УДК 81′322

D. V. Lichargin, K. V. Safonov, N. V. Nikolaeva, E. B. Chubareva

PRINCIPLES OF COMPRESSION AND DECOMPRESSION WITHIN THE FROME OF MULTIDIMENSIONAL ANGUAGE REPRESENTATION APPLIED TO EDUCATIONAL TASKS GENERATION*

In the given work the problem of semantic decompression patterns addition to the meaningful sentences generated as functions in multidimensional vector space of the notions of the natural language is considered. The model of adding semantic decompression patterns based on stylistically-oriented sets of generative grammar rules is offered. The complexity of natural language translation problem is discussed. The considered model provides the algorithm of adding semantic decompression patterns that can be used for improving the performance of educational tasks generation software.

Keywords: Natural Language Generation, Meaningful Sentences Generation, Turing Test.

At the present moment, people are facing a huge amount of information that is not always well absorbed and efficiently used because of the complexity of its structure. Presenting the language as a model of a multidimensional data set can improve the quality of linguistic software. A multidimensional view of data on natural language is important for the construction of electronic translators, abstracting systems, expert systems, generative grammar, etc. In this regard, the analysis of a multidimensional model of language data is relevant at the present stage of development the information technology and mathematical foundations of computer science.

The problem of formal modeling of natural language, particularly, English, is the central task for computational linguistics – a discipline that lies at the intersection of computer science, mathematics, systems analysis, linguistics, philosophy, psychology, etc.

Solving the problems of developing linguistic software successfully implemented numerous theories, concepts and software systems. Numerous works in the field of semantics, discrete mathematics, linguistics and Artificial Intelligence, let people hope for solution many of the problems of formalization of natural language and passing the Turing test in increasingly tough conditions for the test systems in the near future.

For solving the problem of generating the meaningful speech a lot of tools are used today by both Semantics and Artificial Intelligence within the notional apparatus and the various models of mathematical Semantics. In particular, the analysis of natural language was traditionally applied within the following models and tools such as the method of ontology, the method of linguistic classification, the method of multidimensional data, OLAP systems, relational database, frames, generative grammars, in particular, generating Montague grammar, semantic network theory graphs and the resolution method, hybrid systems, and linguistic methods, such as component analysis, the paradigmatic method, the approach of American structuralism, etc [1–8].

There is a topical problem of generating meaningful subsets of the natural language with various approximations. The solution to this problem greatly simplifies tasks such as the construction of expert systems, e-learning systems, automatic transfer systems, programs to support dialogue with users, creating natural-language interface. The solution to this problem is mainly determined by the problem of passing Turing test by software systems, providing identity and the inability to distinguish a dialogue with a person and a dialogue with a software system.

The novelty to offer a classification of generative grammar rules or relational patterns subsets, based on the multidimensional model of the natural language based on the proposed vectorized semantic classification of words and notions of natural language.

The main idea is to view the grammatical and lexical spaces sequence linked by the generation process according to this or that generative grammar rules subset or relational patterns subset. The idea can be applied to solving the task of automatic educational tasks generation algorithms creation.

In the work of D. V. Lichargin “The Methods and Tools for the Generation of Semantic Structures is the Natural Language Interface of Software Systems” and “A Multidimensional View of Data on Vocabulary and Grammar of the English Language” the following model of the meaningful natural language generation is offered.

One can specify the states space for such units of the natural language, as words and notions. The space of the grammar of language is described by the space coordinates (Fig. 1, 2):

– Members of the Sentence (Subject, Predicate, Object, …);

– Parts of Speech (Noun, Pronoun, Verb, …);

– Grammar Categories (Plural, Collective, Superlatives, …).

Next, we construct the lexical space of words of the language (data cube) with the following coordinates is presented:

– Word Order (Doer, Action, Receiver, Property of the Receiver, …);

– Topics (Food, Clothes, Body, Building, Money, …);

– Options for Substitution of Words in a Sentence (Possitive, Neutral, Negative; Maximal, Large, Medium, Little, Minimal, …).

*Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение 14.B37.21.1010.

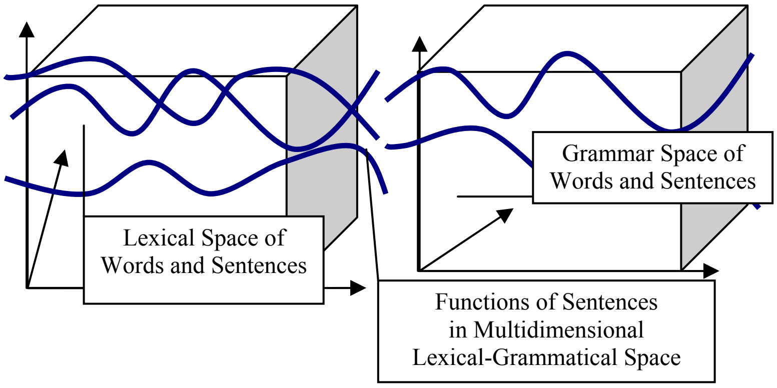

Grammar Space of Words and Sentences
Lexical Space of Words and Sentences
Functions of Sentences in Multidimensional Lexical-Grammatical Space

Fig. 1. The Lexical and Grammatical Space in the Model of Meaningful Language Generation

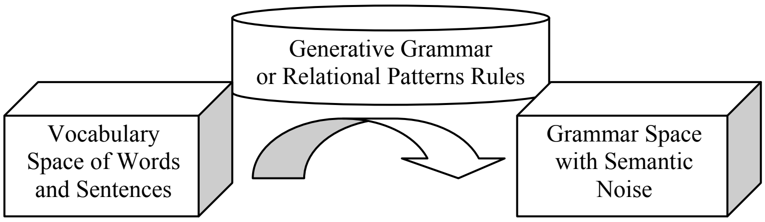

Generative Grammar or Relational Patterns Rules
Vocabulary Space of Words and Sentences
Grammar Space with Semantic Noise

Fig. 2. The principle of Lexical Space – Grammatical Space transformation

Methods of Pattern-Based Semantic Decompression patterns Adding and Text Composition / Decompression

Doer	Action	Object	Substance
I	Eats	The …	With / without …
We	Cooks	Dish	Beef
Bob	Roasts	Potatoes	Fish
I	Sews	The …	From …
They	Knits	Jacket	Wool
The girl	Irons	Shirt	Cotton

Subject	Predicate	Object	Modifier
My DOER's	*ACTION.MAKING-ing*	*Needs / requires / ...*	*Good / nice / ... + SUBSTANCE*
My cook's	roasting	needs	(good) beef
My mother's	sewing	refers to	(brilliant) silk
This / the / the given + SUBSTANCE	*Is good / nice / ideal / ... for*	*For my / his / her / ... + DOER*	*To + ACTION*
Silk	Is good for	My mother	To sew
Fish	Is ideal for	My brother	To cook
ACTION-ing	*Cannot / will not + go on / continue / be done / be all right*	–	*Without + such / this / ... > like this + SUBSTANCE*
Cooking	Will not be done	–	Without beef
Knitting	Will not be all right	–	Without wool

In this respect, when addressing the problem of real texts processing in a particular language as a part of program – generators and analyzers of meaningful speech, it is necessary to solve the problem of removing the semantic decompression patterns. Semantic decompression patterns are present in the texts written in different language styles: from the academic style with a low degree of semantic decompression patterns up to slang with an extremely high degree of semantic decompression patterns.

Semantic decompression patterns are the formal interpretation of emotional expression and depth of the subject. The computer, in particular, can consider it to be unimportant for the task of generating any units of the natural language: words (such as neologisms) sentences and texts.

For the generation of statements with semantic decompression patterns one can also use the method of generating semantic decompression patterns by analyzing the semantic structures of notions and their transformation. For example, the word "to like" corresponds to the vector of coordinates:

[RELATION–CREATURE–X \ ESSENCE \ POSITIVENESS].

The word "beautiful" corresponds to the vector attributes:

[RELATION–X \ ESSENCE \\ RELATION–CREATURE–X \ IDEA \ ON (NOT) LIVE \ POSITIVENESS].

The word "to look" corresponds to the vector of semantic features:

[RELATION–CREATURE–X \ ESSENCE \\ RELATION–CREATURE–X \ IDEA \ ON (NOT) LIVE \ POSITIVENESS].

As a result, it is possible to take an opportunity to regroup the semes of the natural language within a semantic web for each of the word. For example, the phrase "the apple is beautiful" can be transformed into the phrase "I like the form of the apple". In this case, the concept of "beautiful" falls into the group of semes with the meaning "to see" and a group of semes with a value of "good", "love" or, for example, "nice".

Based on the offered model the following scheme of natural language generation is offered (see the table).

For example, the sentence "I cook dinner" can be transformed into ("My cooking dinner" … & "The dinner being cooked by me …" & "It was … for me to cook dinner" & "My dinner after cooking …"). The example of parallel generation in the other topics is presented below: "We build the museum" → ("Our building the museum …" & "The museum being built by us …" & "It was … for us to build the museum" & "Our museum after building …"), similarly "They listen to the music" → ("Their listening to the music …" & "Music being listened to by them …" & "It was … for them to listen the music" & "Their music after listening …").

The generative grammar or relational patterns, used for sentences / text composition and decompression and semantic decompression patterns addition, can be subdivided into stylistic subclasses like the ones below.

1. Common Style.
1.1. Slang.
1.1.1. Tabooed Style.
1.1.2. Criminal Argo.
1.2. Neutral Common.
1.3. Pun.
1.4. Mass Media Style.
2. Artistic Style.
2.1. Poetry.
2.2. Prose.
2.1. Fantasy.
2.2. …
3. Scientific Style.
3.1. Academic.
3.2. General Science.
3.3. Popular Science.
4. Religious Style.
4.1. Orthodox.
4.2. Buddhism.
4.3.Islam.
4.4. …
5. Neutral Style.
5.1. Journalistic Style.
6. Mixed Style.

Below a general scheme of reducing a complex literary phrase to a simplified synonymic equivalent by adding logical, semantic, grammatical, morphological, sound and other noise addition patterns / decompression patterns is presented (Fig. 3).

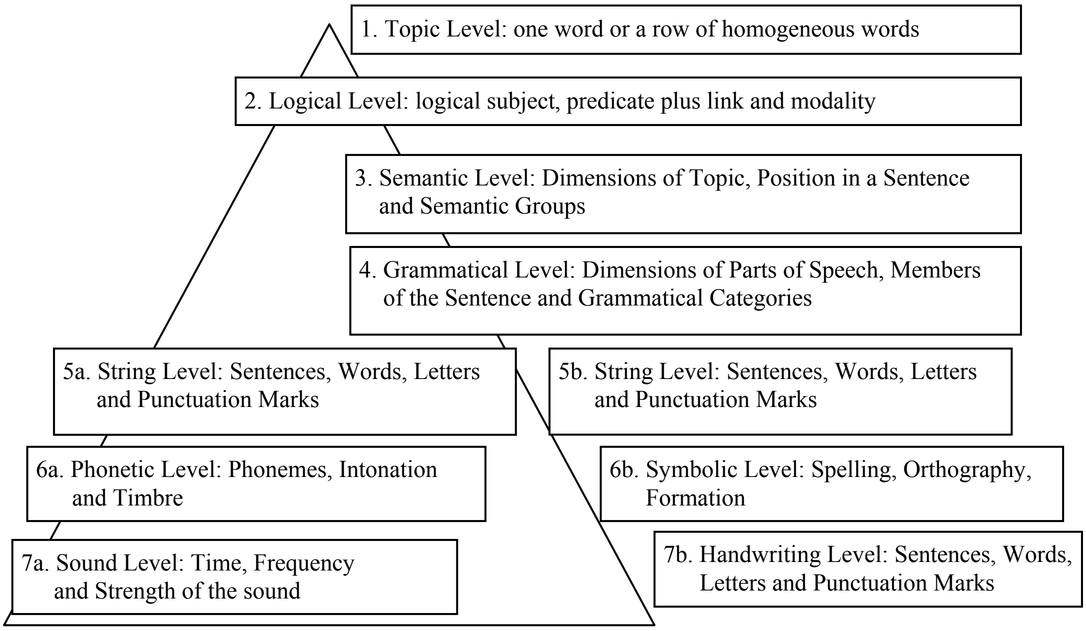

1. Topic Level: one word or a row of homogeneous words
2. Logical Level: logical subject, predicate plus link and modality
3. Semantic Level: Dimensions of Topic, Position in a Sentence and Semantic Groups
4. Grammatical Level: Dimensions of Parts of Speech, Members of the Sentence and Grammatical Categories
5a. String Level: Sentences, Words, Letters and Punctuation Marks
5b. String Level: Sentences, Words, Letters and Punctuation Marks
6a. Phonetic Level: Phonemes, Intonation and Timbre
6b. Symbolic Level: Spelling, Orthography, Formation
7a. Sound Level: Time, Frequency and Strength of the sound
7b. Handwriting Level: Sentences, Words, Letters and Punctuation Marks

Fig. 3. The decompression patterns tree going through semantic, grammatical, morphological, phonetic multi-attribute descriptions – multidimensional spaces

This approach supposed to be used for creating different kinds of linguistic ware, for example, summarizing systems, electronic translation systems, expert systems, natural texts data extraction systems and others.

Let's view some principles of translation of a complex literary text from the point of art – semantic noise computer interpretation. Translation theory, based on a solid foundation on understanding of how languages work, translation theory recognizes that different languages encode meaning in differing forms, yet guides translators to find appropriate ways of preserving meaning, while using the most appropriate forms of each language.

Translation theory includes principles for translating figurative language, dealing with lexical mismatches, rhetorical questions, inclusion of cohesion markers, and many other topics crucial to good translation, which are supposed to be viewed as lexical, grammatical, morphological, phonetic categories multidimensional spaces data presentation as multilevel means of compression tree generation for example.

1. "The driver carries the cans …" – Водитель привозит консервы … .

1.1. "The cans carried by the driver …" – Консервы, привозимые водителем … .

1.2. "The driver, who carries the cans …" – Водитель, который привозит консервы … .

1.3. "Carrying the cans by the driver …" – Перевозка консервов водителем … .

1.3.1. "Carrying the cans from the side of the driver …" – Перевозка консервов со стороны водителя … .

1.3.2. "Carrying the cans by the efforts of the driver …" – Перевозка консервов усилиями водителя … .

1.3.2.1. "The process of carrying the cans by the efforts of the driver …" – Процесс перевозки консервов усилиями водителя … .

1.3.2.2. "The task of carrying the cans by the efforts of the driver …" – Задача перевозки консервов усилиями водителя … .

1.3.2.3. "Carrying the goods, for example, cans by the efforts of the driver …" – Перевозка товаров, например, консервов усилиями водителя … .

1.3.2.3.1. "Carrying the consumer goods – cans by the effort of the driver …" – Перевозка потребительских товаров – консервов – усилиями водителя … .

1.3.2.3.2. "Carrying the goods – round cans by the effort of the driver …" – Перевозка товаров – круглых консервов – усилиями водителя … .

1.3.2.3.3. "Carrying the goods – metal cans by the effort of the driver …" – Перевозка товаров – металлических консервов – усилиями водителя … .

It is also important for human translation process [9; 10] and teaching future translators and interpreters. The translator should understand perfectly the content and intention of the author whom he is translating. The principal way to reach it is reading all the sentences or the text completely so that you can give the idea that you want to say in the target language because the most important characteristic of this technique is translating the message as clearly and natural as possible. The translator should have a perfect knowledge of the language from which he is translating and an equally excellent knowledge of the language structure into which he is translating. At this point the translator should have a wide knowledge in both languages for getting the equivalence in the target language, because the deficiency of the knowledge of both languages decompression and compression principles will result in a translation without logic and sense.

The translator should avoid the tendency to translate word by word, because doing so is to destroy the meaning of the original and to ruin the beauty of the expression. This point is very important and some reader can express another meaning or understanding in the translation. The translator should use the types of speech in common usage. The translator should bear in mind the people to whom the translation will be addressed and use words that can be easily understood. Similarly electronic translation systems should have an option for simplifying the translation for a definite user – form a child to a specialist in another sphere, an ordinary user. All this makes the problem of translation a task of AI level.

The translation adequacy must be taken into account in fiction literature translation especially carefully, and much more intellectual electronic translators should be used for this in the future. Khatuna Beridze, lecturer of the translation theory and practice at the Batumi State University mentions that there are both linguistic and extralinguistic aspects that hinder to reach adequacy in fiction translation. Semantic information of the text differs essentially from the expressive-emotional information of the text but they have one common trait: both can bear and render extralinguistic information. Extralinguistic information often becomes a stone to stumble over by a translator, as it is a lingvoethnic barrier for a fiction literature translator; misunderstanding or misinterpretation of the extralinguistic information means can be the following:

1) Either what was actually communicated in the SL text, what means the pragmatic core of the SL text may be lost or therefore in the TL text ambivalence may arise for the recipient reader;

2) Or there may be misrepresented the author's communicative intention, the social context of the scene/situation as well as disposition or relationships of the communication participants.

Good examples could be brought from translation of stories and novels by V. P. Astafiev. The works of world famous Siberian writer V. P. Astafiev are very difficult to translate because his language is very distinctive, he uses a lot of non-standard events (phenomena), and the translator must handle the material carefully, for not to lose the value and meaning of the content. All non-standard language features in his works can be divided into the following groups: dialect, vernacular, colloquial vocabulary, considering that the vernacular vocabulary is relatively the largest group.

Abnormalities in the speech of the characters serve as a mean of typification or, alternatively, as a mean of linguistic characteristics of a story hero. Abnormalities in the speech of the characters appear at all levels of the language: phonetic – phonetic distortion of the word shape, morphological – mistakes in coordination of number, gender, declension of nouns, syntactic – the use of specifiers (“эдак”, “нате вам”) to connect the sentence. However, most examples of deviations from the norm are at the level of vocabulary that can be considered as a marker of social status of heroes.

So, the semantic compression and decompression should take into account all the tiny particularities of the natural language.

In the author’s speech non-normative events often serve as a mean of creating imagery, expressing the author’s expression. Sometimes the vernaculars in the speech can be considered as the author’s irony.

Let’s see some examples of vernaculars from Astafiev’s text.

«Варнак» – villain (злодей, негодяй).

«Охламон» – dolt (дурень, болван).

«Поросята хоркают» (хрюкать) – to snort (фыркать).

«Закокать курочкой» (кудахтать) – cackle like a hen.

Sometimes interpreters can’t find the appropriate translation of definite phrases, because of distinctive character of Astafiev’s prose.

«Фулиган» – hooligan.

«За что жисть погубил» – What have I made such a mess in my life for.

«А тютюшеньки-тютю» – A-diddums, a-diddums, a-diddums!

«А люлюшеньки-люлю» – A-doodle, a-doodle, a-doodlums!

«А малюшеньки-малю» – A-doodle, a-doodle, a-doodlums!

In some cases interpreters translate words and phrases in such case that readers can understand the sense, but originality of author’s language is lost.

«Жили худо, бедно, натужно и недружно, вразнопляс» –

We had a mean, poor, strained, discordant life, dancing to different tunes.

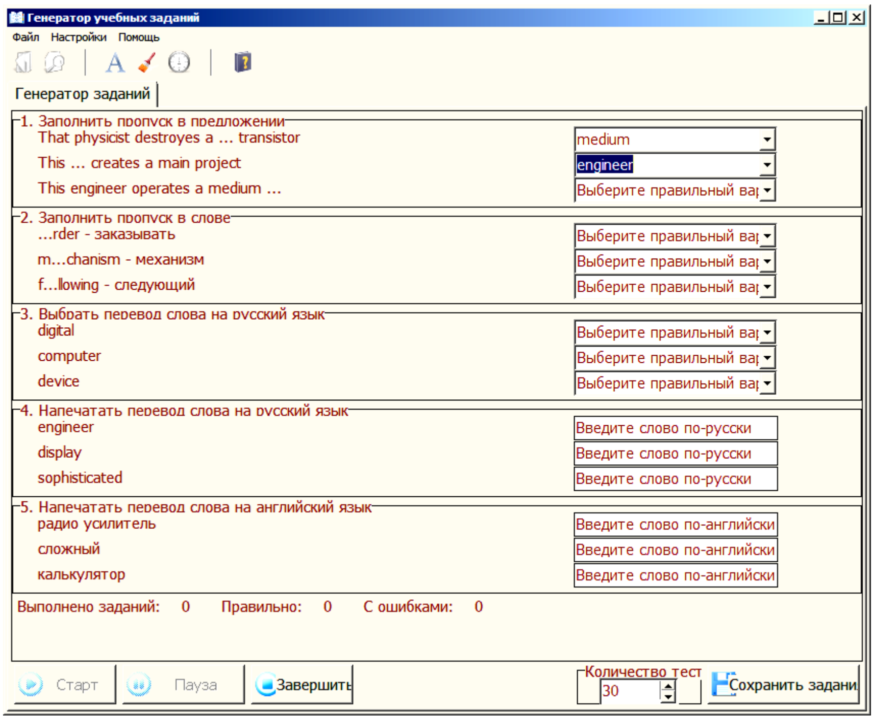
«Исподличался совсем» – I had turned into a real little scoundrel.

The algorithm of more effective translation will be viewed as the following sequence of actions.

1. Writing generative grammar rules for potentially fully generation of a natural language subset in a definite approximation.

2. Potential generation of all possible states of the language around the standard classifications of the needed natural language units – standard words, new, occasional words, simple sentences, sentences with artistic phenomena (conditionally) interpreted by computer as semantic noise.

Because of the fact that deep emotional connotation is a very difficult task for linguistic software, the task to create the software, algorithms and approaches for automatic task generation was selected for applying the semantic noise addition model. It can be very useful in e-learning courses systems [10]. Such a task was realized based on language combinability theory and now such educational task generation systems are improved (Fig. 4). One of the next steps is to use second level patterns for language generation (see the table).



1. Заполнить пропуск в предложении

That physicist destroyes a ... transistor	medium
This ... creates a main project	engineer
This engineer operates a medium ...	Выберите правильный ва

2. Заполнить пропуск в слове

...rder - заказывать	Выберите правильный ва
m...chanism - механизм	Выберите правильный ва
f...llowing - следующий	Выберите правильный ва

3. Выбрать перевод слова на русский язык

digital	Выберите правильный ва
computer	Выберите правильный ва
device	Выберите правильный ва

4. Напечатать перевод слова на русский язык

engineer	Введите слово по-русски
display	Введите слово по-русски
sophisticated	Введите слово по-русски

5. Напечатать перевод слова на английский язык

радио усилитель	Введите слово по-английски
сложный	Введите слово по-английски
калькулятор	Введите слово по-английски

Выполнено заданий: 0 Правильно: 0 С ошибками: 0

Количество тест: 30

Fig. 4. Software for Automatic Educational Tasks Generation

In this way the two-levels patterns can be used for example for generating an educational task based on word combinability principles: “the user + takes + some wire” (Topic: Details of Equipment, Positions: Doer + Action + Receiver, Variants: Passive + Presentation-Passive + Metal-Long-Flexible); based on the phrase a task can be formed: “the user + takes / eats / wears / lives in + some wire”. The head of the task can be like: “select a proper word from the list” or “fill in the gaps with one of the variants, offered below”, etc. But further the decompression of the phrase can be made by adding semantic noise: “the user’s taking the wire was necessary” or “the wire after taking by the user was given to me” (see the table). Corresponding tasks could be made based on the decompressed sentences of the non-reduced form, for example, “the wire after taking / eating / wearing / living by the user was given to me”. The database of the second level semantic patterns should be created for the applications in e-learning linguistic software.

In the article the observation of multidimensional data by the natural language, particularly English, is given. It is possible to apply a multidimensional model of the natural language, semantic vectorized classification of words and notions of the natural language to make an assumption about the stylistic classification structure of the generative grammar rules set or relational word-sentences pattern sets, used for natural language generation. The structure of such rules has been analyzed. They can be used for text and sentences compression and decompression. The methods of using decompression by second level speech patterns for language generation are offered for improving linguistic software for automatic tasks generation for examples for the lessons of a foreign language. Further investigation in the sphere is necessary.

References

1. Verma A., Kumar A. Voice Fonts for Individuality Representation and Translation // ACM Translation on Speech and Language Processing. 2005. Vol. 2, № 1. Article 4.

2. Agamdjanova V. I. Contextual Redundancy of the Lexical Meaning of a Word. M. : Higher School, 1977.

3. Apresyan Yu. D. Ideas and Methods of Modern Structural Linguistics. M. : Science, 1966.

4. Evaluating Discourse Understanding in Spoken Dialogue Systems / R. Higashinaka, N. Miyazaki, M. Nakano, K. Aikawa. // ACM Translation on Speech and Language Processing. 2004. Vol. 1. P. 1–20.

5. Towards Efficient Human Machine Speech Communication: The Speech Graffiti Project / S. Tomko, T. K. Harris, A. Toth et al. // ACM Translation on Speech and Language Processing. 2005. Vol. 2, № 1. Article 2.

6. Lichargin D. V., Taranchuk E. A. Hierarchy Structure of Educational Electronic Course and its Variability for Teaching a Foreign Language // J. Distant and Virtual Education. 2011. № 4. P. 56–75.

7. Lichargin D. V., Bachurina E. P. Generalized Hierarchy Structure of the Educational Electronic Course and Viewing an Electronic Course of the English Language of RIYa IKIT SibFU Based on It // Informatization of Education and Science. 2012. № 3. P. 20–26.

8. Lichargin D. V., Sumaneeva Ya. A., Yuryeva E. V. Methods of Substitution Tables and its Application in the Sphere of Teaching Russian Language to the Foreigners // Bulletin of the Surgut State Pedagogical University. 2012. № 6.

9. Sdobnikov V. V. New view on the strategy of translation: communicative-functional approach // J. of SibFU. Krasnoyarsk, 2011. Vol. 4, № 10. P. 1444–1453.

10. Teaching Mathematics in the Moodle Environment by example of an electronic learning course / T. V. Zykova, A. A. Kytmanov, G. M. Tsibulskiy, V. A. Shershneva // The Bulletin of Krasnoyarsk State Pedagogical University named after V. P. Astafiev. 2012. № 1. P. 60–63.

Д. В. Личаргин, К. В. Сафонов, Н. В. Николаева, Е. Б. Чубарева

ПРИНЦИПЫ КОМПРЕССИИ И ДЕКОМПРЕССИИ В РАМКАХ МНОГОМЕРНОГО ПРЕДСТАВЛЕНИЯ ЕСТЕСТВЕННОГО ЯЗЫКА В ПРИМЕНЕНИИ К ГЕНЕРАЦИИ УЧЕБНЫХ ЗАДАНИЙ

В данной работе рассматривается проблема добавления шаблонов семантической декомпрессии к осмысленным предложениям, порождаемым в виде функций на многомерном векторном пространстве понятий естественного языка. Предлагается модель добавления шаблонов семантической декомпрессии, на основе стилистически выделенных множеств правил порождающей грамматики. Рассматривается сложность проблемы перевода естественных языков. Рассмотренная модель обеспечивает выполнение алгоритмов добавления шаблонов семантической декомпрессии, которые могут быть использованы для улучшения производительности программного обеспечения, предназначенного для генерации учебных заданий.

Ключевые слова: порождение естественного языка, генерация осмысленных предложений, тест Тьюринга.