UDC 004.519.7 (045)

Anastasia Vavilenkova

A SELF-SYSTEM TO IDENTIFY CONCEPTUAL RELATIONSHIPS IN A TEXT

National Aviation University
1, Kosmonavta Komarova avenue, Kyiv, 03680, Ukraine
E-mail: a_vavilenkova@mail.ru

Abstract. *The purpose of this study is to identify conceptual relationships in texts based on the use of logic and linguistic models of textual information. There are analyzed tools for building meaningful relationships within text documents. This analysis compares the creation of thematic progressions and a finite state transition network. This work demonstrates the disadvantages of the augmented finite state transition network and highlights the advantages of constructing logic and linguistic models of natural language sentences as the key operation mechanism of self-system. Identifying of conceptual relationships is possible by matching and replacing the predicate variables and constants.*

Keywords: concptual relationships; logic and linguistic model; natural language; transition network; semantic modling.

1. Introduction

Today, there is a public need for the development of effective language technologies, which are based on technology of knowledge operating. In particular, there is unsolved problem considering creation of modern automatic means of knowledge extraction from electronic texts based on constructing formal models and the problem of comparative analysis of electronic texts by content. Modern automated systems that analyze the semantic structure of the text at a level above sentences are at the experimental stage. Their main drawback – the fact that they work with the contents of individual words and phrases, at least – with relationships between keywords, but the structure of the sentence and the text as a whole is not analyzed. Detection of conceptual relationships in the text is based on the description of the explication of concepts in the text. Thus, by analyzing the frequency of words using, it is extracted a group of lexical units, occupying a strong position. Quantitative analysis is extended and supplemented by qualitative analysis of keywords and content. Thus, the conceptual analysis of the text is moving from concept to lexical field of its explication and from vocabulary to concept, and within content of which there are various types of information synthesized [1].

2. Analysis of the latest sources and publications

More and more scientists in the field of computational linguistics today must turn to the analysis of concepts and search patterns of their interaction in the texts. Otherwise, it is impossible to execute content analysis of text documents.

For example, to search for synonyms American computer linguists [2] use the distance between concepts in the thesaurus, introducing the definition of "synonymous distance." Analysis of conceptual structures is also involved in the field of information technology, where there is a need of knowledge extraction from electronic texts [3-5], because contents relationships come up exactly between concepts, reflecting the particular situation.

At the annual conference on computational linguistics and intellectual technologies "Dialogue 2014", most reports are focused on semantic analysis of textual information [6]. Features and functions of the concepts in terms of cognitive linguistics are described in Maslova work [7].

However, specificity of different languages is not evident at the level of concepts, but at the level of grammatical forms [8]. Therefore, an actual task is the development of methods for processing text information that would associate conceptual and syntactic structure of text.

3. Analysis of tools for building semantic relationships in text by using logic and linguistic models

Logic and linguistic modeling is a method for building knowledge-based systems with a learning capability using conceptual models and modal predicate logic.

Logic and linguistic models have a superficial similarity to Sowa's Conceptual Graphs, both use bubble style diagrams, both are concerned with concepts, both can be expressed in logic and both can be used in artificial intelligence. However, logic and linguistic models are very different in both logical form and in their method of construction.

For future reference it is proposed to stop at the following definition of logic and linguistic model: an expression or expressions in terms of a natural language and linguistic variables (which can be not only numbers but also the words and phrases of natural or artificial language). The basis for the logic and linguistic models is the logic of predicates where the predicate – a *P*-function, which takes the value 0 or 1, and the arguments which ranges value from an arbitrary set of M: $P(x_1, x_2, \ldots x_n)$).

Logic and linguistic model covers all conceptual relationships that may be encountered in the text and reflects the syntactic structure of arbitrarily complex sentence of natural language, that allows you to extract knowledge from text information [9].

Usually, the substance of the sentence is associated with the presupposition – component content of the sentence that must be true in order that the sentence is not perceived as abnormal. Presuppositions which define the content of the statements may be contained in the preceding context. Then, identifying of conceptual relationships is possible by matching and replacing the predicate variables and constants in the logic and linguistic models [10].

Chains of proper names – proper name take its specific substantive value, only coming into the chain of titles. So, there are means that enhance the informative quality of text and belong to the formal parameters of relationships of text structural units [11].

For example, "Louis Armstrong was one of the most famous and best-loved jazz musicians of all the time. Armstrong did a great deal to popularize this type of music". Logic and linguistic model for this fragment is:

Was (Louis & Armstrong, one [musicians {most [famous], best-loved, jazz} [all [time]]]) / Did (Armstrong, deal {great}) & Popularize (Armstrong, music {jazz}).

It was made replacement "this type of" for "jazz". In the first sentence of the text it was given an explanation of who was Louis Armstrong and how he related to music. Then the second sentence explained that he made a great contribution to popularize jazz.

Deictic repetition or anaphoric links point to already called objects, features and circumstances using special words, pronouns, adverbs, semantically devastated words that are used in place of a word or phrase from the previous content, numerator as the subject and others.

Let there is a text "Louis Armstrong loved music from a very early age. His music teacher Mr. Davis gave him bugle and concert lessons, and the boy had never been happier. He learned quickly and was soon made the leader". Logic and linguistic model for this fragment is:

Loved (Louis & Armstrong, music [age {very & early}])/Gave (Mr. & Davis[music[teacher]], Louis & Armstrong [lessons { bugle & concert }]) & Had & never & been (Louis & Armstrong, happier) / Learned (Louis & Armstrong, quickly) & Was & made (Louis & Armstrong, leader, soon).

In the second sentence, the pronoun "his" and "him" indicate "Louis Armstrong". The words "boy" and "Louis Armstrong" are synonymous because the pronoun "he" from the third sentence also refers to "Louis Armstrong". Therefore deictic repetition allows to make replacement in the logic and linguistic models.

Gerunds phrase at the beginning of the sentence serves as the circumstances of the time value or hue reason that relates to the entire sentence as a whole. Gerunds phrases establish the relationship of one event to another, indicate the sequence of actions, can mean the action that takes place already mentioned, but prior to the events described hereinafter. This tool of cohesion allows to depict events without describing them separately, but considering the starting point for the other action.

Let there is a text "Armstrong began travelling all over with his band. Louis was becoming known as the best player around New Orleans". Logic and linguistic model of the sentence will look like:

Began& travelling (Armstrong, all [over [band {his}]]) → Was & becoming& known (Louis, player {best} [around [New & Orleans]]).

That tour with band made him well-known around New Orleans, so in the logic and linguistic model there is an implication operation used.

Discourse words – coordinated and subordinate conjunctions. Conceptual relationships between sentences in the text are expressed by means of conjunctions that stay at the beginning of sentences. Choosing of the conjunction relates to the nature of semantic relationships in the text (simultaneity or sequence of events, the contrast dependence, alternative choice, and so on).

For example, there is text "Before he became a writer, O. Henry had been a bank office worker, a cowboy, a reporter and a tramp trying to find a way to make a living".

In this fragment, conjunction "before" indicates that some event occurred in the past, and the first part of a complex sentence follows from the second one. As you can see from the example, discursive words affect operations used in the logic and linguistic models:

Had & been (O. & Henry, bank [office [worker]], cowboy, reporter) & Tramp& trying & to & findv (O. & Henry, way[make[living]]) → Became (O. & Henry, writer).

Ellipsis or contextually incomplete syntactic structures – cohesion tools, which means a space in the text of a particular linguistic unit that can be reproduced within the meaning of the preceding sentence. Semantic and structural incompleteness of sentences related to the fact that the description of the new situation, some elements are already mentioned in the previous context and do not require re-designation. Incomplete designs cannot be used independently and implement its communicative function only with other sentences of the text.

For example, in the sentence "The purpose is to provide remote electronic banking, teaching, shopping, taxpaying, game playing, video-conferencing, film ordering, medical help – the list goes on".

Dash makes it possible to understand that there are listed not all facilities of remote control. Logic and linguistic model of the sentence:

Provide & remote (purpose, electronic [banking], teaching, shopping, shopping, taxpaying, game [playing], video-conferencing, film [ordering], help {medical}) ~ Goes & on (list).

Parentheses execute formation function of the text, indicating the order of idea presentation, f.e. first, second, finally, consequently, by the way, in other words, and so on.

Parentheses do not appear in the logic and linguistic model, but they are essential for relationships between natural language sentences and also create stylistic coloring of the text.

Suppose, there is a passage of text "Managers need a network of contacts and human relationships, because to achieve organizational goals. Therefore, human-relation skills are very important skills for a manager".

Parenthesis "therefore" indicates that the second sentence contains the result, as mentioned earlier in the text, so the logic and linguistic model of the fragment will look like:

Need (managers, network [contacts] & network [relationships {human}) → Achieve(,organizational [goals]) → Are & very & important (human-relation [skills], skills [manager]).

4. Creating a finite state transition network to interpret semantic relationships in text

Content unity of the text at the level of natural language is implemented as a sequence of sentences linked by semantic relationships. Abstract models, which underlie the construction of texts and provide such a semantic unity called thematic progression [12]. By F.Danesh there are five types of thematic progressions. In order to reveal the conceptual relationships in the text, we present each type of thematic progressions as a finite state transition network (FTN).

FTN is represented by set of nodes and directed arcs connecting them. These nodes correspond to nonterminal symbols and arcs to terminal symbols [13]. Where S is initial node, and S* is final node. Terms of use of the finite state network:

1) We have to choose one of the directed arcs, which comes from this node and go through it.

2) When the arc is passed, we have to choose one of the terminal symbols subset corresponding to that arc.

3) Continue the process until a node S* would be reached.

Write down the procedure for consideration of FTN's nodes to track semantic-conceptual relationships in texts of various kinds of thematic progressions. Sentence is a minimal and basic communication unit of language. The offer must be holistic and transmit information throughout the complexity of dependencies and relationships [14]. Therefore each FTN's node will reply to the natural language sentence, and arcs reply to the type of syntactic relationship between sentences. Thus, the substantival relationship will occur when sentences are related by repeat of co-root words or synonyms; pronoun relationship – if in each following sentence pronoun is taken instead of the object, which was discussed in the previous sentence; conjunctive relationship – when there are compound and complex sentences with appropriate conjunctions, verbal relationship – all the sentences come with the same tenses, resulting the text becomes well-defined space frame.

Simple linear progression – there is typical consistent deployment of information when rheme of preceding sentence becomes the theme of the next sentence. Sot the deployment of text proceed from present(theme) to new (rheme).

Suppose, there is a passage of text "Louis Armstrong loved music from a very early age. His music teacher Mr. Davis gave him bugle and concert

lessons, and the boy had never been happier. He learned quickly and was soon made the leader. He managed to save enough money to buy an old cornet. He began to practice and listen to music every chance he got." A finite state transition network for it will look like Fig. 1.

Where A – "Louis Armstrong loved music from a very early age"; B – "His music teacher Mr. Davis gave him bugle and concert lessons, and the boy had never been happier"; C – "He learned quickly and was soon made the leader"; D – "He managed to save enough money to buy an old cornet"; E – "He began to practice and listen to music every chance he got".

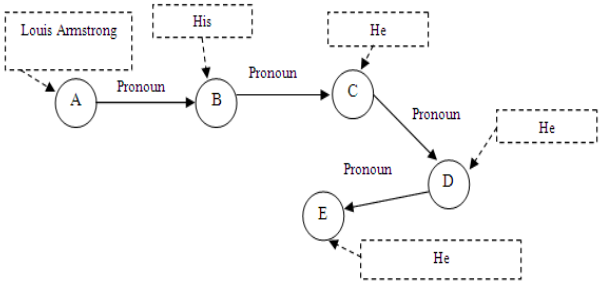

Fig. 1. Example of FTN for simple linear progression

Progression of underlying theme – is characterized by a single theme that is repeated in every sentence of text. Thus, one and the same theme pervades the entire text. In this model, the first element of the chain may be optional.

For example, there is text "Management is a process of managing people. Any manager has some functions. He performs planning, organizing, leading and controlling. Planning is choosing an organizational mission. All other functions depend on this one. Organizing is determining what resources activities are required and delegating the authorities to employees." A finite state transition network for it will look like Fig. 2.

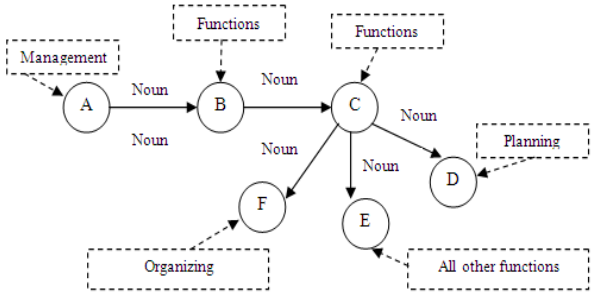

Fig. 2. Example of FTN for progression of underlying theme

Where A – "Management is a process of managing people"; B – "Any manager has some functions"; C – "He performs planning, organizing, leading and controlling"; D – "Planning is choosing an organizational mission"; E – "All other functions depend on this one"; F – "Organizing is determining what resources activities are required and delegating the authorities to employees".

Progression with derived themes – each sentence of text, which doesn't have incorporated elements of consistent lemmatization (the first type of thematic progressions) or cross thematization (second type) is used to express the overall direction of the text content. Thus, a number of selected themes reveal one common hipertheme that can be defined in text, and may be missed.

There is a passage of text "There are different kinds of computers. Some do only one job over and over again. There are special-purpose computers. Each specific application requires a specific computer. One kind of computer can help us build a spacecraft, another kind of computer can help us navigate that spacecraft. A special-purpose computer is build for this purpose alone and cannot do anything else." A finite state transition network for it will look like Fig. 3.

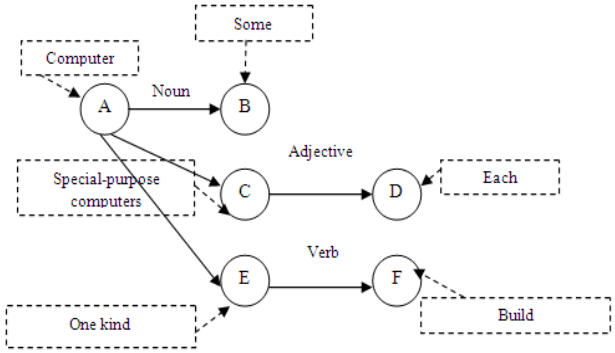

Fig. 3. Example of FTN for progression with derived themes

There A – "There are different kinds of computers"; B – "Some do only one job over and over again"; C – "There are special-purpose computers"; D – "Each specific application requires a specific computer"; E – "One kind of computer can help us build a spacecraft, another kind of computer can help us navigate that spacecraft"; F – "A special-purpose computer is build for this purpose alone and cannot do anything else".

Progression of split theme – based on double rheme, components of which form (during thematization) initial points for the development of specific thematic progressions.

For example, there is text "Indians had been living in North America long before any Europeans arrived. Indians who settled in northern areas hunted and fished. Those who settled in the east and southwest farmed. Despite their many differences, most Indians shared the belief that people should live in harmony with nature. They believed that people should not own land because the land, like the air, stars and water, belonged to everyone." A finite state transition network for it will look like Fig. 4.

Where A – "Indians had been living in North America long before any Europeans arrived"; B – "Indians who settled in northern areas hunted and fished"; C – "Those who settled in the east and southwest farmed"; D – "Despite their many differences, most Indians shared the belief that people should live in harmony with nature"; E – "They believed that people should not own land because the land, like the air, stars and water, belonged to everyone".

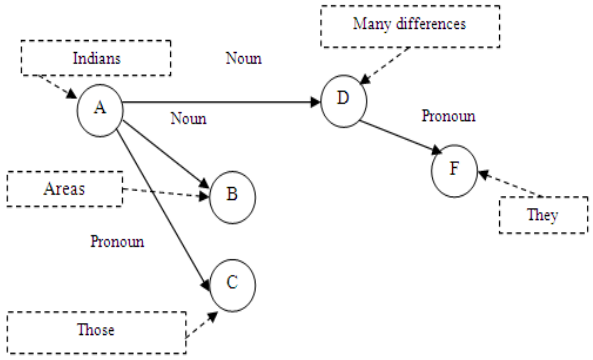

Fig. 4. Example of FTN for progression of split theme

Progression of thematic jump – requires a break in theme-rheme chain that can be restored from the context.

There is a passage of text "The English are great lovers of competitive sports. Britain was the home of the modern world's most popular sports. Football is the best example. The game peculiarly associated with England is cricket. Football is the most popular British sport." A finite state transition network for it will look like Fig. 5.

Where A – "The English are great lovers of competitive sports"; B – "Britain was the home of the modern world's most popular sports"; C – "Football is the best example"; D – "The game peculiarly associated with England is cricket"; E – "Football is the most popular British sport".

In real texts the thematic lines intertwine in various ways, so the listed above thematic progressions cannot cover all types of relationships in text. However, the thematic progressions are used to monitor moving of information in text.

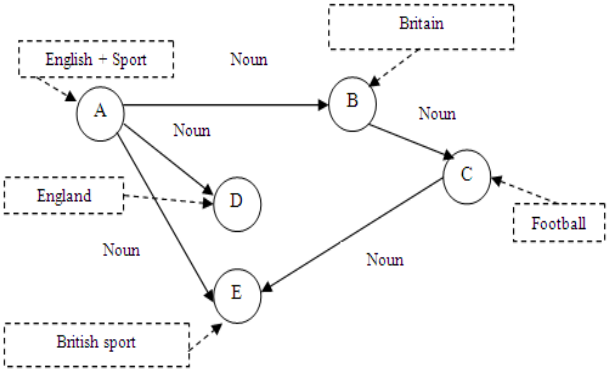

Fig. 5. Example of FTN for progression of thematic jump

5. A self-system to identify conceptual relationships in text

Analysis of tools for building semantic relationships in texts and researching the ways of its displaying and recording, allows to create a project of self-system for identifying conceptual relationships in texts.

In many cases, the basis of the self-organized is requirement to maintain constant a transfer function of a closed system (which corresponds to quality of the management process. With variable settings of object, self-system of automatic control compared to conventional systems has special additional profile of self-loop configuration [15]. The stages of a process of self-adjusting system for identifying conceptual relationships in text showed on Fig. 6.

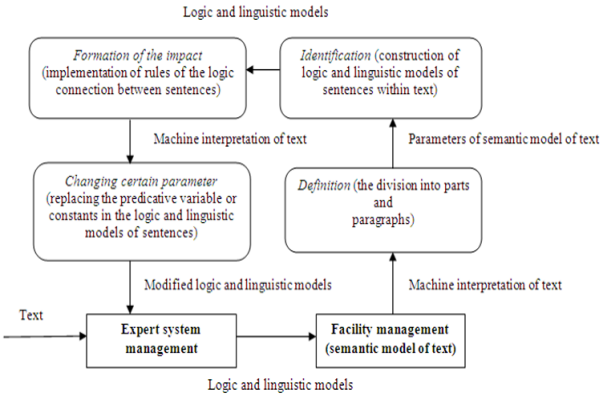

Fig. 6. The stages of a process of self-adjusting system for identifying conceptual relationships in text

1) Definition (change, calculation) of initial factor for the self-organized. At this stage there is certain variable measured, in this case text. It is performed the division of text, and defined the number of parts and paragraphs in it. Measurement of value can directly be the initial factor for the self-organizing.

2) Identification (identifying key parameters or characteristics required for the self-organized) – involves the construction of logico-linguistic models of sentences which included in the given text for the definition of semantic and syntactic structure of logical components.

3) Formation of the impacts on part of management system that is being organized. It is based on the value of tools for building semantic relationships. This stage is to determine how we should change a certain parameter of system, ie what kind of replacements can be done in logico-linguistic model of sentence. As a result, appropriate corrective signal is fed into a part of a system that is configured (organized).

4) Changing certain parameter of logico-linguistic model. At this stage there is a realization of parameter changing or configuring the structure of the system in accordance with set corrective signal.

The created system of identifying conceptual relationships in text actually refers to systems, equivalent to the self-organized because the parameters of its control rule vary depending on tools for building semantic relationships. Self-organized system identifying conceptual relationships in text works with electronic documents, creates its content model by constructing logic and linguistic models of natural language sentences, and then, based on the rules of forming semantic relationships, makes replacements in logic and linguistic models. As a result of these operations there is produced a new informative model of text. Moreover, the self-organized component searches for semantic relationships in text.

6. Conclusions

Analysis of tools for building semantic relationships in text documents, including the use of chains of proper names, deictic repetition or anaphoric links, gerunds phrase, discourse words, ellipsis or contextually incomplete syntactic structures and parentheses, made it possible to identify mechanisms of change logic and linguistic models of natural language sentences.

Reflecting thematic progressions in the form of finite state transition network is a kind of formal record of conceptual relationships in text. These researches (obtained rules, laws) became the basis for creation a project of self-system for identifying conceptual relationships in texts. Thus, the system will give the ability to track the logical links in texts, and to perform automatic, semantic and linguistic analysis.

References

[1] *Proskuryakov M. R.* Conceptual structure of text, disser. St. Petersburg, 2000, 330 p. (in Russian).

[2] *Dan Jurafsky, Christopher Manning.* Natural Language Processing, 2012 (in English). Available at: https://www.coursera.org/ course/nlp (accessed 15.05.2014).

[3] *Greimas A. J.* Structural semantics. Search method. Moscow, Academic Project, 2004, 368 p. (in Russian).

[4] *Gorunova O. N.* Conceptual structure of the text of the advertising message, disser. St. Petersburg, 2005, 183 p. (in Russian).

[5] *Paducheva E. V.* Dynamic models in the semantics of lexical. Moscow, Slavonic languages of culture, 2004, 608 p. (in Russian).

[6] *Computational* Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2014). Issue 13, 2014, ((in Russian)). Available at: http://www.dialog-21.ru/digest/2014/pdf/ (accessed 14.11.2014).

[7] *Maslova V. A.* Cognitive linguistics. Minsk, TetaraSystems, 2008, 272 p. (in Russian).

[8] *Plotnikova S. N.* Concept and concept analysis as a linguistic method of studying social intelligente, vol. 2(18), 2012. (in Russian).

[9] *Vavilenkova A. I.* Knowledge extraction from natural language sentences, Program prodacts and systems, 2012. – P. 87–90 (in Russian).

[10] *Geeraerts Dirk.* Cognitive linguistics: basic readings research / Dirk Geeraerts, Rene Dirven, John R. Taylor. Berlin-New York: Mouton de cruyter, 2006, 486 p. (in English).

[11] *Golovkina C. C.* Linguistic analysis of text. Vologda, 2006, 124 p. (in Russian).

[12] *Valgina N. S.* The theory of text, Moscow, 2003. (in Russian).

[13] *Alfirenko N.* Disputes of semantics, Moscow, 2006, 326 p. (in Russian).

[14] *Phillipov K. A.* Text Linguistics, St. Petersburg: St. Petersburg university Publ., 2008, 336 p (in Russia).

[15] *Trubezhlov D. I., Mchedlova E. S. Krasichkov L. V.* Introduction to the theory of self-organized open systems, Moscow, 2002, 200 p. (in Russian).

Received 17 November 2014.

А. І. Вавіленкова. Самоорганізаційна система виявлення концептуальних зв'язків у тестах
Національний авіаційний університет, просп. Космонавта Комарова, 1, Київ, Україна, 03680
E-mail: a_vavilenkova@mail.ru

Запропоновано проект системи виявлення концептуальних зв'язків у текстах, що базується на використанні логіко-лінгвістичних моделей текстової інформації. У матеріалах статті проаналізовано засоби побудови змістовних зв'язків у текстових документах. Проведено паралель між створенням тематичних прогресій та мереж переходів із кінцевим числом станів. Продемонстровано недоліки роботи методу розширених мереж переходів та висвітлено переваги побудови логіко-лінгвістичних моделей речень природної мови як основного механізму функціонування самоорганізаційної системи. Виявлення концептуальних зв'язків стало можливим за рахунок зіставлення та заміни предикатних змінних і констант.
Ключові слова: концептуальні відношення; логіко-лінгвістична модель; природна мова; розширена мережа переходів; семантичне моделювання.

А. И. Вавиленкова. Самоорганизационная система выявления концептуальных связей в текстах
Национальный авиационный университет, просп. Космонавта Комарова, 1, Киев, Украина, 03680
E-mail: a_vavilenkova@mail.ru

Предложен проект системы выявления концептуальных связей в текстах, который базируется на использовании логико-лингвистических моделей текстовой информации. В материалах статьи осуществлен анализ способов построения смысловых связей в текстовых документах. Проведена параллель между созданием тематических прогрессий и сетей переходов с конечным числом состояний. Продемонстрированы недостатки работы метода расширенных переходов и высветлены преимущества построения логико-лингвистических моделей предложений естественного языка как основного механизма функционирования самоорганизационной системы. Выявление концептуальных связей стало возможным благодаря сопоставлению и замене предикатных переменных и констант.
Ключевые слова: естественный язык; концептуальные отношения; логико-лингвистическая модель; расширенная сеть переходов; семантическое моделирование.

Vavilenkova Anastasia (1984). Candidate of Engineering. Associate Professor.
Department of Computerized Control Systems, National Aviation University, Kyiv, Ukraine
Education: National Aviation University, Kyiv, Ukraine, 2007.
Research area: information technologies, computational linguistics, the formation of the logic and linguistic models.
Publications: 60
E-mail: a_vavilenkova@mail.ru