УДК 519.682

# PRINCIPLES OF SEMANTIC NOISE ADDITION WITHIN MULTIDIMENSIONAL NATURAL LANGUAGE GENERATION REPRESENTATION

**Narytto A.G., Lichargi D.V., Nikolaeva N.V.**
**Scientific supervisor – Associate professor Lichargi D.V., Nikolaeva N.V.**

***Siberian Federal University***

In the given work the problem of semantic noise 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 noise based on stylistically-oriented sets of generative grammar rules is offered. The considered model provides the algorithm of adding semantic noise to the sentences of the natural language. The conclusion of semantic variety of the natural language and the necessity of considering it in the systems of natural language generations is made.

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.

There is an acute 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.

In the works 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:

• 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, …).

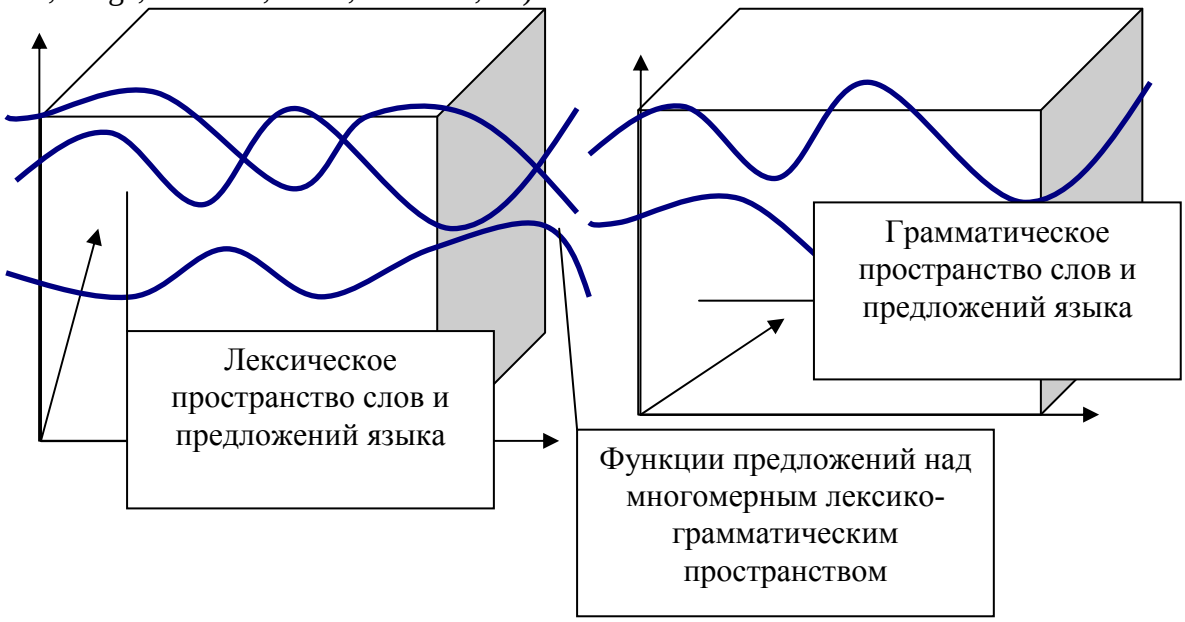


Figure 1.The Lexical and Grammatical Space
in the Model of Meaningful Language Generation

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 noise. Semantic noise is present in the texts written in different language styles: from the academic style with a low degree of semantic noise up to slang with an extremely high degree of semantic noise. Semantic noise is 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.

Table 1 – Sections of the Viewed Multidimensional Space as Tables of Speech Generation by «Lulia-Palmer» Method

| | | | | |
|---|---|---|---|---|
| I<br>***я*** | wish to<br>***желать*** | take<br>***занять*** | the<br>***этот*** | medal<br>***медаль*** |
| we<br>***мы*** | want to<br>***хотеть*** | win<br>***выиграть*** | a(n)<br>***некоторый*** | golden medal<br>***золотая медаль*** |
| you<br>***вы*** | decided to<br>***решил*** | take<br>***взять*** | this<br>***этот*** | silver medal<br>***серебренная медаль*** |
| they<br>***они*** | happen to<br>***случилось*** | gain<br>***завоевать*** | that<br>***тот*** | copper medal<br>***бронзовая медаль*** |
| | need to<br>***нуждаться*** | award<br>***присудить*** | one<br>***один*** | diploma<br>***грамота*** |

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

[RELATION-SUBJECT -X \ ESSENCE \ POSITIVENESS]

The word «beautiful» corresponds to the vector attributes:

[RELATION-X \ ESSENCE \\ RELATION-SUBJECT-X \ IDEA \ ON (NOT) LIVE \ POSITIVENESS].

The word «to look» corresponds to the vector of semantic features:

[RELATION-SUBJECT-X \ ESSENCE \\ RELATION-SUBJECT –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.

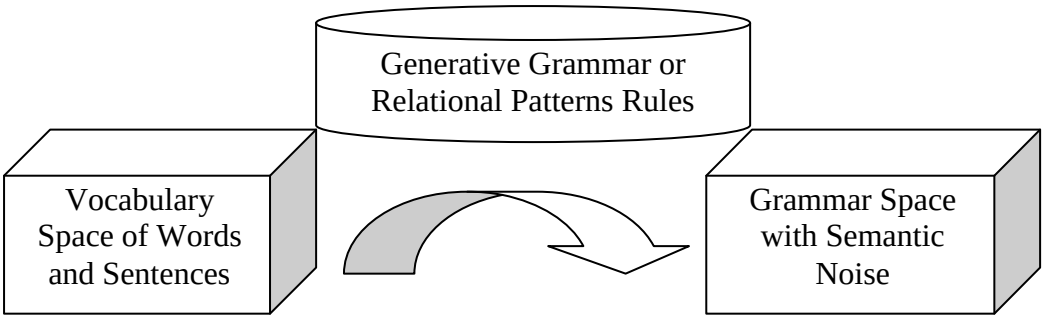


Figure 2.The principle of Lexical Space – Grammatical Space transformation

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 to…»).

Table 1

Methods of Pattern-Based Semantic Noise 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 |

The generative grammar or relational patterns, used for sentences / text composiotion and decompression and semantic noise 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;
5. Neutral Style;
5.1. Journalistic Style;
6. Mixed Style.

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 structure of the generative grammar rules or relational pattern, used for natural language generation. The structure of such rules has been analyzed. They can be used for text and sentences compression and decompression. Further investigation in the sphere is necessary.