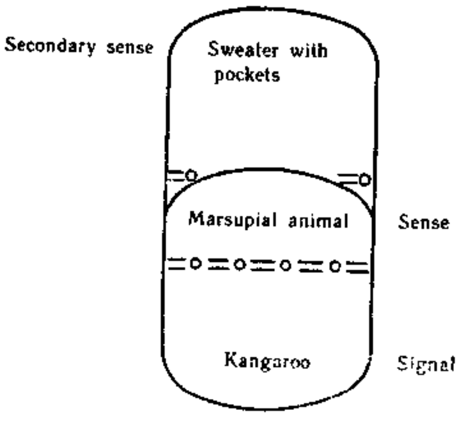


Fig. 2. Connotative semiosis

(3) on the base of thesauri and linguistic competence, some mysterious mechanism of actualisation realises the verbalisation and linearisation of the image and the plan, thus forming the message itself.

In Figure 4, a scheme of the hypothetical semiotic model is presented. This scheme shows how a sender (speaker of writer) passes through a number of levels in message formation — from the awareness of extralinguistic 'cognitive experience' received from his environment to its expression in external speech (or writing) [5, p. 11-13; 6, p. 36; 12, p. 9—10],

Although the message is generated by the sender's individual semiosis as a bilateral (meaning-form) entity, it enters the channel as a sequence of signals to be perceived and decoded by the receiver.

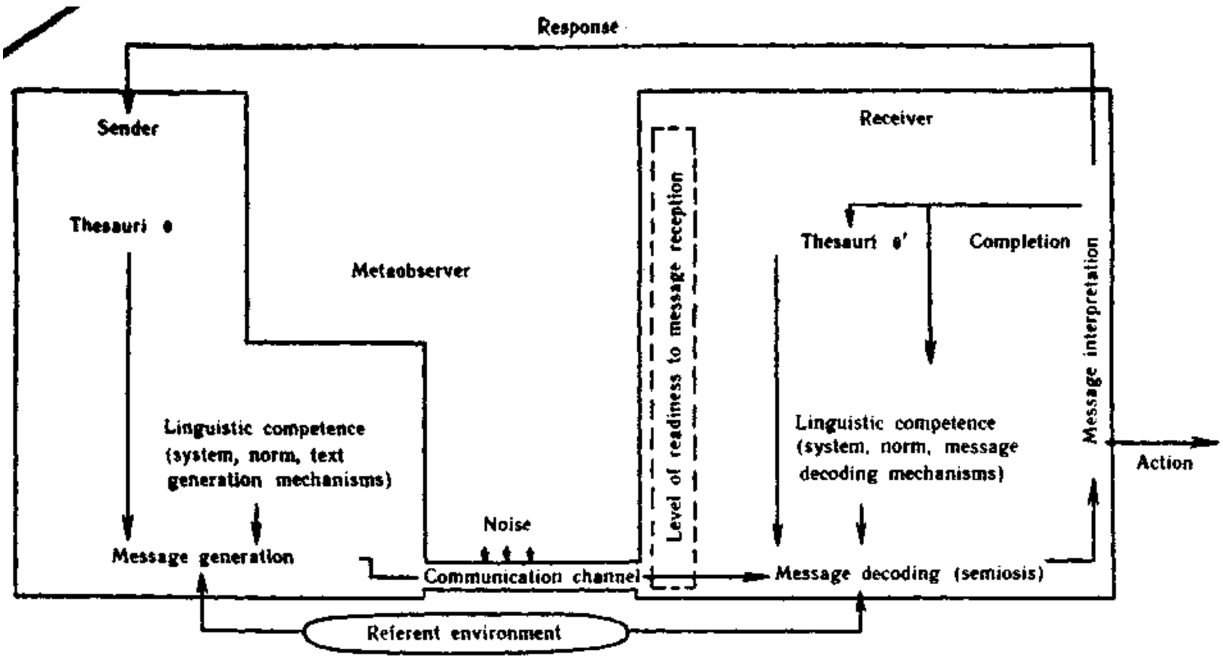


Fig. 3. Human communication scheme

The receiver performs decoding by means of certain, no less mysterious, mechanisms using thesauri, linguistic competence, estimates of linguistic environment, subconscious goal-setting, priority strategies, etc. This decoding is none other than a new semiosis realised by the message receiver himself.

Thus, the natural-language communication process implies two semiosis 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 message sender and receiver use different thesauri, or when message generator's presupposition while perceiving the environment is not the same as the recipient's one. Environment perception 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 normalised lexical and grammatical language resources are not able to express and convey the rhematic novelty of all situations. So it is clear why practically any meaningful sentence implies resignificance, i. e. the metaphoric sense-shift of lexical and grammatical units conventionally included in the language system.

A good illustration of connotative semiosis is provided by the simple and trivial cliche sentence

Cats adore fish

in T. Winograd's description of computer grammar [13]. This sentence narrates neither of cats nor 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* symbolises 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 communication in natural language. While analysing our semiotic communi-

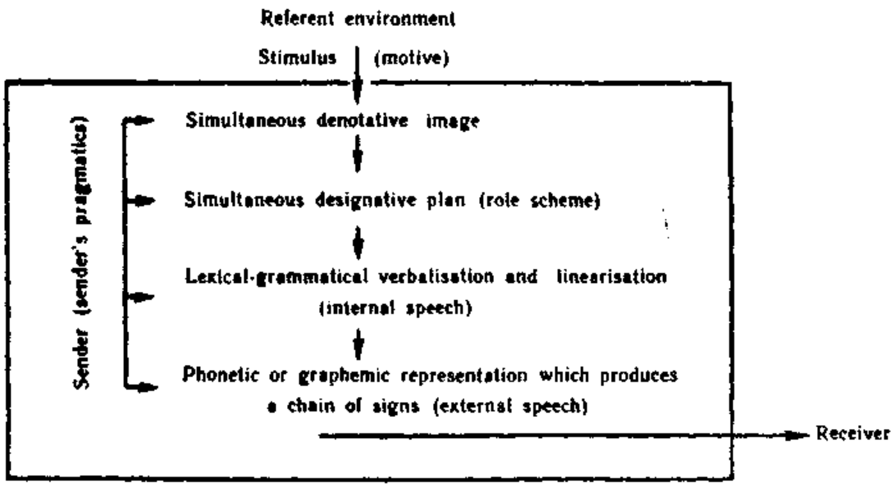


Fig. 4. Semiotical levels of message generation

cation map (Figure 3) in the light of this communication, 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 the computer,

(3) recognising the connotative semiosis and understanding the meaning of resignified signs by the computer.

These problems, now practically irresolvable, form, the nucleus of man-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 the man-computer communication system.

As it is well known, there are five kinds of information:

(1) potential information (pre-information), 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 texts (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 [14, p. 39—41,203—205, 214—225; 15, p. 41—45].

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 [16, p. 4—5], Syntactic and meaning information is measured by means of questioning subjects under experiment (17, p. 325—363; 18, p. 14—28; 19, p. 137—138]. The entire procedure is based on the counts of signals' statistical frequencies. As to measuring 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 [20, p. 25—34]. Informational and semiotical analysis of these procedures and their comparison with the possibilities of a linguistic automaton* (LA) [21, 22] show that the computer, being the recipient of information, can receive and process not only the pre-information (potential information) and the syntactic information, which are determined by statistic and combinatorial properties of signals (names). Having entered a sufficiently comprehensive linguistic data base [5, p. 92—149; 23] into computer, we may simulate the reception and processing of meaning (sigmatic, semantic and pragmatic) information by an LA.

* The concept of a linguistic automaton refers to a combination of some formal language model, algorithms and programs which describe its operation, and the respective computer for its implementation.

The sigmatic information can be received and processed by an LA on condition that its data base is constructed after the denotatum principle, i. e. this base includes the denotata analogs of lexical and grammatical signs. At the same time, the semantic information becomes accessible to an LA on condition that the linguistic data axe arranged in the LA data base as a semantic network, which represents a system of semantic relations (cf. Saussure's 'valeurs') of signs and their meaningful combinations [24—26].

Theoretically, the goal-setting and priority strategies of some group of receivers can be also simulated by an LA which in this way obtains the possibility to perceive and process pragmatic information [27, p. 235—237; 28, p. 227—253].

But how can one explicate the mysterious mechanism for resignifying the linguistic sign meaning by the message sender and for decoding this secondary semiosis by the receiver who employs his own previous experience and information extracted from the non-linguistic and referential environment? These problems comprise the central philosophical question of machine translation theory and that of artificial intelligence. Until the theoretical speculations on this problem are embodied in actually functioning computer-based linguistic models generating real texts, all discussions in connection with high-quality machine translation will resemble the search for the philosopher's stone.

REFERENCES

1. Piotrowski, R. The antinomies of linguistics and automatic interpretation of the text. In: *3rd International Congress of Applied Linguistics, Copenhagen, 1972.*
2. Dreyfus, H. L. *What computers can't do. A critique of artificial reason.* New York: Harper & Row, 1972.
3. Bektaev, K. B.; Kenesbaev, S. K,; Piotrowski, R. G. Engineering linguistics. In: *Linguistics 200.* The Hague-Paris: Mouton, 1977.
4. Zarechnak, M. The history of machine translation. In: Henisz-Dostert, B.; Macdonald, R.; Zarechnak, M. *Machine translation.* The Hague: Mouton, 1979.
5. Lesokhin, M. M.; Lukyanenkov, K. F.; Piotrowski, R. G. *Introduction to mathematical linguistics. The linguistic application of mathematics fundamentals.* Minsk: Nauka i Tekhnika Publishers, 1982 (in Russian).
6. Hoffmann, L.; Piotrowski, R. *Beiträge zur Sprachstatistik. Linguistische Studien.* Leipzig: VEB Verlag Enzyklopädie, 1979.
7. Morris, Ch. W. Signs and the act. In: Morris, Ch. *Writings on the general theory of signs. Approaches to semiotics.* (Ed. by Th. A. Sebeok), 16. The Hague-Paris: Mouton, 1971, p. 401—414.
8. Hjelmslev, L. *Prolegomena to a theory of language. Supplement to International Journal of American Linguistics, vol. 19. Indiana University Publications in Anthropology and Linguistics.* Indianapolis: Indiana University Press, 1953, § 22.

9. Melnikov, G. P. *Systemology and linguistic aspects of cybernetics.* Moscow: Soviet Radio Press, 1978 (in Russian).

10. Mackay, D. M, Operational aspects of intellect. In: *Proceedings of N. P. L. Conference on Mechanization of Thought Processes. Vol. 1.* London, 1959.

11. Adrew, J. I. J. *Communication and meaning. An essay in applied modal logics.* Dordrecht: D. Reidel Publishing Company, 1983.

12. Piotrowski, R. G.; Rakhubo, N. P.; Khazhinskaja, M. S. *Systemic investigation of the lexicon of scientific text.* Chişinau: Ştiinţa, 1981 (in Russian).

13. Winograd, T. *Understanding natural language.* Cambridge, Mass.: Massachusetts Institute of Technology, 1972, § 3.3, example No. 97.

14. Nauta, D. *The meaning of information. Approaches to semiotics.* (Ed. by Th. A. Sebeok), 20. The Hague-Paris: Mouton, 1972.

15. Georgiev, H. C.; Piotrowski, R. G. A new method of measuring meaning. *Language and Speech,* 1976, vol. 19, p. 1.

16. Boguslawskaja, G.; Korzeniec, T.; Piotrowski, R. Language and information. In: Poznańskie towarzystwo przyjaciól nauk. Wydział filologiczno-filozoficzny. Komisja językoznawcza. *Biuletyn fonograficzny,* XIII. Poznań, 1972.

17. Novak, L.; Piotrowski, R. Experimento di predizione ed entropia della lingua rumena. In: *Statistica linguistica (con l'aggiunta di due appendici). Linguistica. Collezione di monografie originali o tradotte di linguistica generale, speciale ed applicata diretta da Carlo Tagliavini.* 3. Bologna: Casa editrice Prof. R. Patron, 1971.

18. Petrowa, N.; Piotrowski, R.; Giraud, R. Caractéristiques informationelles du mot français. *Bulletin de la Société de linguistique de Paris,* 1970, t. LXV, f. 1.

19. Piotrowski, R. Meaning information and its measures. In: *Soviet studies in language behavior.* Amsterdam: North-Holland Publishing Company, 1976.

20. Piotrowski, R. *Text, Computer, Mensch. Quantitative linguistics. Vol. 24.* Bochum: Studienverlag Dr. N. Brockmeyer, 1984.

21. Belyaewa, L. N.; Piotrowski, R. G. The structural approach to machine-aided translation and modelling of a unitary linguistic data base. In: *Sixth International Conference on Computer and the Humanities, June 6-8, 1983.* Raleigh, North Carolina: North Carolina State University, 1983.

22. Piotrowski, R.; Mikhailova, I. Procesamiento automático de textos españoles. In: *Homenaje al Prof. Dr. D. Alvaro Galmés de Fuentes. Vol.III.* Madrid: ed. Gredos S. A., 1984.

23. Vaina, L. Towards a computational theory of semantic memory. In: *Cognitive constraints on communication. Processes and representations.* (L. Vaina, J. Hintikka, eds.). *Synthese Language Library, 18.* Dordrecht: D. Reidel Publishing Company, 1983.

24. Goncharenko, V. V.; Shingaryova, E. A. *Frames for text meaning recognising.* Chişinau: Ştiinţa, 1984 (in Russian).

25. Brown, G. Linguistic and situational context in a model of task-oriented dialogue. In: *Cognitive constraints on communication. Processes and representations.* (L. Vaina, J. Hintikka, eds.). *Synthese Language Library, 18.* Dordrecht: D. Reidel Publishing Company, 1983.

26. Minsky, M. Jokes and the logic of the cognitive unconscious, *ibid.*

27. Piotrowski, R. G.; Beliayeva, L. N.; Popescul, A. N.; Šingareva, E. A. Bilingual annotating and abstracting. In: *Advances in science and technology. Series Informatics. Vol. 7. Automation of document indexing and abstracting.* Moscow: VINITI, 1983 (in Russian).

28. Brown, J. C.; Burton, R. R.; de Kleer, J. Pedagogical, natural language and techniques in SOPHIE I, II, and III. In: *Intelligent tutoring systems.* (Ed. by D. Sleeman and J. S. Brown). London, New York, etc.; Academic Press, 1982.