Semantic Approach to NLP Problem Solving

Aliaksandr Hardzei
Minsk State Linguistic University
Minsk, Belarus
Email: alieks2001@yahoo.com

***Abstract*—On the basis of the third edition of the second version of Theory for Automatic Generation of Knowledge Architecture (TAPAZ–2), a new approach to the semantic markup of an event and the syntax formalization of Chinese, English and Russian sentences is proposed [1].**

***Keywords*—combinatory semantics, semantic markup, semantic case, semantic classifier, knowledge graph, role list of individs, subject, object, action, macroprocess, specialized process, world model, TAPAZ-algebra, TAPAZ-unit**

I. Introduction

Almost simultaneously with the advent of the Internet, the idea of building algorithms for relevant information retrieval according to strictly specified semantic rules was born in the American scientific community. In 2001, the creator of the World Wide Web and the current head of the W3C Consortium Timothy Berners-Lee, together with James Hendler and Ora Lassila, published a keynote article in Scientific American "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities" [1]. The article was focused on the development of semantic technologies for searching and processing information on the Internet and through the Internet: "To date, the World Wide Web has developed most rapidly as a medium of documents for people rather than of information that can be manipulated automatically. By augmenting Web pages with data targeted at computers and by adding documents solely for computers, we will transform the Web into the Semantic Web. Computers will find the meaning of semantic data by following hyperlinks to definitions of key terms and rules for reasoning about them logically. The resulting infrastructure will spur the development of automated Web services such as highly functional agents. Ordinary users will compose Semantic Web pages and add new definitions and rules using off-the-shelf software that will assist with semantic markup" [1: 36].

In fact, the authors offered an alternative to the statistical methods of data processing that were gaining popularity in Artificial Intelligence: "The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. Such an agent coming to the clinic's Web page will know not just that the page has keywords such as "treatment, medicine, physical, therapy" (as might be encoded today) but also that Dr. Hartman works at this clinic on Mondays, Wednesdays and Fridays and that the script takes a date range in yyyy-mm-dd format and returns appointment times. And it will "know" all this without needing artificial intelligence on the scale of 2001's Hal or Star Wars's C-3PO" [1: 37], and then: "For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. Artificial-intelligence researchers have studied such systems since long before the Web was developed. Knowledge representation, as this technology is often called, is currently in a state comparable to that of hypertext before the advent of the Web: it is clearly a good idea, and some very nice demonstrations exist, but it has not yet changed the world. It contains the seeds of important applications, but to realize its full potential it must be linked into a single global system" [1: 37].

II. TOWARDS COMBINATORY SEMANTICS

Since with the help of N. Chomsky's transformational grammar, on which context-dependent and context-free grammars were built, and that became the basis for higher-level object-oriented programming languages, such as Java, C++, C#, etc., it was possible only with varying success to formalize syntax, but not semantics, the views of American scientists turned to the semantic cases of Ch. Fillmore [2–5], Jackendoff's early work [6, 7] and Stowell's "theta-grids" [8]. It is curious that at the same time the formal grammar of R. Montague [9] with its PTQ (Proper Treatment of Quantification in Ordinary English) and lambda abstraction was pushed into the background, although it was this grammar in a number of cases that generalized in terms of mathematical logic the achievement of generative semantics by G. Lakoff [10] and the interpretive semantics of R. Jackendoff. It is also curious that the semantic syntax of L. Tesnière [11] with verb nodes of actants and syrconstants, in fact

[1]The article develops the scientific provisions formulated by the author in the following works: A. Hardzei, "Plagiarism Problem Solving Based on Combinatory Semantics". *Communications in Computer and Information Science (CCIS).* Switzerland: Springer Nature Switzerland AG, 2020, vol. 1282, pp. 176–197. Available: https://link.springer.com/book/10.1007%2F978-3-030-60447-9 and A. Hardzei, "Semantic Markup of the Event and its Display by Means of the Chinese and Russian Languages". *Foreign Languages in Tertiary Education*, 2021, no. 2(57), pp. 5–26.

– analogs of the semantic cases of Ch. Fillmore, was also taken out of the brackets – the works of L. Tesnière were published for 30 years earlier, and the monograph by V. V. Martynov "Cybernetics. Semiotics. Linguistics" with a prototype of the Universal Semantic Code (USC) and a description of the roles of signs in the nuclear semantic string *subject – action – object* (SAO) – for 2 years earlier than the case grammar of Ch. Fillmore [12]. Note that at present linguistics has only one synthetic (sequentially deductive and procedural) model of language – the Panini grammar, dating from the 5th century BC, in which 3959 short sutras (rules) totally described the generation, construction and transformation of all Sanskrit units, starting from the phonetic-phonological level and ending with the semantic-syntactic level [13–15]. It is still not clear what formalisms were used as the basis for such an accurate description of a natural language and how it was possible to achieve this in such ancient times, just as, for example, it is not unknown, what kind of mathematics were used to collect the hexagrams of the ancient Chinese "Book of Changes" (易經), the analysis of which led Leibniz to the idea of binary calculus, that became the basis of modern computing, one thing is clear – European linguistics, first of all, French, in a hidden form borrowed a number of postulates of Panini's grammar, in particular, that the case is not so much morphological as semantics-syntactic category – the founders of structural linguistics, of course, knew about Panini's grammar, the departments of Sanskritology were in many European universities. However, we emphasize that semantic cases were important, but not the only achievement of Panini. They were calculated by some algebra and organically fitted into the entire architecture of grammar. Without this algebra, it was possible, albeit with difficulty, to translate Panini's grammar from one language to others, but it was impossible to describe other languages, like Panini: the formalization of languages and translation are rather different tasks. Therefore, V. V. Martynov started looking for such an algebra.

The first version of USC was published at 1974, 1977 – the second, 1984 – the third, 1988 – the fourth, 1995 – the fifth, 2001 – the sixth [16–21]. From version to version the algebraic apparatus and the list of semantic primitives were improving. Thus, the list of tasks to equip computer with encyclopedic knowledge bases was narrowed, and finally the list consists of five components:

1. "To calculate semantic primitives, i.e. semantically irreducible kernel words and define rules of their combinatorics.
2. To define the necessary and sufficient set of formal characteristics constituting 'dictionary entry'.
3. **To define a set of semantic operations for calculating a subject domain of any kind.**
4. To propose heuristic teaching rules to work with the system.
5. To build a system of mutual references based on semantics" [21: 42].

In 1993 achievements in the approach allowed the researchers of the center "Semantics" of Minsk Linguistic State University, headed by V. V. Martynov, to begin an intensive research of ways to expand the basic semantic classifier to the encyclopedic knowledge base. In 1994, the first procedure of calculating the subject domains in the form of a directed graph of complex strings was proposed by A. Hardzei [22]. Use of the procedure has required the establishment of a one-to-one (vector) transition between actions in basic semantic classifier and has led to the creation of the automatic generation of knowledge architecture theory (TAPAZ) which was founded on: the formal theory; the semantic counterpart; the set of macroprocesses (actions) as semantic primitives; the algorithm defining roles of individs, and the knowledge graph for searching processes through macroprocesses (see Fig. 1) [23, 24].

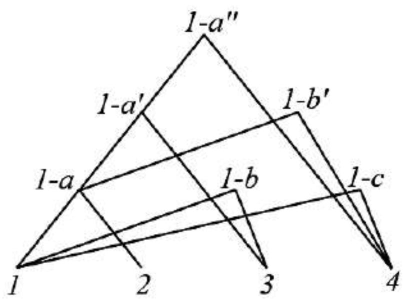

Figure 1. TAPAZ Knowledge Graph (the Semantic Classifier Graph). where: *1* – active macroprocess; *2, 3, 4* – clarifying macroprocesses; *1-a, 1-b, 1-c* – derivative processes with *1-a* as the active derivative process; *1-a'* and *1-b'* – derivative processes of the second level with *1-a'* as the active derivative process of the second level; *1-a"* – the active derivative process of the third level.

For example, the macroprocess '*restore*' may be considered as a set and the processes '*treat*', '*repair*', '*adjust*' as its subsets. Such subsets represent isomorphism of subject domains and create a knowledge structure where subsets of processes fill cells of the knowledge structure with a concrete content [1]. TAPAZ–2 as the new version of the Theory for Automatic Generation of Knowledge Architecture differs from the previous version in several ways: simplified algebraic apparatus, increased number of rules for interpretation of the standard superposition of individs, and minimized semantic calculus. The number of operations with the strings of semantic code are reduced to two and it is now the algebra type:

$$A = \langle M, *, - \rangle \quad (1)$$

[1]For a detailed description of the TAPAZ Knowledge Graph, see: A. Hardzei, A. Udovichenko, "Graph of TAPAZ–2 Semantic Classifier". In: V. V. Golenkov et al. (eds.) CONFERENCE 2019, *Open Semantic Technologies for Intelligent Systems (OSTIS)*. Minsk: Belarussian State University of Informatics and Radioelectronics Publ., 2019, iss. 3, pp. 281–284. Available: http://tapaz.by.

where: M is a set of elements, '*' is operation of superposition, '–' is operation of extension [1].

Examples of TAPAZ formulas:

$$((((((((((((((X*\bar{X})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{Z}})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}}; \tag{45}$$

$$((((((((((((((X*\bar{X})*\bar{\bar{X}})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}}; \tag{46}$$

$$((((((((((((((X*\bar{X})*\bar{\bar{\bar{Y}}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{\bar{Z}}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W})*\bar{\bar{\bar{W}}})*\bar{W}; \tag{47}$$

$$(((((((((((((((X*\bar{X})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Y}})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}})*\bar{W}; \tag{48}$$

$$(((((((((((((((X*\bar{X})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{Z}})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}})*\bar{W}; \tag{49}$$

$$(((((((((((((((X*\bar{X})*\bar{\bar{X}})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}})*\bar{W}; \tag{50}$$

$$(((((((((((((((X*\bar{X})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}})*\bar{W})*W; \tag{51}$$

$$((((((((((((((((X*\bar{X})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Y}})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}})*\bar{W})*W; \tag{52}$$

$$((((((((((((((((X*\bar{X})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{Z}})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}}*\bar{W})*W; \tag{53}$$

$$((((((((((((((((X*\bar{X})*\bar{\bar{X}})*\bar{\bar{Y}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{Z}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{W}})*\bar{W})*W)*\bar{W})*\bar{\bar{W}})*\bar{W})*W; \tag{54}$$

$$((((((((((((((((X*\bar{X})*\bar{\bar{\bar{Y}}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{\bar{Z}}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W}; \tag{55}$$

$$(((((((((((((((((X*\bar{X})*\bar{\bar{\bar{Y}}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{\bar{Y}}})*\bar{\bar{\bar{Z}}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W}; \tag{56}$$

$$(((((((((((((((((X*\bar{X})*\bar{\bar{\bar{Y}}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{\bar{Z}}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{\bar{Z}}})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W}; \tag{57}$$

$$(((((((((((((((((X*\bar{X})*\bar{\bar{\bar{X}}})*\bar{\bar{\bar{Y}}})*\bar{Y})*Y)*\bar{Y})*\bar{\bar{\bar{Z}}})*\bar{Z})*Z)*\bar{Z})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W})*\bar{\bar{\bar{W}}})*\bar{W})*W)*\bar{W}. \tag{58}$$

TAPAZ–2 is a tool for generating a world model in a form suitable for Natural Language Processing in systems of Artificial Intelligence. The Intellectual Knowledge Base built in a computer combines the Semantic Classifier – a final ordered (vector) set of semantic primitives (actions and roles of individs) and the Semantic Ontology – an algorithm for generating new sense units based on the original set of primitives, presented in the form of TAPAZ Knowledge Graph.

An Intelligent (Expert) Search System based on the TAPAZ–2 Semantic Classifier may consist of an intelligent search engine that selects and reviews content on a given topic from the Internet, and a dialog user interface that allows the system to process user requests and transform them in the canonized text corresponding to the machine-readable the World Model, and the user will confirm whether this conversion was performed correctly, and if not, then offer his own decoding through the Semantic Classifier.

This Intelligent Search System can be used to solve various problems, including the task of automatic identification of semantically equivalent fragments of text documents, which will be discussed below.

The main components of this Intelligent Search System are: an online content monitoring module with adequate crawler and stapler; automatic lexical analysis module with a tagger on a semantic (Parts of Language), not on a morphological (Parts of Speech) basis; dynamic syntactic analysis module with a recursive reconstruction algorithm (parser) for sentence string elaborated in combinatory semantics technology; a module for direct and reverse conversion of syntactic expressions into TAPAZ–2 algebraic formulas, as well as the Intellectual Knowledge Base (IKB) consisting of TAPAZ-units (IKB-taxonomy) assembled in the order indicated by the Knowledge Graph of the Semantic Classifier (IKB-ontology). The knowledge base taxonomy also serves as a corresponding subject domain semantic dictionary during automatic lexical analysis.

III. ROLE LIST OF INDIVIDS

Combinatory semantics studies the linguistic mapping of the dynamics of individs' roles in an event. Its founder, as we consider, is Z. Harris, who put forward the nuclear semantic string *subject – action – object* as a starting point of formalizing sentences [25]. Research in this direction was continued at Minsk Semantic School under the guidance of V. V. Martynov and A. Hardzei. Combinatory semantics should not be confused with combinatorial semantics, which studies co-occurrence of signs using statistical methods, the founder of which is also Z. Harris [26].

The updated vector role list of individuals ordered by the TAPAZ-algebra (The TAPAZ Role List) [2] is the following set:

subject *(initiator → spreader → inspirer → creator) →* ***instrument*** *(activator → suppressor → enhancer → converter) →* ***mediator*** *(landmark → locus → carrier → adapter → acceptor → stock → separator → material → model → retainer → resource → stimulus → regulator → chronotope source → indicator) →* ***object*** *(coating → hull → interlayer → kernel) →* ***product*** *(billet → semi-product → prototype → end item),*

were: ***subject*** – the originator of the action, varieties of the subject: *initiator* – initiates the action, *spreader* – spreads the action, *inspirer* – involves into the action, *creator* – completes the action by making a product from the object; ***object*** – the recipient of the action, varieties of the object: *coating* – the outer insulation of the individ's shell, *hull* – the individ's shell, *interlayer* – the inner insulation of the individ's shell, *kernel* – the core of the individ; ***product*** – the result of the subject's impact (action) on the object (the individ adapted to a given role in a new action), varieties of the product: *billet* – the object turned into a raw material, *semi-product* – the product half-made from raw materials, *prototype* – the prototype product, *end item* – the finished product; ***instrument*** – the performer of the action, the closest individ to the subject, varieties of the instrument: *activator* – directly affects the mediator, *suppressor* – suppresses the resistance of the mediator, *enhancer* – increases the effect on the mediator, *converter* – converts the mediator into the instrument; ***mediator***, i.e. the mediator of the action – the closest individ to the object; varieties of the mediator: *landmark* – orientates the impact on the object, *locus* – the closest environs of the object partially or completely surrounding the object that localizes the object in space and thereby containing (enclosing) it, *carrier* – carries the object, *adapter* – adapts the instrument to affect the object,

[1] For a detailed description of the new version of TAPAZ-algebra, see: A. Hardzei, Theory for Automatic Generation of Knowledge Architecture: TAPAZ-2. Rev. English edn. Minsk: RIHE, 2017, 50 p. Available: http://tapaz.by.

[2] To date, each of the 32 TAPAZ-algebra role formulas have been deciphered. M. I. Svyatoshchik provided all possible assistance in the interpretation of some formulas of the TAPAZ Role List [27].

acceptor – catches the object, *stock* – the object collected for processing, *separator* – sorts the object, *material* – the object used as a raw material for making a product, *model* – the physical or informational original sample for making a product from the object, *retainer* – turns a variable locus of the object into a constant one, *resource* – feeds the instrument, *stimulus* – reveals the parameter of the object, *regulator* – serves as an instruction in making a product from the object, *chronotope* – localizes the object in time, *source* – provides instructions for the instrument, *indicator* – displays a parameter of impact on the object or a parameter of the product as the result of subject's impact on the object.

The algorithm for extracting specialized terminology from the Internet content of selected subject domain and constructing TAPAZ-units assumes answers to the key questions:

Who? With which tool? In relation to whom / what? In what place? Arriving on what? Adjusting by what? Accepting by what? Stocked up (on) what? Selecting by what? Making of what? Following what example? Fixing by what? Spending what? Stimulating with what? Guided by what? In what period? Knowing wherefrom? At what parameter? Affecting who / what? Produces whom / what?

To facilitate the work of experts in the construction of TAPAZ-units on the basis of the updated TAPAZ Role List for Chinese, English and Russian sectors of IT-Industry, a new ExpertTool version 1.0.0.0 was developed by the efforts of software engineer A. A. Matsko (see Fig. 2).

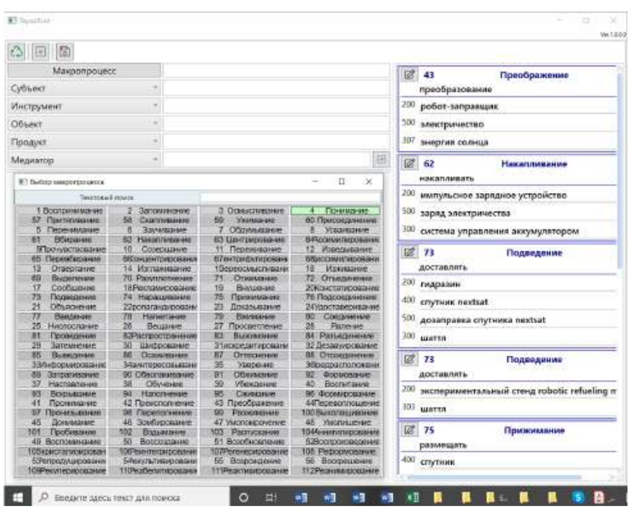

Figure 2. The working window of the software tool with an expanded tab.

There must be a one-to-one correspondence not only between noun phrases and thematic roles, as N. Chomsky mentioned [28], but a one-to-one correspondence between roles of individs in the event, parts of the sentence, mapping this event [1], and parts of the language [2], playing relevant roles in this sentence, otherwise, we will not be able to implement machine learning algorithms, such as, for example, an artificial neural network or a random forest and thereby provide automatic semantic markup of texts collected in the Knowledge Base. TAPAZ Semantic Dictionary consists of subject domains' TAPAZ-units assembled in the order indicated by TAPAZ Knowledge Graph. TAPAZ-units simultaneously form the taxonomy of the Intellectual Knowledge Base. Initially, TAPAZ-units are manually assembled by experts until the training data is sufficient to implement machine learning algorithms.

IV. Examples of Semantic Markup

Media report: *Yesterday at 10:30 am the Belarusian spacecraft was launched from the Baikonur cosmodrome.*

Text preprocessing by an expert using the TAPAZ software tool: *Specialists of the Federal State Institution "Roskosmos" on July 22, 2012 at 10:30 am from the Baikonur cosmodrome using the "Soyuz" carrier rocket into the near-earth orbit delivered the Belarusian spacecraft;* specified remote sensing process – *deliver,* TAPAZ macroprocess – *(73) approach,* subject – *the Federal State Institution "Roskosmos",* instrument – *the "Soyuz" carrier rocket,* object – *the Belarusian spacecraft,* locus – *the Baikonur cosmodrome*, landmark – *the near-earth orbit,* chronotope – *July 22, 2012; 10:30 am.*

Technical description: *The Belarusian spacecraft is similar to the Russian one.*

Stable expression *the Belarusian spacecraft* plays the role of a grammatical subject, but at the same time it is not mapped into a subject, because the Belarusian spacecraft does nothing with the Russian spacecraft. Correct semantic reconstruction of the role structure of the sentence: *What does someone do with information about the Belarusian spacecraft? Compares it with information about the Russian spacecraft and states the identity;* TAPAZ macroprocess – *(20) state,* subject – virtual *someone*, object – *information about the Belarusian spacecraft*, landmark – *information about the Russian spacecraft,* product – *coincidence of information.*

Let us show the solution to the problem of attributive 的 *de* when parsing of a Chinese sentence by means of TAPAZ technology: 皮球是红的 *Píqiú shì hóngde (The ball is red)*. The fact is that the Chinese sign 红 *hóng* (red) in the language system denotes a property, and in the sentence plays a specialized role of a grammatical attribute, therefore it does not require 的 *de*, which

[1]Syntactical rules for dynamic syntactic analysis module with a recursive reconstruction algorithm (parser) for sentence string is described in [29, 30].

[2]For definitions of Parts of Language, their paradigm, and semantic delineation procedures, see: [31, 32].

transforms nominal semantics into attributive, for example: 石头 *shítou (a stone)* → 石头的 *shítoude (stony)*. The presence of 的 *de* in the sentence 皮球是红的 indicates the omission of the grammatical direct object 东西 *dōngxi (thing)*, i.e. ***the ball is a red object, not the redness***. Indeed, from the classical logic point of view, identity should be established between homogeneous objects: an individ and another individ, or a feature of an individ and another feature of an individ. In this case, the Chinese language turns out to be more accurate in representing the world model than English or Russian. Correct parsing of the sentence: 皮球 *píqiú (the ball)* is a grammatical subject, 是 *shì (to be)* is a grammatical predicate, 红的 *hóngde (red)* is a grammatical attribute, 东西 *dōngxi (thing)* is a reconstructed grammatical direct object. However, the role structure of the event will be different:

What does someone do with the color information about the ball? Compares it with information about the red color and states their identity; TAPAZ macroprocess – *(20) state,* subject – virtual *someone,* object – *color information about the ball,* landmark – *information about the red color,* product – *coincidence of information.*

Due to the fact that semantic primitives lie at the core of the language system, one should achieve the minimum depth of recursive reconstruction, observing the strict requirements of **order, clarity and simplicity,** when the missing parts of a sentence are restored at the first or, at most, at the second step, because the deep recursion and the complex reconstruction create a high probability of error.

Another example: 苹果多少钱两斤？*Píngguǒ duōshao qián liǎng jīn? How much is a kilogram of apples?* → Recursive reconstruction: 苹果[是]多少钱，[少要买]两斤[苹果]。*Píngguǒ [shì] duōshao qián, [wǒ yào mǎi] liǎng jīn [píngguǒ].* Word for word translation: *Apples [are] how much money, [I need to buy] one kilogram [apples].* 苹果 *píngguǒ (apples)* is the 1st grammatical subject, 是 *shì (are)* is the reconstructed 1st grammatical predicate, 钱 *qián (money)* is the 1st grammatical direct object, 多少 *duōshao (how much)* is the 1st grammatical quantitative attribute of the 1st grammatical direct object; 少 *wǒ (I)* is the reconstructed 2nd grammatical subject, 要 *yào (need)* is the reconstructed modal component of the 2nd grammatical predicate, 买 *mǎi (buy)* is the reconstructed main component of the 2nd grammatical predicate, 两斤 *liǎng jīn (one kilogram)* is the 2nd quantitative attribute of the reconstructed 2nd grammatical direct object, which is 苹果 *píngguǒ (apples)*.

To determine the role structure of the event, it is necessary to transform the interrogative sentence into a narrative one: *The buyer asks for information about the cost of one kilogram of apples in order to compare it with information about the amount of money he has, if the information coincides, the buyer will buy apples;* 1st TAPAZ macroprocess – *(1) perceive,* object – *information about the cost of one kilogram of apples,* source – *product price of a seller;* 2nd TAPAZ macroprocess – *(24) certify,* object – *information about the cost of one kilogram of apples,* landmark – *the amount of money a buyer has;* 3d TAPAZ macroprocess – *(60) attain,* subject – *buyer,* object – *one kilogram of apples.*

We note that the power of the TAPAZ semantic markup, only in terms of the typical roles of individs, not even talking about the TAPAZ-algebra and generated by it Paradigm of Actions and the Knowledge Graph, almost 5 times exceeds the power of the closest analogue – the technology of Active Vocabulary [33], standardized and adopted by W3C Consortium in 2017 within the framework of Semantic Web project [34] and then Schema.org [35]. This technology is predominantly based on the theory of semantic cases of Fillmore and Jackendoff's early work, that we mentioned above, the inventory of which is:

"*Agent* – the initiator of some action, capable of acting with volition, and *actor* – supertype of agent which performs, effects, instigates, or controls the situation denoted by the predicate;

patient – the entity undergoing the effect of some action, often undergoing some change of state;

theme – the entity which is moved by an action, or whose location is described;

beneficiary – the entity for whose benefit the action was performed;

experiencer – the entity which is aware of the action or state described by the predicate but which is not in control of the action or state;

percept or *stimulus* – the entity which is perceived or experienced;

instrument – the means by which an action is performed or something comes about;

source – the entity from which something moves, either literally or metaphorically;

goal – the entity towards which something moves, either literally or metaphorically, and *recipient* – subtype of goal involved in actions describing changes of possession;

location – the place in which something is situated or takes place" [36].

It is not difficult to see that *experiencer, source, percept* or *stimulus, goal* and *recipient,* in fact, represent the same typical role of *landmark* in TAPAZ–2; *patient* and *theme* – the role of *object; agent* and *actor* – the roles of *subject* and *creator; beneficiary* – the role of *mediator; percept* or *stimulus* – the role of *source*; *location* – the role of *locus;* the role of *instrument* in both theories is almost the same, if one does not take into account the varieties of the instrument in TAPAZ–2. There is no any algebra in the substantiation of Fillmore's "case frames" or Stowell's "theta-grids", all these semantic categories were empirically distinguished, so it is impossible to establish their consistency, independence and completeness, thereby avoid the Russell's paradox, which inevitably arises from a mixture of theory and metatheory, language and metalanguage, semantics and metasemantics. It is for these reasons that the developers of the Semantic Web, despite titanic efforts to standardize technology, have so far failed to reduce various subject ontologies to a top-level ontology, which, as many commentators emphasize, is "critical to the whole concept" [37: 94]. This was partially acknowledged in 2006 by T. Berners-Lee himself in a joint article "Semantic Web Revisited" with N. Shedbolt and W. Hall: "The Semantic Web is a Web of actionable information – information derived from data through a semantic theory for interpreting the symbols. The semantic theory provides an account of "meaning" in which the logical connection of terms establishes interoperability between systems. This was not a new vision. Tim Berners-Lee articulated it at the very first World Wide Web Conference in 1994. This simple idea, however, remains largely unrealized" [38: 96].

TAPAZ Semantic Classifier is just such a top-level ontology. It includes the Ordered Set of Macroprocesses as Semantic Primitives (Paradigm of Actions), Role List of Individs and TAPAZ Knowledge Graph.

The Paradigm of Actions consists of informational and physical macroprocesses ordered by TAPAZ-algebra. The physical macroprocesses are shaded (see Fig. 3). Note separately that the construction of the TAPAZ Universal Problem Solver [39–41] is carried out using the TAPAZ-algebra and the TAPAZ Semantic Classifier, that is by combinatory methods, and not statistical, since all statistical methods, including artificial neural networks, only imitate the intellectual or inventive human activity, guessing the correct solutions with more or less degree of reliability, but in the fact that neural networks are able to effectively scale the solutions found by combinatory methods – we have no doubts. Moreover, it was precisely with the advent of deep learning algorithms for multilayer neural networks proposed by Geoffrey Hinton in 2007 [42, 43] that it became possible with the help of only one scientific laboratory to solve such large-scale tasks as compiling vast collections of texts of various subject domains and operating with big data, whereas before this required transnational scientific conglomerations and global interstate associations.

		I	II	III	IV
A	*a*	1 perceive	2 reflect	3 comprehend	4 understand
		attract 57	cumulate 58	constrict 59	attain 60
	b	5 adopt	6 memorize	7 contemplate	8 learn
		absorb 61	accumulate 62	center 63	assimilate 64
	c	9 feel	10 behold	11 feel profoundly	12 experience
		over absorb 65	concentrate 66	centrifuge 67	dissimilate 68
	d	13 reject	14 erase	15 rethink	16 overcome
		expel 69	decompress 70	force off 71	disassociate 72
B	*a*	17 notify	18 advertise	19 instill	20 state
		approach 73	joint 74	press down 75	connect 76
	b	21 explain	22 propagandize	23 prove	24 certify
		insert 77	pump 78	press in 79	link 80
	c	25 reveal	26 prophesize	27 enlighten	28 divine
		conduct 81	spread 82	squeeze out 83	disconnect 84
	d	29 darken	30 encode	31 discredit	32 disavow
		take out 85	pull up 86	push out 87	unlink 88
C	*a*	33 inform	34 interest	35 assure	36 predispose
		touch on 89	envelope 90	clamp 91	mold 92
	b	37 admonish	38 teach	39 convince	40 nurture
		rip up 93	fill up 94	press 95	form 96
	c	41 pierce	42 intend	43 transfigure	44 reincarnate
		penetrate 97	overflow 98	unclamp 99	eviscerate 100
	d	45 pester	46 mesmerize	47 lose conscious	48 go mad
		punch 101	uplift 102	disband 103	annihilate 104
D	*a*	49 recollect	50 recreate	51 restart	52 render
		recrystallize 105	reintegrate 106	regenerate 107	restore 108
	b	53 reproduce	54 reclaim	55 renew	56 revive
		recuperate 109	rehabilitate 110	reactivate 111	reanimate 112

Figure 3. Paradigm of Macroprocesses (Actions).
where: *A* – activation group, *B* – exploitation group, *C* – transformation group, *D* – normalization group; *a* – surroundings-shell subgroup, *b* – shell-core subgroup, *c* – core-shell subgroup, *d* – shell-surroundings subgroup; *I* – initiation raw, *II* – accumulation raw, *III* – amplification raw, *IV* – generation raw.

V. Conclusion

The TAPAZ technology offers a search by event fragments or technological cycles, which are described by special TAPAZ-units, which are macroprocesses [1] in the assembly, when specialized subject domain processes are algorithmically correspond to TAPAZ macroprocesses and the roles of all participants in the events are algorithmically calculated [2].

This approach provides maximum accuracy and speed of search, relevance of search results and simultaneously solves the problem of automatically identifying the semantic equivalence of text documents and borrowing scientific ideas in order to curb the spread of plagiarism and prevent clogging the information space under the conditions of its globalization. In addition, it allows you to find similar technological cycles in close (adjacent) and distant subject domains, thereby providing support to the user in analytical activities, which greatly expands the functionality of the search engine, shifting it towards inventive level.

Judging by the rapid development since 2011 of activity-based technology for the international public resource Schema.org by Google, Microsoft, Yahoo and Yandex, as well as since 2017 – the Activity Vocabulary by the W3C Consortium for the Semantic Web, in the next 10–15 years, the main efforts of international scientific and financial centers will be focused on the

[1]We emphasize that macroprocess is one of 112 extremely abstract processes that are isomorphic to any subject domain and are calculated and encoded by the TAPAZ-algebra.

[2]"There are such concepts as "culprit", "tool", "product of labor" ⟨. . .⟩ We are here in the field of various categories, apparently ontological, but essentially semantic" [44: 11].

creation of knowledge graphs for automatic extraction of semantically relevant information from search pages, in other words, on the stage-by-stage development of a language capable for representing and transforming information in a readable machine form. Such a language should describe both the data that exists in any branch of knowledge (subject domain), and the rules for reasoning about this data, as well as the rules for displaying data on the Internet and back. The transition to the seventh technological order depends on this.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 284, pp. 34–43, 2001.

[2] C. J. Fillmore, *"The Case for Case". Universals in Linguistic Theory*. New York : Holt, Rinehart and Winston, 1968.

[3] ——, "Frame semantics and the nature of language," in *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, vol. 280, 1976, pp. 20–32.

[4] ——, "Frame semantics," in *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co., 1982, pp. 111–137.

[5] C. J. Fillmore and S. Atkins, "Starting where the dictionaries stop: The challenge for computational lexicography," in *Computational Approaches to the Lexicon*. Oxford: Oxford University Press, 1994, pp. 349–393.

[6] R. S. Jackendoff, *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press, 1972.

[7] ——, *X-bar Syntax: A Study of Phrase Structure*. Cambridge, Mass.: MIT Press, 1977.

[8] T. A. Stowell, "Origins of phrase structure," in *Thesis (Ph.D.)*, Massachusetts Institute of Technology. Dept. of Linguistics and Philosophy., 1981, p. 496, Available: http://www.ai.mit.edu/projects/dm/theses/stowell81.pdf.

[9] R. Montague, in *Formal Philosophy: Selected Papers of Richard Montague*. New Haven, London: Yale University Press, 1974, pp. 222–247.

[10] G. Lakoff, "Instrumental Adverbs and the Concept of Deep Structure," *Foundations of Language*, vol. 4, no. 1, pp. 4–29, 1968.

[11] L. Tesnière, *Éléments De Syntaxe Structurale*. Paris: Librairie C. Klincksieck, 1959, english translation: L. Tesnière, Elements of structural syntax. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2015, 698 p.

[12] V. Martynov, *Semiologicheskie osnovy informatiki [Semiological foundations of computer science]*. Minsk: Nauka i tekhnika [Science and technics], 1966.

[13] Ashtadhyayi or Sutrapath of Panini [Electronic resource]. Mode of access: https://sanskritdocuments.org/doc_z_misc_major_works/aShTAdhyAyI.pdf. — Date of access: 12.06.2021.

[14] V. N. Toporov, "On some analogies to the problems and methods of modern linguistics in the works of ancient indian grammarians," *Brief communications of the Institute of Asian Peoples*, pp. 123–133, 1961.

[15] O. A. Voloshina, "On the structure and linguistic terminology of panini grammar and its influence on european linguistics," *Bulletin of the Russian State University for the Humanities, Series: History. Philology. Culturology. Oriental studies*, vol. 9, no. 52, pp. 161–172, 2010.

[16] V. Martynov, *Semiologicheskie osnovy informatiki [Semiology foundations of informatics]*. Minsk: Nauka i tekhnika [Science and technics], 1974.

[17] ——, *Universal'nyj Semanticheskij Kod. Grammatika. Slovar', Teksty [Universal Semantic Code. Grammar. Dictionary, Texts]*. Minsk: Nauka i Tekhnika [Science and Technics], 1977.

[18] ——, *Universal'nyj Semanticheskij Kod. USK-3 [Universal Semantic Code: USC–3]*. Minsk: Nauka i Tekhnika [Science and Technics], 1984.

[19] ——, *Universal'nyj Semanticheskij Kod. USK-4 [Universal Semantic Code: USC–4]*. Institute of Linguistics of Academy of Science of BSSR, Minsk: Nauka i Tekhnika [Science and Technics], 1988, no. 2, Preprint.

[20] ——, *Universal'nyj Semanticheskij Kod. USK-5 [Universal Semantic Code: USC–5]*. Minsk State Linguistic University, Minsk: Izdatel'stvo Minskogo Gosudarstvennogo Lingvisticheskogo Universiteta [Minsk State Linguistic University Publ.], 1995, no. 4, Preprint.

[21] ——, *Osnovy Semanticheskogo Kodirovaniya. Opyt Predstavleniya i Preobrazovaniya Znanij [Foundations of Semantic Coding. Experience of Knowledge Representation and Conversion]*. Minsk: Izdatel'stvo Evropejskogo gumanitarnogo universiteta [European Humanitarian University Publ.], 2001.

[22] A. Hardzei, "Procedural Semantics and Calculus of Subject Domains," in *CONFERENCE 1994, Language: Semantics, Syntactics, Pragmatics*, M. Leschenko, Ed. Minsk: Minsk State Linguistic University Publ., 1995, pp. 16–17, pt. 1.

[23] ——, *The Deductive Theory of Language*. Minsk: Belarus Science, 1998.

[24] ——, *The Principles of Calculating the Semantics of Subject Domains*. Minsk: Belarusian State University Publ., 1998.

[25] Z. Harris, *String Analysis of Sentence Structure*. The Hague: Mouton, 1962.

[26] ——, "Co-occurrence and transformation in linguistic structure," *Language*, vol. 33, no. 3, pp. 283–340, 1957.

[27] M. I. Svyatoshchik, "Translating Multi-Object Semantics into Single-Object Semantics as a Method of Minimizing Semantic Calculation," in *Uchonyye zapiski VGU im. P. M. Masherova: sb. nauch. trudov [Scientific notes of Vitebsk State University named after P. M. Masherov: collection of scientific papers]*. Vitebsk: Vitebsk State University Publ., 2020, vol. 31, pp. 180–184.

[28] N. Chomsky, *"Lectures on government and binding: The Pisa lectures". Studies in generative grammar*. New Jersey : Foris Publications, Dordrecht and Cinnaminson, 1981, no. 9.

[29] A. Hardzei, "Virtual string as a syntactic code of a sentence (on the example of the Chinese language)," in *CONFERENCE 2007, Language, society and problems of intercultural communication*, L. M. Sereda, S. V. Grinev-Grinevich, and I. F. Uhvanova-Shmygova, Eds. Grodno: Grodno State University Publ., 2007, pp. 349–358, pt. 2.

[30] ——, "Dynamic Syntax: A Semantic View," *Foreign Languages in Tertiary Education*, vol. 4, no. 43, pp. 26–34, 2017.

[31] ——, "The Paradigm of Parts of Language," in *CONFERENCE 2003, Word Formation and Nominative Derivation in Slavic Languages*, S. Emelyanova, L. Rychkova, and A. Nikitevich, Eds. Grodno: Grodno State University Publ., 2003, pp. 173–179.

[32] ——, "Parts of Language and the Procedures of Its Delineation," in *The Paths of the Middle Kingdom*, A. Hardzei and W. Hongbin, Eds. Minsk: Belarusian State University Publ., 2006, ch. 1, pp. 69–75, iss. 1.

[33] Activity Vocabulary. W3C Recommendation [Electronic resource]. Mode of access: https://www.w3.org/TR/activitystreams-vocabulary/#dfn-activity. — Date of access: 12.06.2021.

[34] Semantic Web [Electronic resource]. Mode of access: https://www.w3.org/standards/semanticweb. — Date of access: 12.06.2021.

[35] Schema.org Vocabulary, version 8.0 [Electronic resource]. Mode of access: https://schema.org. — Date of access: 12.06.2021.

[36] B. Aarts, in *English Syntax and Argumentation*, 2nd ed. Houndmills, Basingstoke, Hampshire, New York: Palgrave (Macmillan), 2001, pp. 94–95.

[37] S. P. Rippa and O. M. Lyashenko, "Semantic platforms of knowledge bases software in informatics," in *Pratsi Odes'koho politekhnichnoho universytetu [Proceedings of Odessa Polytechnic University]*, vol. 1, no. 40. Odessa: Odessa Polytechnic University Publ., 2013, pp. 91–96.

[38] T. Berners-Lee, N. Shedbolt, and W. Hall, "Semantic Web Revisited," *IEEE Intelligent Systems*, vol. 21, pp. 96–101, 2006, iss. 3.

[39] A. Hardzei, *Theory for Automatic Generation of Knowledge Architecture: TAPAZ-2*, Rev. English ed. Minsk: The Republican Institute of Higher School Publ., 2017, Available: http://tapaz.by.

[40] ——, "Plagiarism Problem Solving Based on Combinatory Semantics," *Communications in Computer and Information Science (CCIS)*, vol. 1282, pp. 176–197, 2020, Available: https://link.springer.com/book/10.1007%2F978-3-030-60447-9.

[41] ——, "Semantic Markup of the Event and its Display by Means of the Chinese and Russian Languages," *Foreign Languages in Tertiary Education*, vol. 2, no. 57, pp. 5–26, 2021.

[42] G. E. Hinton. "Boltzmann machine". Scholarpedia [Electronic resource]. Mode of access: http://www.scholarpedia.org/article/Boltzmann_machine. — Date of access: 12.06.2021.

[43] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[44] T. Kotarbinski, *Praxiology: An Introduction to the Sciences of Efficient Action*. Oxford: Pergamon Press, 1965.

Семантический подход к решению проблемы обработки данных на естественном языке

А. Н. Гордей

На основе третьей редакции второй версии Теории автоматического порождения архитектуры знаний (ТАПАЗ-2) предложен новый подход к семантической разметке события и формализации синтаксиса китайских, английских и русских предложений.

Ключевые слова: комбина́торная семантика, сематическая разметка, семантический падеж, семантический классификатор, граф знаний, ролевой лист индивидов, субъект, объект, акция, макропроцесс, специализированный процесс, модель мира, ТАПАЗ-алгебра, ТАПАЗ-юнит.

Received 15.06.2021