Linguistic Simulation of Semantic Invariants for Multilingual Knowledge Management Systems

Elena Kozerenko

Institute for Problems of Informatics of the Russian Academy of Sciences,
Moscow, Russia,
e-mail: kozerenko@mail.ru

Keywords: semantics, natural language, contrastive study, linguistic simulation, multilingual access to knowledge

Introduction

The problem addressed in this paper is establishment of semantic invariants to serve as a kind of metalanguage of "senses" valid for the natural language systems under consideration. We see the key objective of natural language processing in developing multilingual facilities of computer access to knowledge contained in texts. Our experience in design and implementation of natural language processors for knowledge-based systems allows us to assert that the focus of linguistic research is simulation of semantic-syntactic invariants [1-3]. For that we employ both theoretic study (contrastive analysis) of a subset of European languages (at present Russian, English and Italian) and computer experiment. The main hardships in NL "understanding" by a computer program lie not so much in lexical semantics but in the semantics of structures. The given paper states out basic conclusions of our latest work in linguistic simulation for the development of the natural language knowledge-based system IKS. At present our research efforts are continued in the going on innovative project dealing with multilingual document handling.

The focus of the given paper is consideration of semantic invariants which result from experimental linguistic simulation and theoretic study. The proposed solutions are deployable in knowledge management systems of various kinds. The insistent urge to develop multilingual access to knowledge contained in natural language (NL) texts (presented in electronic form) dictates the use of contrastive study for two and more languages. Contrastive study of the language systems enable us to single out the most general features which are characteristic for all the considered languages and, therefore could be a source for some basic semantic model invariant for all these languages. The current state of the art in natural language processing technology and theoretical and applied linguistics is taken into account. The given research was greatly inspired by the tradition of the strong semantic approach of Moscow linguistic school [4-6]. The principal models considered are semantic cases of Fillmore [7,8] relational grammar [9], applicative grammar [10]. Our research led us to some conclusions which accord with the Minimalist program of Chomsky [11]. Our point is that a valid basis for natural language simulation can be provided by the balance of the example-based and the rule-based approach. It is obvious that an excessive language resource is indispensable for quality solution of linguistic problems, but designing heuristics for structure analysis is not less important. The problem of semantic presentation of syntax was addressed by researches in our country and abroad [2,5-16], and models viewing the problem from different points were proposed: formal, logical, cognitive and functional. However, the urge of the present day tasks requires a synthetic approach.

©Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

Contrastive Study of Language Structures as a Source for Simulation Decisions

The principal hardship arising in various projects connected with natural language simulation is the sumultaneous involvement of different language levels (morphological, lexical and syntactic) in conveying the same meaning, i.e. if a certain notion in one language is expressed by means of a word, in another language it can be expressed by a morpheme, word combination, or even a phrase, for example:

IZBUSHKA (Russian) → CASETTINA (Italian) → A WEE LOG HUT (English)

i.e. the meaning of "diminutiveness" in some languages is regularly presented by morphological facilities (suffixes -ushk, -in), and in some languages by lexical means: the word "wee" in English. Thus we see that synonymity can be realized across different language levels. The same phenomenon can be observed inside the system of one and the same language: it is possible to say in Russian "malen'kaia izba", though the variant employing the suffix is more preferable as concerns the usage. Thus we see that the question of semantics can't be ignored even at the morphological level, and the the meanings of word-forming morphemes should be reflected in the language model. Here we encounter with the lexical units derivational process, and our approach is to simulate derivation starting with the initial most simple form, and to introduce semantics and establish synonymity where possible. A finite set of derivational history models is introduced for the lexical units and is stored in the dictionary, and interlanguage correspondences between these models are established.

However, the key question of the given paper is not so much the problem of lexical semantics, but the investigation of the functional units which could be called syntactic structures, or "syntaxemes" [10]. Consider the meanings of the system of grammatical cases in the Russian and the English languages which give us a good example of contrast between synthetic (i.e. employing inflectional endings) and analytic (i.e. using word order and prepositions) languages. The case meanings (object, recipient, instrument, etc.) are given by means of inflections in Russian, and via prepositions in the English and Italian languages (prepositions are considered to be lexical means - separate words, though some linguists tend to view them as morphemes, or even submorphems [16] on the grounds of their functional identity with morphemes). Let us illustrate the means of conveying the idea of "instrumentality" in these languages: Russian: "otkryt' dver' kliuchOM"; English: "open the door WITH a key"; Italian: "aprire la porta CON una chiave". In other words, for each sentence in a given language there exists a semantic equivalent in another language which, however, is not necessarily equal in length to the initial sentence.

The most important subgoal of our research is the construction of an integral presentation of sentence semantics structure. For that end we should establish the sense structure of a proposition, and investigate the semantic structure of various types of predicate expressions which constitute proposition. Sentence (S) is the basic unit of communication, and therefore, of meaning transmission. We consider sentence as a structural macrounit of "sense" which can be dissolved into component parts, and sentence structure elements will be studied from the point of view of their conveying deep semantic relations. The simulation efforts in our case are directed towards creation a unified sentence semantic structure which could serve a base for designing algorithms of natural language texts analysis allowing the transition from a natural language to another one without considerable "sense" losses in conveying relations presented by the structures of these natural languages.

The most frequently used predicate expressions besides "regular verbal predicates" in real natural language texts are those in which non-finite verb forms are used. Non-finite verb forms (further they will be called "verboids") are participles, infinitives and gerunds (verbal nouns are also related to this class in our system). A unified model for verboids is proposed. The principal characteristic of verboids is their "hybrid" nature, i.e. combination of verbal and nominal features. And in all three languages under consideration the meanings conveyed by these verbals and verbal phrases form a regular system with certain typological properties which serve the basis of the proposed model. The following meaning types conveyed by verboids in the Russian, English and Italian languages are singled out: Verbal Definition (V_D), Verbal Circumstance (V_C), Verbal Entity (V_E). Substancial similarity of form (morphology) and meaning of verbals is observed in all three languages. Let us give a comment on each form.

a) Verbal Definition (V_D):
a1) Active (the action is performed by an active agent):
e.g. The boy playing in the garden;
mal'chik, igraiuschii v sadu;
il bambino giocante nel giardino.
a2) Passive (the action is performed with the object):
e.g. The performed work; vypolnennaia rabota; il lavoro fatto.
b) Verbal Circumstance (V_C):
b1) Imperfective (the action accompanying the action expressed by the finite verb form in a sentence):
e.g. igraia, playing, giocando;
Working at the library, I didn't attend the lectures.
Lavorando alla biblioteca, non frequentavo le lezioni.
b2) Perfective (completed before the action of the finite verb form started):
e.g. poigrav, having played, avendo giocato
Having answered the question, he sat down at the table.
Avendo risposto alla domanda, lui si siede alla tavola.
c) Verbal Entity (V_E) (action having nominal character),
c1) Verbal Entity-Name (V_E_N is expressed by the infinitive form):
e.g. igrat', to play, giocare.
c2) Verbal Entity-Process (V_E_P is expressed by gerunds and verbal nouns:
e.g. chtenie, reading.

Primarily the language systems of Russian and English along with problem-oriented texts in these languages were subjected to contrastive study. Later a few experiments were carried out for the Italian language. Our experience shows that of prime importance is the focus on the semantics of syntax. Another important point is interrelation of different natural language levels: morphological, lexical and syntactic in conveying a certain meaning. The excessive study of language synonymity was undertaken, with special stress on the synonymity of structures. Lexical semantics was studied for subject areas under consideration, the thesauri were worked out presenting lexical units as a family tree of open class words. The most considerable in size was the thesaurus for the subject area of politological forecast (the areas of research establishment management and criminal police reports also underwent simulation).

Implementation Experience

Application knowledge-based systems were implemented on this foundation. The key problem was simulation of a NL sentence as a whole unit to provide its interpretation and inner representation as a globally interconnected structure, and not as an unstructured conglomeration of words and phrases. The representational mechanism employed in our work is the extended semantic networks. The nodes of semantic networks may contain not only elementary objects (such as words) but other networks representing compound structures (such as infinitive, participial, gerundial constructions and clauses). Simulation of embedded structures of any degree of complicity is supported by the extended semantic networks. A sentence is formed by a predicate construction, mainly presented by a verb or a verbal phrase. A verb determines the framework of a sentence. We worked out its representation as a case frame, where cases are the source of constraints with double nature: functional and lexical semantics. The first dictates the syntactic roles of eligible arguments and the second determines admissible classes of words.

The major difficulties for machine understanding of NL sentences resulted from transformations of finite (we take them for a "norm") verbal forms into nonfinite. The shift of cases in verbal transformations has the similar character in all the languages under study. Here the language units of different languages display semantic equivalence and congruency of form. We single out the "primary" semantic element (the meaning conveyed by a "normal" form), then secondary, terciary etc. meanings conveyed by derivative forms. The clear and precise description of the mechanism of syntactic derivation and formulation of it as an algorithm is our primary concern. We postulate the realization of a certain language unit meaning as a juxtaposition of its primary, secondary, etc. meanings, determined by the derivational history of a given language unit. The derivational history is fixed in production rules of the specialized logical programming language DECL which represent the program of semantic networks processing in the course of NL sentence interpretation. Of primary importance is the study of nonfinite verbal forms as their transformations inflict the shift of the whole case frame in a sentence. For example, let us consider the phrase: (1) Department N develops hybrid systems. Here the verb "develops" in its finite form displays "normal" distribution of syntactic-semantic cases. If we consider the transformations: (2) The hybrid systems developed by Department N (3) Development of hybrid systems by Department N ... we see that the semantic case of Agent (Department N in our case) which in example (1) was expressed by the subject relation R1 in (2) and (3) becomes R3 (indirect object). When verbal case frame arguments are considered, special attention is paid to sentencial arguments. Their syntactic functions are similar to those of nouns and nominal phrases which can occupy the same argument positions. Thus a compound sentence with different types of subordinate clauses is treated by our processor as a simple sentence in which some members can be expressed by other sentences whose syntactic function is similar to that of the corresponding elementary member.

We propose the hybrid semantic-syntactic model which employs six basic syntactic functions (relations) which together with their derivational histories are put into correspondence with semantic cases (a selected subset of semantic cases proposed by Fillmore). The relations namely are: R1 - subject, R2 - direct object, R3 - indirect object, R4 -attribute, R5 - verbal or attribute modifier, R6 - macrolink (a relational pronoun acting as a conjunction of a subordinate clause). The selected semantic cases are A: agent, O: object, D: characteristic, M: method, L: locative, P: purpose. Our selection of these types of semantic cases and relations was dictated by the principle of "rational sufficiency". The semantic cases are employed to constitute the semantic invariant of a sentence at the level of knowledge-base structures. The open class words produce the lexicalization of case frames. This model serves as an intralingua providing the common algorithmic platform for a multilingual processor which performs the procedures of NL analysis and synthesis and the shift from one natural language to another. The major results presented in this paper have been employed and evaluated in a series of projects dealing with construction of the natural language hybrid expert system environment DIES, expert system shell LOGOS, and the knowledge-based system IKS. A number of application expert and knowledge-based systems were developed for the subject areas of politological forecast, research establishment management, criminal police reports and press-relises handling. These systems operate on IBM PC computers under MS DOS and the version for Windows is being developed.

Acknowledgements

This research is supported by grant number 00-06-80069 of the Russian Foundation for Basic Research.

Conclusion

>From the very beginning of our research our goal was designing a multilingual system supporting a variety of European languages (Russian, English and Italian - which are representative languages of Slavonic, German and Roman language groups respectively). This meant that the language model underlying our simulation technique should be a unified language system valid for all the envisaged languages. Our experience demonstrates that the main difficulty lies not so much in the simulation of lexical semantics (especially for the domain-specific lexics, though a lot of problems not stated in the given paper arise here as well: for example the problems of lexical "holes" in one language compared with another), but in the simulation of structures having the same functional meaning across different natural language levels.

The original contribution of our work as compared to the work of other researchers is the development of a hybrid semantic-syntactic model for invariant sentence structure presentation valid for a number of European languages. As to our previously reported works [1-3], in the present paper we formulate our latest simulation results. For the first time we introduce the idea of sense realization as a juxtaposition of "primary", "secondary", etc. meanings of a language unit. We consider the derivational history of language units.

Facing present days tasks of natural language processing it is clear that the most required practical impact of structural semantic research would be the development of a semantic retrieval engine for the needs of processing corporation knowledge presented in natural language text form, for the World Wide Web Consortium activities and others.

References

[1] *Kouznetsov, I.P.* Semanticheskie Predstavleniia. M.: Nauka, 1986.

[2] *Kouznetsov, I.P., E.B. Kozerenko.* In Search for Language Universals: Linguistic Simulation Based on Extended Semantic Networks. In Dialogue'99: Proceedings of the International Workshop "Computational Linguistics and its Applications", A.S. Narinyani (ed.) Tarusa, 1999, Vol.2, 164-172.

[3] *Kozerenko, E.B.* O Podhode k vyiavleniu universalnyh semanticheskih kategorii i sposobov ih vyrazhenia v razlichnyh iazykovyh sistemah. In "Sistemy i sredstva informatiki", Vol.5, Moscow: Nauka, 1993, pp. 53-61.

[4] *Apresian, Y.D. Leksicheskaia Semantika.* Sinonimicheskie sredstva iazyka. Moscow: Nauka, 1974.

[5] *Melchuk, I.A.* Opyt Teorii Lingvisticheskih modelei "Smysl-Text". Moscow: Nauka, 1974.

[6] *Boguslavsky, I.M., Tsinman L.L.* Semantics in a Linguistic Processor. Computers and Artificial Intelligence, Vol. 11, No 4, pp. 385-408.

[7] *Fillmore, C.* The case for case reopened. In P. Cole & J.Sadok, Eds. Syntax and Semantics. 1977. New York: Academic Press.

[8] *Fillmore, Ch.J., P. Kay.* Construction Grammar Course Book. Berkeley: Univ. of Calofornia, 1992.

[9] *Perlmutter, D. and P. Postal.* The 1-Advancement Exclusiveness Law. In D. Perlmutter and C. Rosen (eds.) Studies in Relational Grammar 2, 81-125. Univ. Of Chicago, Chicago Press, 1984.

[10] *Shaumyan, S.* A Semiotic Theory of Language. Indiana University Press, 1987.

[11] *Chomsky, N.* A Minimalist Program for Linguistic Theory. In K. Hale & S.J. Keyser (eds.) The view from Building 20. 1-52, MIT Press, Cambridge, MA, 1993.

[12] *Paducheva, E.V.* O Semantike Sintaksisa. Moscow: Nauka, 1974.

[13] *Shank, R., Birnbaum, L., Mey J.* Integrating semantics and pragmatics. Quaderni di semantica. Vol. VI, N2, 1985.

[14] *Bock K., Loebell H.* Framing sentences. Cognition, Vol.35, 1990, pp.1-39.

[15] *Stabler, E.P.* The logical approach to syntax: foundations, specifications and implementations of theories of government and binding. MIT Press, Cambridge, MA, 1990.

[16] *Langacker, R.* Foundations of Cognitive Grammar. Vol.II: Descriptive Application. Stanford: Stanford University Press, 1991.

[17] *Kurylowich, Jerzy.* Derivation lexicale et derivation syntaxique. Contribution a la theorie des parties du discours.1936, B.S.L. 37:74-92. Reprinted in Jerzy Kurilowicz, Esquises linguistiques I. Munich: Wilhem Fink Verlag.