

The logic and linguistic model for automatic extraction of collocation similarity

N. Khairova¹, S. Petrasova², Ajit Pratap Singh Gautam³

National Technical University "Kharkiv Polytechnic Institute"; e-mail: nina_khajrova@yahoo.com

Received November 15 2015; accepted November 25 2015

Abstract. The article discusses the process of automatic identification of collocation similarity. The semantic analysis is one of the most advanced as well as the most difficult NLP task. The main problem of semantic processing is the determination of polysemy and synonymy of linguistic units. In addition, the task becomes complicated in case of word collocations. The paper suggests a logical and linguistic model for automatic determining semantic similarity between collocations in Ukraine and English languages. The proposed model formalizes semantic equivalence of collocations by means of semantic and grammatical characteristics of collocates. The basic idea of this approach is that morphological, syntactic and semantic characteristics of lexical units are to be taken into account for the identification of collocation similarity. Basic mathematical means of our model are logical-algebraic equations of the finite predicates algebra. Verb-noun and noun-adjective collocations in Ukrainian and English languages consist of words belonged to main parts of speech. These collocations are examined in the model. The model allows extracting semantically equivalent collocations from semi-structured and non-structured texts. Implementations of the model will allow to automatically recognize semantically equivalent collocations. Usage of the model allows increasing the effectiveness of natural language processing tasks such as information extraction, ontology generation, sentiment analysis and some others.

Key words: automatic extraction, identification of collocation similarity, finite predicates algebra, logical-algebraic equations, grammatical and semantic features.

INTRODUCTION

This is a particularly exciting time to be working on computer linguistic or natural language processing. Nowadays linguistic technologies have become not only tools for modelling language but also a production factor. Computer linguistics is now one of the most strongly developing directions of information technologies. In fact, almost every intelligent information system with a user interface, both text and web-content processing systems, uses linguistic technologies [1].

The vast amount of textual data on the Web and social media has made it possible to build lots of new and interesting applications.

Important tasks of computer linguistic include: Information extraction (IE), Sentiment analysis, Machine translation, Information retrieval, Ontology generation and some others.

One important task of natural language processing is information extraction. IE is the task of automatically extracting structured information from unstructured and/or semi-structured textual information. In fact, the task of IE is to identify instances of a particular prespecified class of entities, relationships and events in natural language texts, and the extraction of the relevant properties of the identified entities, relationships or events [2].

Another application of this kind of IE, involves sentiment analysis. Sentiment analysis (also known as opinion mining) refers to the use of NLP to identify and extract subjective information in texts. This can be used for lot of tasks [3]. For example:

- such information can become an additional powerful source for predicting the expected stock market changes;
- such information can become a source to predict election outcomes;
- such information can help a corporation to determine what people think about some (new) products;
- such information can help politics to determine what people think about candidates or issues;
- and many others tasks.

Another task of computer linguistics that is very important nowadays is ontology generation [4]. Ontology generation (aka ontology acquisition) is the automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between those concepts from a corpus of natural language text [5]. This task typically involves:

- technology for automated concepts extraction using linguistic processor [6]
- and extraction of the semantic relations between concepts using linguistic processor.

Another modern application of NLP is to identify text clones, for example, in the technical documentation [7]. A text clone means a block of text that is repeated in various degrees of similarity across the documentation [8].

A number of things make natural language understanding difficult. These are problems with ambiguity, idioms, segmentation of words and sentences, non-standard language that we frequently see in texts of

Twitter, SMS, blog, social media and others. And of course we also have a lot of problems with entity names, synonymy and co-reference.

The above-mentioned problems include challenge of collocation extraction and semantic equivalence recognition. Solving this problem applied to tasks of Automatically Ontology Generation, Sentiment Analysis and Information Extraction is still quite hard.

THE ANALYSIS OF RECENT RESEARCHES AND PUBLICATIONS

The notion of a collocation differs across linguistic traditions. For instance, a collocation is a recurrent word combination [9]. By contrast, a collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components [10].

In this study, a collocation is considered as a combination of two lexical units that co-occur in the text non-randomly. The available variety of collocations extraction methods can be divided into two groups.

The methods from the first group are statistical methods. Statistical measures have become extremely widespread in modern linguistic research. These measures are based on co-occurrence frequencies of word pairs and frequencies of each constituent [11].

The *window-based methods* rely on a linear word order model, in which the collocation candidates are extracted from a fixed-size window [12].

Mutual information (MI) and *Pointwise mutual information (PMI)* measures are used to determine the significance of the occurrence of two words by comparing the frequency of their co-occurrence with the product of frequencies of their independent occurrence in the text [13].

The *T-score measure* takes into account the frequency of co-occurrence of a keyword and its collocate. Words with the highest T-score occur frequently, so we must set a list of stop words to reject the most frequent words.

The *Chi-squared distribution* uses the Pearson χ^2 -test to evaluate how likely it is that any observed difference between the sets arose by chance. The four values of a contingency table are:

- frequency of a collocation;
- frequency of a collocation with the first word (without the second one);
- frequency of a collocation with the second word (without the first one);
- frequency of all other collocations.

The drawbacks of statistical methods are extraction of noise and ignoring of syntactic correlations between words in long distances.

The methods from the second group are based on the analysis of the syntactic structure of collocations [14]. The analysis of the syntactic structure allows to filter out false collocates as well as to extract collocates located in a long distance from each other. It should be noted that this extended precision is achieved by a careful description of all possible syntactic constructions for two collocates.

It is also worth noting that methods of collocation extraction have become widely used in modern corpus linguistics [15].

Far fewer studies are aimed at solving the identification of collocation similarity problem [16, 17, 18].

OBJECTIVES

In the article we are focusing on the problems of collocation extraction and semantic equivalence recognition. Semantic equivalents can be defined as words with a similar meaning. The main aim is detecting that two collocations mean the same thing or the identification of collocation similarity.

Two-word phrases formed by pairs of semantic equivalents may be semantically similar (Fig.1)

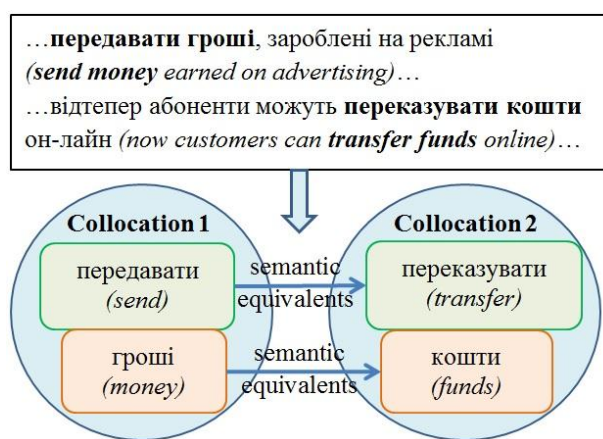


Fig. 1. Collocation similarity.

Collocations may be semantically dissimilar, even if they are collocates, which are equivalents (Fig.2).

The proposed logical-linguistic model formalizes semantic equivalence of collocations by means of semantic and grammatical characteristics of the collocates. The basic idea of this approach is that there is common content (meaning) between collocates that have semantic correlations. And this meaning expresses similarity of denoted concepts or phenomena. We consider verb-noun and noun-adjective collocations in Ukrainian and English languages. To formally express Collocation Similarity we use logical-algebraic equations of the finite predicates algebra.

...Класифікатор системи позначень **одиниць виміру** та обліку (*The classifier of the notation system of **units of measure and calculation***)...
 ...**Визначити величину** валового і чистого доходів (*To **determine the amount** of gross and net revenue*)...

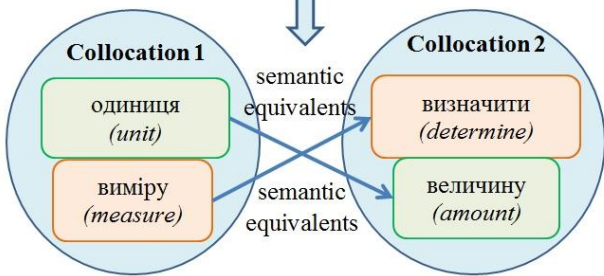


Fig.2. Collocation dissimilarity.

BASIC MEANS OF THE MODEL

Basic mathematical means of our model are logical-algebraic equations of the finite predicates algebra [19]. Let U be a universe of elements. The universe U contains various elements of the language system: lexemes, sentences, phrases, word-combinations, words etc. The universe is finite, as the sets of the elements are finite and determinate. The set $M = \{m_1, \dots, m_n\}$ is a subset of grammatical and semantic features of a collocate, and n is amount of system features. Predicates P_i are defined over the Cartesian products $M_1 \times M_2 \times \dots \times M_n$. They designate relations between grammatical and semantic features of collocates by formal tool of the finite predicates algebra [20]. Predicate $P(x) = I$, if the main word features of the collocation have a certain grammatical and semantic characteristics. Predicate $P(y) = I$, if the dependent word features of the collocation have a certain grammatical and semantic characteristics. And both predicates equal zero otherwise.

Variables x_1, x_2, \dots, x_n are called subject variables and their values are called subjects. The recognition predicate of the subject a by the subject variable x_i is the basic one for the algebra of predicates:

$$x_i^a = \begin{cases} 1, & \text{if } x_i = a \\ 0, & \text{if } x_i \neq a \end{cases} \quad (1) \quad (1 \leq i \leq n),$$

where $i = \{1, 2, \dots, n\}$, a is any of the universe elements.

MODELING OF COLLOCATION SIMILARITY IDENTIFICATION IN UKRAINIAN LANGUAGE

We can define a set of grammatical and semantic characteristics of collocates for Ukrainian language using two subject variables (1). The variable a defines grammatical categories of Ukrainian language:

$$a^{NNom} \vee a^{NGen} \vee a^{NAcc} \vee a^{NDat} \vee a^{Nln} \vee a^{NPr} \vee a^{ANom} \vee a^{AGen} \vee a^{AAcc} \vee a^{ADat} \vee a^{Aln} \vee a^{APr} \vee a^{VRef} \vee a^{VNonRef} = 1,$$

where: a^{NNom} is a noun, nominative case; a^{NGen} is a noun, genitive case; a^{NPr} is a noun, prepositional case; a^{ANom} is an adjective, nominative case; a^{AAcc} is an adjective,

accusative case, a^{ADat} is an adjective, dative case; a^{VRef} is a verb, reflexive; $a^{VNonRef}$ is a verb, non-reflexive.

The subject variable c defines semantic categories:

$$c^{Ag} \vee c^{Att} \vee c^{Pac} \vee c^{Adr} \vee c^{Ins} \vee c^M = 1,$$

where: c^{Ag} – an agent, c^{Att} – an attribute, c^{Pac} – an patient, c^{Adr} – an addressee, c^{Ins} – an instrument, c^M – a location or content.

As we mentioned above predicate $P(x)$ defines grammatical and semantic characteristics of the main word of collocations:

$$P(x) = a_x^{NNom} c_x^{AG} \vee a_x^{NGen} c_x^{Att} \vee a_x^{NAcc} c_x^{Pac} \vee a_x^{NDat} c_x^{Adr} \vee a_x^{Nln} c_x^{Ins} \vee a_x^{NPr} c_x^M \vee a_x^{VNonRef}. \quad (2)$$

Whereas predicate $P(y)$ defines grammatical and semantic characteristics of the dependent word of collocations:

$$P(y) = a_y^{NGen} c_y^{Att} \vee a_y^{NAcc} c_y^{Pac} \vee a_y^{NDat} c_y^{Adr} \vee a_y^{Nln} c_y^{Ins} \vee a_y^{NPr} c_y^M \vee a_y^{ANom} \vee a_y^{AGen} \vee a_y^{AAcc} \vee a_y^{ADat} \vee a_y^{Aln} \vee a_y^{APr}. \quad (3)$$

Double predicate $P(x, y)$ describes a combination of semantic and grammatical information of words in two-word collocations:

$$P(x, y) = (a_y^{ANom} \vee a_y^{AGen} \vee a_y^{AAcc} \vee a_y^{ADat} \vee a_y^{Aln} \vee a_y^{APr}) (a_x^{NNom} c_x^{AG} \vee a_x^{NGen} c_x^{Att} \vee a_x^{NAcc} c_x^{Pac} \vee a_x^{NDat} c_x^{Adr} \vee a_x^{Nln} c_x^{Ins} \vee a_x^{NPr} c_x^M) \vee a_x^{VNonRef} a_y^{NAcc} c_y^{Pac} \vee a_x^{NNom} c_x^{AG} a_y^{NGen} c_y^{Att}. \quad (4)$$

The predicate equals unity, if the both words that have a certain grammatical and semantic features form a collocation. And predicate equal zero otherwise. For example, the last conjunction of the predicate describes the semantic and grammatical characteristics of the following collocations:

- мова_x ^a NNom c_x Ag розмітки_y ^a NGen c_y Att (*a markup language*);
- період_x ^a NNom c_x Ag користування_y ^a NGen c_y Att (*a usage period*).

A predicate of semantic equivalence can be defined between collocations. The ratio of semantic equivalence of two two-word collocations can be defined as:

$$P(x_1, y_1) * P(x_2, y_2) = \gamma_i(x_1, y_1, x_2, y_2) \bullet P(x_1, y_1) \bullet P(x_2, y_2), \quad (5)$$

where: $*$ indicates semantic similarity, \bullet defines the Cartesian product, $\gamma_i(x_1, y_1, x_2, y_2)$ predicate eliminates collocations between which semantic equivalence cannot be identified.

For example predicate γ_1 defines the semantic similarity between the collocations:

“to store data” and to “keep indicators”

semantic equivalence of two two-word collocations can

$$\begin{aligned}
 & \text{“a wireless device”} \approx \text{“a cordless machine”} \approx \text{“a} & (11) \\
 & \text{device is wireless”}; \\
 & \text{“a standard is created”} \approx \text{“the criterion is} \\
 & \text{formed”}.
 \end{aligned}$$

Predicate γ_2 :

$$\begin{aligned}
 \gamma_2(x_1, y_1, x_2, y_2) = & a_{x1}^{NSub} c_{x1}^{Ag} a_{y1}^{NObj} c_{y1}^{Att} \vee \\
 & \vee a_{x2}^{NObj} c_{x2}^{Att} a_{y2}^{NSub} c_{y2}^{Ag}
 \end{aligned}$$

be defined as equation (5). The predicate γ_i eliminates collocations between which semantic equivalence cannot be identified in the equation.

In English, collocations γ_i can be identified as:

$$\begin{aligned}
 & \gamma_1(x_1, y_1, x_2, y_2) = \\
 & = a_{y1}^{AAtt} a_{x1}^{NSub} c_{x1}^{Ag} \vee a_{x2}^{NSub} c_{x2}^{Ag} a_{y2}^{APr}. & (10)
 \end{aligned}$$

Predicate γ_1 shows, for example, the semantic similarity between the following collocations:

shows, for example, the semantic similarity between the following collocations:

$$\begin{aligned}
 & \text{the usage of data} \approx \text{the application of information} \approx \\
 & \approx \text{the data usage}; \\
 & \text{a content provider} \approx \text{a maintenance supplier}.
 \end{aligned}$$

Predicate γ_3 :

$$\begin{aligned}
 & \gamma_3(x_1, y_1, x_2, y_2) = \\
 & = a_{x1}^{NSub} c_{x1}^{Ag} a_{y1}^{VTt} \vee a_{x2}^{NSub} c_{x2}^{Ag} a_{y2}^{VIntr} & (12)
 \end{aligned}$$

shows the semantic similarity between the following collocations:

$$\begin{aligned}
 & \text{an equipment detects} \approx \text{an appliance finds}; \\
 & \text{broadcast happened} \approx \text{transmission occurred}.
 \end{aligned}$$

Predicate γ_4 :

$$\begin{aligned}
 & \gamma_4(x_1, y_1, x_2, y_2) = \\
 & = a_{x1}^{VTt} a_{y1}^{NObj} c_{y1}^{Pac} \vee a_{x2}^{VTt} a_{y2}^{NObj} c_{y2}^{Pac} & (13)
 \end{aligned}$$

shows the semantic similarity between the collocations:

$$\text{provide aid} \approx \text{give support}.$$

Thus, a predicate of semantic equivalence between collocations consisted of semantically equivalent pairs of collocates in the English language can be defined as:

$$\begin{aligned}
 & \gamma(x_1, y_1, x_2, y_2) = \\
 & = (a_{y1}^{AAtt} a_{x1}^{NSub} c_{x1}^{Ag} \vee a_{x1}^{NSub} c_{x1}^{Ag} a_{y1}^{APr}) \cdot \\
 & \cdot (a_{y2}^{AAtt} a_{x2}^{NSub} c_{x2}^{Ag} \vee a_{x2}^{NSub} c_{x2}^{Ag} a_{y2}^{APr}) \vee \\
 & \vee (a_{x1}^{NSubOf} c_{x1}^{Ag} a_{y1}^{NObj} c_{y1}^{Att} \vee \\
 & \vee a_{x1}^{NObj} c_{x1}^{Att} a_{y1}^{NSub} c_{y1}^{Ag}) (a_{x2}^{NSubOf} c_{x2}^{Ag} \cdot \\
 & \cdot a_{y2}^{NObj} c_{y2}^{Att} \vee a_{x1}^{NObj} c_{x2}^{Att} a_{y2}^{NSub} c_{y2}^{Ag}) \vee \\
 & \vee a_{x1}^{NSub} c_{x1}^{Ag} (a_{y1}^{VTt} \vee a_{y1}^{VIntr}) \vee \\
 & \vee a_{x2}^{NSub} c_{x2}^{Ag} (a_{y2}^{VTt} \vee a_{y2}^{VIntr}) \vee \\
 & \vee a_{x1}^{VTt} a_{y1}^{NObj} c_{y1}^{Pac} a_{x2}^{VTt} a_{y2}^{NObj} c_{y2}^{Pac}, & (14)
 \end{aligned}$$

where: $a_x^{NSub} c_x^{Ag}$ is a normalized form of the subject variable a_x^N . Since the subject variable c_x does not influence semantic equivalence between collocates in such predicates as γ_1 and γ_2 , we can neglect it for nouns a_x , which are main words in collocates.

As a result, the predicates of collocations that satisfy these characteristics will be equal unity. Otherwise, the predicate equals zero when two collocations are semantically dissimilar. For example, semantically dissimilar collocations:

$$\begin{aligned}
 & \text{a tale} a_{x1}^{NSub} c_{x1}^{Ag} \text{is checked} a_{y1}^{APr}, \\
 & \text{verify} a_{x2}^{VTt} \text{a story} a_{y2}^{NObj} c_{y2}^{Pac},
 \end{aligned}$$

where: words *a tale* and *a story* are similar, *to check* and *to verify* are similar too, though the collocations *a tale is checked* and *to verify a story* are dissimilar because it does not satisfy equation 14. Although according to 10 and 14 collocations *a tale is checked* and *a verified story* are similar:

$$\begin{aligned}
 & \text{a tale} a_{x1}^{NSub} c_{x1}^{Ag} \text{is checked} a_{y1}^{APr} \approx \\
 & \approx \text{a verified} a_{y2}^{AAtt} \text{story} a_{x2}^{NSub} c_{x2}^{Ag}.
 \end{aligned}$$

Let's take a look at another example:

$$\begin{aligned}
 & \text{a decision} a_{x1}^{NSub} c_{x1}^{Ag} \text{influences} a_{y1}^{VTt} \text{and} \\
 & \text{affect} a_{x2}^{VTt} \text{a resolution} a_{y2}^{NObj} c_{y2}^{Pac},
 \end{aligned}$$

where: words *a decision* and *a resolution* are similar, *to influence* and *to affect* are similar too, though the collocations *a decision influences* and *to affect a resolution* are dissimilar because it does not satisfy the equation 14.

Although according to 12 and 14 collocations *a decision influences* and *a resolution affects* are similar:

$$\begin{aligned}
 & \text{a decision} a_{x1}^{NSub} c_{x1}^{Ag} \text{influences} a_{y1}^{VTt} \approx \\
 & \approx \text{a resolution} a_{x2}^{NSub} c_{x2}^{Ag} \text{affects} a_{y2}^{VTt},
 \end{aligned}$$

and according to 13 and 14 collocations *influence a decision* and *affect a resolution* are similar too:

$$\begin{aligned}
 & \text{influence} a_{x1}^{VTt} \text{a decision} a_{y1}^{NObj} c_{y1}^{Pac} \approx \\
 & \approx \text{affect} a_{x2}^{VTt} \text{a resolution} a_{y2}^{NObj} c_{y2}^{Pac}.
 \end{aligned}$$

CONCLUSIONS

The main result of the study is the logical-linguistic model of collocation similarity for Ukrainian and English languages. The model allows extracting semantically equivalent collocations from semi-structured and non-structured texts in Ukrainian or in English. Implementation of the model will allow to automatically recognize semantically equivalent collocations. Usage of the model allows increasing the effectiveness of natural language processing tasks such as information extraction, ontology generation, sentiment analysis and some others.

In the future research we intend to broaden the scope of the study on semantic equivalence. This study has shown that the grammatical dependency of main and dependent words should be taken into account together

with their grammatical and semantic characteristics. The main challenge is to discover semantic similarity between NN and N^{of}N collocations.

REFERENCES

1. **N. Khairova, G. Shepelyov, S. Petrasova. 2014.** Evaluating effectiveness of linguistic technologies of knowledge identification in text collections. – Transactions on business and engineering intelligent applications. ITHEA. – Rzeszow – Sofia. 2014. – 71-75.
2. **Thierry Poibeau, Horacio Saggion, Roman Yangarber (Eds.). 2008.** Multilingual Information Extraction and Summarization Proceedings of MMIES-2: the Second Workshop on Multi-Lingual, Multi-Source Information Extraction and Summarization, at COLING-2008: the 22nd International Conference on Computational Linguistics.
3. **Bo Pang, Lillian Lee. 2008.** Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2 (1-2). 1–135.
4. **Y. Burov. 2014.** Business process modelling using ontological task models. *Econtechmod. An international quarterly journal*, Vol. 1, No. 1. 11–22.
5. **V. Lytvyn. 2013.** Design of intelligent decision support systems using ontological approach. *Econtechmod. An international quarterly journal*, Vol. 2, No. 1. 31–37.
6. **V. Lytvyn, O. Semotuyk, O. Moroz. 2013.** Definition of the semantic metrics on the basis of thesaurus of subject area. *Econtechmod. An international quarterly journal*, Vol. 2, No. 4. 47–51.
7. **M. Ericsson, A. Wingkvist, and W. Löwe. 2012.** Visualization of Text Clones in Technical Documentation. In *Proceedings of the Swedish Chapter of Eurographics (SIGRAD)*. 79-82.
8. **A. Wingkvist, M. Ericsson, and W. Löwe. 2011.** Making Sense of Technical Information Quality: A Software-based Approach. *Journal of Software Technology*, 4(3). 12-18.
9. **John Sinclair. 1991.** *Corpus, Concordance, Collocation*. Oxford University Press, 179.
10. **Christopher D. Manning, Hinrich Schütze. 1999.** *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, 680.
11. **V. Brezina, T. McEnery, S. Wattam. 2015.** Collocations in context: a new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20, 2. 139-173.
12. **W. Church, P. Hanks. 1990.** Word association norms, mutual information, and lexicography. – *Computational Linguistics*, 16(1). 22–29.
13. **S. Evert, B. Krenn. 2001.** Methods for the qualitative evaluation of lexical association measures – Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse. 188–195.
14. **S. Koshcheeva, V. Zakharov. 2014.** Comparing methods of automatic verb-noun collocation extraction. – *Computational Models for Business and Engineering Domains*. ITHEA. Rzeszow-Sofia. 158 – 171.
15. **S. Evert., 2008.** Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. 1212-1248.
16. **Muller P., Hathout N., Gaume B. 2006.** Synonym Extraction Using a Semantic Distance on a Dictionary. Workshop on TextGraphs, at HLT-NAACL. Association for Computational Linguistics. 65–72.
17. **Hua WU, Ming ZHOU. 2003.** Synonymous Collocation Extraction Using Translation Information. *Proceeding ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, V. 1. 120-127.
18. **D. Hindle, M. Rooth. 1993.** Structural ambiguity and lexical relations. *Association for Computational Linguistics*, 19, 1. 103-120.
19. **Bondarenko M., Shabanov-Kushnarenko J. 2007.** The intelligence theory. Kharkiv: “SMIT”, 576. (In Russian).
20. **N. Khairova, N. Sharonova. 2009.** Use of Predicate Categories for Modelling of Operation of the Semantic Analyzer of the Linguistic Processor. *Proceedings of IEEE EAST-West Design & Test Symposium (EWDTS'09)*. 204- 207.