Improving Semantic Component of the Meaning ⇔Text Model

Valery D. Solovyev

One possible way to get a detailed description of natural language for computer use is the construction of Meaning ⇔ Text models, the model by I. Mel'čuk being one of the most developed among them. This paper proposes certain tools to overcome some difficulties the Mel'čuk's model has met with. A language of semantics representation is specified, to perfect Mel'čuk's model theoretically and to promote finally a full computer realization of the model.

1 INTRODUCTION

The Meaning ⇔ Text model developed by I. Mel'čuk in 70s (partially represented in English in [1]) consists of elaborated semantic, syntactic, and morphologic components, and it seems a good pattern of the computer-oriented language description, with Russian as the most developed example. It has exerted a great influence on the development of linguistics in Russia, but is practically unknown in the West and, thus, fallen out of the mainstream of modern linguistics as a whole. By this work we try to discuss the most important, in our opinion, reason of such situation. In addition, some approaches to the further model development are provided.

We propose a general approach to modernization of semantic representation. This implies essential extension of tools, in comparison with those used by Mel'čuk, with simultaneous imposing restrictions on the admissibility of different semantic graph structures. The semantic representation is constructed by combination of well-known and enough developed tools: frames, taxonomic structures, and feature spaces.

This paper shortly outlines the main elements of the project aimed at the further development of Mel'čuk's Meaning ⇔ Text model. Our approach can be also used for construction of language processors based on other theories.

2 DIFFICULTIES IN REALIZATION OF "SEMANTIC-RUSSIAN" DICTIONARY

While passing from the meaning to the text, the essential point is the transformation of the semantic representation (in the shape of a graph) to the syntactic one (in the shape of another graph). The process involves division of a semantic graph into fragments that may be changed by lexemes of a concrete language. In accordance with original work by Mel'čuk [2, p.178], such transformation should be done "by means of the semantic – specific language dictionary introduced as a list of rules X $\Rightarrow$ Y, where X is a certain semantic graph (i.e. a subgraph of the semantic graph allied to the original semantic representation), and Y is a subtree of the syntactic tree having a generalized lexical unit of the considered language, i.e. a lexeme, in the head." Regrettably, it is not clear how to realize these issues. Mel'čuk continues: "No investigation of the formal side of such operations (i.e. "reading" a tree from a compound graph) was carried on." In reality, such a semantic-language dictionary was not constructed in [2], and even theoretical basics of its construction were not laid.

During 25 years passed since that time (some recent studies in this field can be found in [3]) the situation has not changed much. Indeed, we have not ever a zero-level instance of the semantic-language dictionary. In our opinion, it is implied neither by the absence of theoretical developments, nor by difficulties in constructing the mentioned dictionary. The cause of it is the absolute impossibility of such construction, at least in the form represented by [2]. The fact is that in the semantic language we can describe too many situations – much more than a number of words in every natural language, and much more than it needs in practice. We pay tribute for the unlimited universality of the semantic language. Let us exemplify the matter.

We shall consider emotions which could be characterized in the next way: "Annoying state of the mind in the situation of absence of what a human wishes, when he or she thinks that his wish is unreachable" [4]. We shall denote this group of emotions by the symbol G. To express these emotions, the Russian language includes words of the following synonymy set – {*toska, unynie, pechal', grust'*} [4]. Of course, they are not absolute synonyms and differ in duration of the state, its deepness, measure of annoyance, manner of the outer display, etc. [4]. Enumerated properties can be considered, with some simplification, as parameters taking one of several discrete values. Learning on the analysis given in [4] for that synonymy set, we may distinguish values of the enumerated parameters as follows:

- Duration: <transitive state (marked with the digit 0), prolonged state(1)>.
- Deepness: <not deep (0), deep (1)>.
- Measure of annoyance: <not annoying (0), annoying (1), sickly (2)>.
- Manner of display: (1) being thoughtful, feeling slack, absence of animation, seriousness; (2) bad mood; (3) apathy, gloom, weakness, reluctance to do something.

Then according to [4], values of words of that synonymy set may be figured by the following table.

Emotion	Duration	Deepness	Measure of Annoyance	Manner of display
Toska	1	1	2	2
Unynie	1	1	1	3
Pechal'	1	0-1	0-2	1
Grust'	0	0	0	1

Being universal, the semantic language must permit to express any parameter given above and, accordingly, any combination of them. The use of just a dictionary (i.e. of a static data structure instead of some algorithm) in transition from the semantic representation to the syntactic one, means that input set of semantic-Russian dictionary has to be formed by semantic graphs according to all possible combinations of those parameters. But a whole number of combinations is $2 \times 2 \times 3 \times 3 = 36$, that is much more than a number of words in the synonymy set. Meanwhile, not all of the parameters were enumerated here, since we have such related ideas as “a wish to change the situation”, “a specification of causes”, etc.

It is quite possible that certain, or many, combinations of the feature values make no sense, since they do not correspond to any situation in the real world. For example, a man can scarcely feel emotions with the following combination of parameters: the measure of annoyance 0, the manner of display 3.

However we have no mechanism to reject nonsensical combinations. As for Mel’čuk’s MTM, it principally denies to analyze the semantic representation on the subject of its meaningfulness, consistency, etc. Therefore, the number of entries for such dictionary may be incredibly high (probably, hundreds of million units and even more), whereas the most part of its points has no sense and are inapplicable.

Thus, in its entire original form, the MTM seems unrealizable.

3 ALGORITHMIC SINGLING OUT A LEXEME FROM THE SEMANTIC REPRESENTATION

The above mentioned problem caused by enormous number of semantic structures cannot be solved by means of the dictionary based on direct correspondence. But it may be settled by using an algorithmic approach. The algorithm having finite (and possibly not so large) size can be applied to the potentially infinite range of input data. Let us first demonstrate possibility of the algorithmic approach on the emotions from the G group.

In English we can find the word *blues* corresponding to the emotion of that kind and being characterized in [5] as “fits of bad mood that are often annoying but short-lived and not deep” (based on [6]). From the given description we find out that the parameter “duration” takes the value 0, the parameter "deepness" – also 0, the parameter " manner of display " – 2. As to the parameter "measure of annoyance", to determine its value exactly is difficult. It ought to be either 2 or 1.5, if we accept intermediate values. Considering the case with account of intermediate values, let us characterize *blues* by the vector {0, 0, 1.5, 2}.

This word should replace the corresponding semantic graph in a proper way. How could we transform this semantic representation to the Russian words? The exact translation is absent because there is no Russian word to reflect the emotion from the G group with such a set of properties. The following approach is possible in this case.

It is necessary to represent the words under consideration by point in multi-dimension feature space and to calculate distances in this space. In our approach, we use Euclidean metrics. Of course, it is one of possible kinds of metrics. Further investigations will prove which metrics is the most appropriate for the task. The word from the target language standing at the minimal distance from the source word may be taken as its translation equivalent.

When using this approach, we have some troubles.

Case 1. It is possible that distances from the source word to several target words – translation candidates – are equal. Then there may be two options:

- If the source word and the candidates have the same difference between feature values, for each feature, it is possible to choose any candidate.
- There are at least two features with the following condition: the value of the first feature of the source word is nearer to one candidate, while the value of the second feature to other candidate. Then we can solve the problem by introducing the weights reflecting the importance of each feature. In our example of translating from English to Russian, the analysis shows that the "measure of annoyance" is more important than the "manner of display". Hence, we assign a larger weight to the parameter "measure of annoyance".

Choosing appropriate weights, we can exclude cases with equal distances. Introduction of weights will give an opportunity to avoid ambiguity of the choice and to raise the quality of translation. Weights of the same feature may have changed in different fields, of course.

Case 2. Some cells of the table can contain intervals instead of specific values. At the extreme, these intervals cover all possible values of the feature (we may call this feature as irrelevant). Henceforth, we shall consider the word with this property as having general meaning (in comparison with other words from the chosen synonymy set). In our consideration, it is applicable to the Russian word *pechal'*.

We can formulate two methods for operating with the interval valued cells.

Method 1. To change the interval with the average, and then to calculate a distance in usual way. Applying this method to the word *blues,* we get the following.

Step 1: Parameter value assignment for the word *pechal'* is (1, 0.5, 1, 1).

Step 2: Calculation of Euclidean distances gives: ρ(*blues, toska*) = sqrt(2.25), ρ(*blues, unynie*) = sqrt(3.25), ρ(*blues, pechal'*) = sqrt(2.5), ρ(*blues, grust'*) = sqrt(3.25).

Thus, the best candidate for the translation of the word *blues* is *toska*.

Method 2. We operate with the main parameter values only, without taking into account intermediate points. Then two general principles are taken for this method:

1. to stay within interval is better than to reveal a divergence between feature values;
2. coincidence of concrete values is better than covering with an interval.

The principles should be realized in the next way.

Principle 1. If one of compared words S1 and S2 has *k* be the value of a certain parameter when another word has interval [n1, n2], and $k \in$ [n1, n2], then we should put *k* instead of interval [n1, n2] in the formula for distance calculation. Hence, the coincidence of values of this parameter for the words S1 and S2 is reached.

Principle 2. If distances from the source word to the candidates calculated in accordance with the principle 1 are equal, then one should prefer the candidate with narrower interval.

This formulation for realization of the principle 2 is not strict, since it is necessary to introduce a linear order on the set of candidates. A more precise definition is the subject of further research.

Let us now apply the method 2 to the lexeme *blues*. Its set of parameter values will be represented by the string (0, 0, 2, 2). Realizing the principle 1 on the word *pechal'*, we get the vector (1, 0, 2, 1). Distances are calculated accordingly: ρ(*blues, toska*) = sqrt(2), ρ(*blues, unynie*) = sqrt(4), ρ(*blues, pechal'*) = sqrt(2), ρ(*blues, grust'*) = sqrt(5). Since the distances from the word *blues* to the words *toska* and *pechal'* are equal, whereas the word *toska* is more concrete in meaning, we realize the principle 2 by the translation of *blues* to *toska*. It is characteristic that both methods have given the same result.

We have to discuss translation of the words with general meaning separately. For this case it would be natural to choose the word-candidate with general meaning too. The English word *sadness* being characterized by string of the value (0-1; 0-1; 0-2; 1-3) has the most general meaning among others from the considered class of emotions G. All English-Russian dictionaries provide this word with the Russian translation *pechal'* (which is as it should be), though some dictionaries include other possible translations as well. In the most part of instances from [5] the word *sadness* is translates as *pechal'.* Other translations are widely circumstantiated with additional informational being got from the context (or semantic graph) and concretizing values of parameters (incorporating common knowledge of world).

If the language does not possess the word with general meaning (from the field needed), then to get an additional information becomes an actual problem.

We can upgrade the quality of translation if using not the only lexeme but the whole noun phrase, with having provided the possible word interpretation with a range of concretizing definitions. Thus, starting from the lexeme *toska* with modifiers *neglubokaja* 'not deep' and *kratkovremennaja* 'short-lived', we shall get a very good correspondence to the semantics of the word *blues.* Modifiers within such an approach can be interpreted as an operator on lexemes changing values of certain parameters, similar to a lexical function by Mel'čuk. In fact, we need only two operators: increasing values of parameters and decreasing them. The algorithm forming an appropriate combination of operators is rather simple.

In generally case, singling out the proper lexeme can be done by an expert system operating with common knowledge of the world. However, in most cases a simple algorithm calculating minimal distance approximates does the job well enough.

4 GLOBAL SEMANTIC STRUCTURE

For Meaning ⇒ Text transformations, certain requirements upon the structure of the semantic representation should be set. It is rather evident that the semantic representation ought to be more heterogeneous in comparison with usual marked graphs and more structured, to simplify the procedure of singling out lexical and syntactic information. At the first step, a structure with following properties may be offered.

There must be described three metalanguages: of semantic concepts, of differential features and of graph structures. All they represent the semantic component of the model, and their destiny is to delimit and to describe precisely different elements of the semantic representation in order to overcome difficulties mentioned above.

4.1 Semantic concepts

The term “semantic concept” infers the essence treated indivisible (primitive) within a range of this formalism. It may be either a concrete object (e.g., *Lennon*) or an object class (e.g., *emotion*) able to be further detailed. Set of concepts is a hierarchy of ordered taxonomic structures. Instances of semantic concepts are shown with their internal structures.

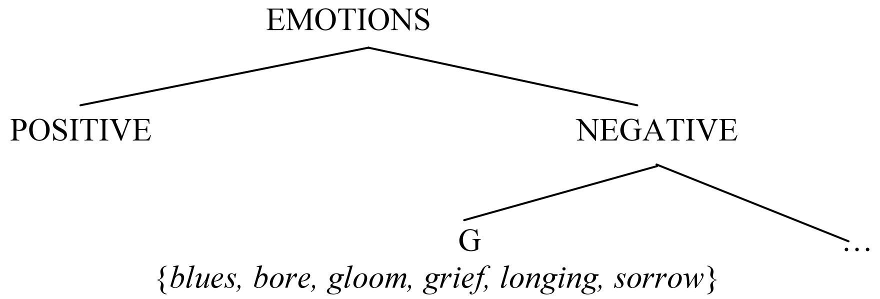

Figure 1. Leaves of this tree (sets of lexemes in brackets) represent quasi-synonymy sets.

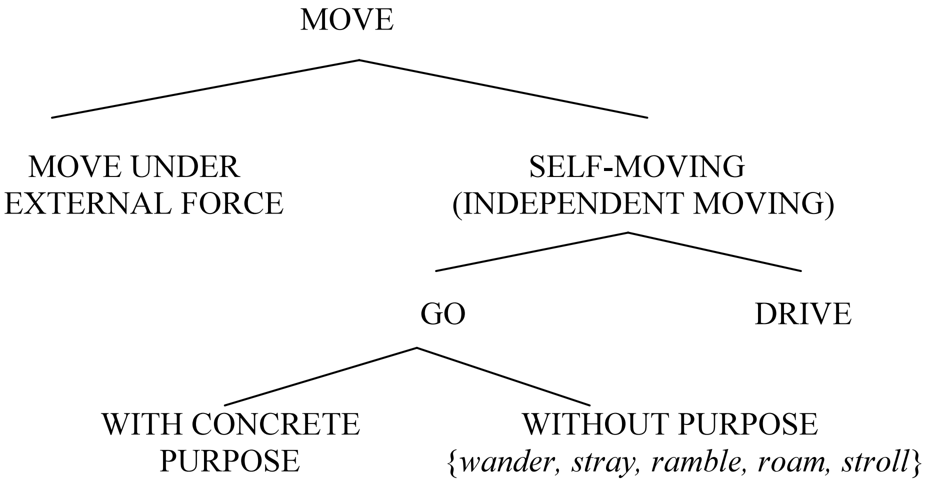
MOVE
MOVE UNDER EXTERNAL FORCE
SELF-MOVING (INDEPENDENT MOVING)
GO
DRIVE
WITH CONCRETE PURPOSE
WITHOUT PURPOSE
{*wander, stray, ramble, roam, stroll*}

Figure 2. Leaves of this tree also represent quasi-synonymy sets.

In this way, the model uses not only “genuine” primitives, but also a considerably enlarged class of concepts representable by appropriate lexemes. We can see a strong divergence between this viewpoint and such “minimalist” positions like promoted by Wierzbicka [7]. Our approach seeks for simplicity and transparency of the representation. Its realization supposes development of a full taxonomic classification for the language under investigation.

4.2 Differential features

Differential features are parameters attributed to the concepts. A specific feature value makes possible to identify a concrete lexeme from a synonymy set represented as a leaf.

For other semantic concepts, to choose value of the feature attributed to the concept means to make the concept more concrete so well as we could move down the true of taxonomic classes. For example, let us introduce the feature "independence" with possible values: "under external force" and "independently" attribute this feature to the concept "moving." It is obvious that the second level of concept on the Figure 2 deals with the mentioned values of the feature introduced.

Success of such approach will depend, to a great extent, on how large is the list of necessary concepts and features. One may hope it will prove not to be enormously enlarged by repeated using the same features with different concepts. Returning to our example, the feature "independence" can be associated not with the concept "moving" only, but with many others also. This feature differentiates values of the words *kormit'* 'feed' and *est'* 'eat', and, generally, makes the difference between causative and non-causative verbs. There are certain grounds to guess that the sufficiently exact representation of the semantics can be obtained using of no more than several hundreds features.

4.3 Graph structure

Each node of the graph is an independent semantic concept not decomposable into atoms. Each semantic concept is connected with its description formalized in the shape of frame, with pointers to subframes corresponding to the concepts of the low-level, slots associated with differential features, etc. Then, we shall get a graph of frames in the result.

General semantic concepts of the highest level ("moving" in that number) have been described in [8]. In [9] a sufficiently detailed description of the frame "independent moving" was proposed wherein 24 slots are enumerated with the restrictions on their values. Our extended possibilities in filling the semantic graph nodes allow us to impose the restrictions on structure of the graph itself. This could simplify reading the syntactic tree from the graph. We have to define a set of permitted structures. One possible variant appears with graphs representing Schank's conceptual dependency [10].

In the whole, keeping its universality, the semantic representation should be simpler and easily processed. We try to reach the result by making a balanced distribution of sense load between different components of the semantic representation (concepts, taxonomic connections, differential features, graph structure); for using only graph and semantic primitives (for example) inevitably complicates each of the components. Another methodological idea consists in complex application of well-studied constructions separately used before: frames, taxonomic classifications, n-dimensional space of differential features, etc.

The problem of whether the semantic concepts are language-independent is relevant. It is so for their higher levels. However, lower level of description stays language-dependent. In [11] the author has shown that structure of the semantic field of emotions in Tartar differs from the same structure in English and Russian.

Such language relativism extremely complicates the problem of constructing a universal metalanguage for semantics. However, a number of studies made in the field of linguistics and cognitive science render us optimistic prospects. First, we mean classical studies by Osgood on the semantic differentials [12] wherein the universality (i.e., language and culture independence) of some differential scales (“pleasant – not pleasant”, “active - passive”, etc.) has been proved.

In the case when appropriate concepts from different languages do not coincide, one may use the following procedure to describe the lexeme set which the proper translation should be chosen form.

Let the level *i* have concepts K1 and K2 (from different languages) with coincidence of the semantic contents; and let the level *i+1* fix the other structure of World (differing in turn from one language to another). Let K1 = $L_1 \cup \ldots \cup L_n$ and K2 = $M_1 \cup \ldots \cup M_s$. Let $\{M_{i_1}, \ldots, M_{i_t}\}$ be minimal set for which content of L_j is covered by content of $M_{i_1} \cup \ldots \cup M_{i_t}$. In this case we select the set $M_{i_1} \cup \ldots \cup M_{i_t}$ as the set to choose the translation lexemes from L_j. And after all steps we have got to apply the procedures described above in concern with synonymous rows.

Yet for bilingual translation systems the problem of describing the semantics of two languages in offered terms seems rather cumbersome and requiring large resources (both machine and human), but principally solvable. To solve it, we must, first, describe all synonymous rows in traditions of Apresyan’s school [4]. The close viewpoint has been stated earlier in [13].

5 CONCLUSIONS

The paper exposes certain ideas for increasing competitiveness of the Meaning ⇔ Text model. An improved semantic representation is featured:

- by creation of new semantic metalanguage, on the basis of the frame theory, admitting Apresyan’s synonymy sets as its organic component;
- by transition from the vocabulary approach to the algorithmic one in solution of the problem of transforming the semantic graph to the syntactic tree.

The paper outlines the semantic metalanguage consisting of three sublanguages: of concepts and frames, of differential features, and of graph structures. Methods of the algorithmic selection of an adequate sexeme from a synonymy set is also outlined.

REFERENCES

1. Mel'čuk, I. *Meaning-Text Models: A Recent Trend in Soviet Linguistics.* Annual Review of Antropology, v.10, 1981, p.27-62.
2. Mel'čuk, I. A. *Opyt teorii lingvisticheskih modelej "Meaning <-> Text"* (in Russian). Moscow, Nauka Publ., 1974.
3. Wanner, L. (ed.) *Recent trends in Meaning-Text theory* (d.), John Benjamins Publ. Com., 1997.
4. Apresjan, Yu. D. et al. *Novyj ob'jasnitel'nyj slovar' sinonimov russkogo jasyka* (in Russian). Moscow, Jasyki russkoj kul'tury Publ., 1997.
5. Apresjan. Yu. D. et al. *Anglo-russij sinonimicheskij slovar'* (in Russian). Moscow, Russkij jasyk Publ. 1979.
6. *Webster's Dictionary of Synonyms*. Springfield, Mass. 1963.
7. Wierzbicka, A. *Semantic Primitives*. Linguistische Forschungen. N.22. Frankfurt: Atheneum, 1972.
8. Dirven, R., M. Verspoor. *Cognitive Exploration of Language and Linguistics*. Amsterdam: John Benjamins Publ. 1998.
9. Solovyev, V. D. *Cognitive metalanguage to describe the semantics of movement verbs.* Proc. of the 2nd Int. Conf. on Cogn. Science. Tokyo: JCSS, 1999. P. 495-498.
10. Schank, R. *Conceptual dependency: A theory of natural language understanding*. Cognitive Psychology, v.3, N.4.
11. Solovyev, V. D. *Semantica glagolov pechali v tatarskom jasyke* (in Russian). Proc. conf. "Cognitive Modeling", in print.
12. Osgood Ch., C. J. Susi, P. H. Tannenbaum. *The measurement of meaning.* Urbana, 1957.
13. Gelbukh, A. *Between text and meaning* (in Russian, abstract in English). Proc. Annual International Conf. on Applied Linguistics Dialogue-99, May 30 – June 5, 1999, Moscow, Russia.

Valery D. Solovyev is a professor of Kazan State University (Russia). He is the author of about 90 publications in the theory of algorithms and cognitive linguistics, the President of the Russian Association of Cognitive Technology and Text Processing, member of the Association of Computational Linguistics (ACL). He can be reached at solovyev@tatincom.ru