ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2005, Vol. 34, No. 3

FORMAL DESCRIPTION OF THE SYNTAX OF THE LITHUANIAN LANGUAGE

Daiva Šveikauskienė

Institute of Mathematics and Informatics
Goštauto st. 12-209, LT-01108 Vilnius, Lithuania

Abstract. Artificial intelligence is a type of the software, capable of performing the mental work of a man. Machine translation belongs to the very same sphere. The first stage of the machine translation is the grammatical analysis of a sentence. The article presents the methods of how the syntactic analysis of a Lithuanian sentence should be performed by a computer. In the course of the analysis, very specific features of the Lithuanian language, namely, its great inflexion and the free word order in a sentence, should be taken into account. For the purpose of alleviating the tasks of the programming, the syntactic rules of the Lithuanian language are written in the BNF (Backus-Naur Form), following the formal rules of the context-free grammar (if we bear in mind the rules of N. Chomsky formal grammar classification).

Keywords: artificial intelligence, natural languages processing, machine translation, automatic syntactic analysis.

Indroduction

Nowadays a great number of the systems of machine translation have been created to serve the needs of many languages. The Lithuanian language cannot pride itself on having its own system of the machine translation yet. The main reason to be mentioned might be the following: the Lithuanian language has not been sufficiently prepared, i.e., sufficiently formalized to be accessible for the purposes of the computerized usage.

If we choose to remember the already completed works dedicated to the task of the formalization of the Lithuanian language, the lemmatizing created by V. Zinkevicius should be the first to be mentioned. It could be used to serve the first stage of the system of machine translation, which would be the stage of the morphological analysis of a Lithuanian text. The program mentioned above could be also used while generating the sentence, which had been already translated into the Lithuanian language during the stage of the morphological synthesis.

The second stage of the machine translation is that of the syntactic analysis. The automatic syntactic analysis of the Lithuanian language has not been prepared yet, which fact precludes the creation of the system of the machine translation of that language. That is why this work attempts to present the formalized description of the syntax of the Lithuanian language, with the view to this description forming the basis for the automatic syntactical analysis.

The already created systems of the syntactic analyses, which serve the needs of other languages, could be of little use when the needs of the Lithuanian language are considered. The differences between the Lithuanian language and other Indo-European languages, which have been using their own systems of machine translation already, are too big.

This work attempts to evaluate the specific qualities of the Lithuanian language – its great inflexion and the free word order in a sentence. The work also aspires to create the methodology, enabling a good quality automatic syntactical analysis of the Lithuanian sentences to be performed.

The new in the work is the consideration of the specificity of the Lithuanian language. The syntactical functions are differentiated in accordance with the morphological categories of words. Attention is paid to a very great inflexion of the Lithuanian language. For the purposes of the improvement of the results of the syntactical analysis of other languages usually the semantics of the words in those languages is made recourse to. For that purpose nobody uses morphological data, which would mean centering the attention on the flexions of words. At least the author of this work is not familiar with any literary source, describing the morphological methodology of the analysis. The Lithuanian language presents a contrast to the general pattern. The other very specific feature of the Lithuanian language, which is its free word

order in a sentence, is evaluated with the help of the formal parameter THREAD, which determines the word order of the syntactically linked words in a sentence with regard to each other as well as with regard to the words which do not belong to that link.

The work should be of service when creating the system of the machine translation of the Lithuanian language. Speaking more precisely, this work should be of a great help when preparing the stage of the syntactic analysis of that language.

1. History

The very first ideas pertaining to the mechanization of translation were put down in the seventeenth century. In 1629, Rene Descartes suggested writing the books which would be made up of ciphers, whereas in the dictionaries the corresponding words in all the languages would be given the same code number. In J.J. Becher's dictionary, which appeared in 1661, 10,000 of Latin words were provided with coding, but finding the equivalents of the very same words in the Greek, Hebrew, German, French, Slavonic and Arab languages turned out to be not such a simple thing.

It was in 1933, when the idea of creating the "translating machines" was suggested for the first time. At that time the patents, securing the rights of mechanical dictionaries, appeared both in France and Russia. The French engineer George Artsrouni planned to use the paper bands for the purpose of finding the equivalent of a word in another language. A Russian P.Smirnov-Trojanski envisaged the three stages of the mechanical translation [4]:

1. Editor, knowing the source language exclusively, had to perform the "logical" analysis of the words, indicating their base forms, which would be the Nominative case of nouns and the Infinitive of verbs, etc. The syntactical functions of the words, that is, Subject, Predicate, Object, etc., had also to be indicated.
2. The sequence of the base forms and their syntactical functions in the source language had to be changed by a machine into an equivalent sequence in the target language.
3. Editor, knowing only the target language, had to change the disparate words produced by a machine into the requisite forms of that language.

The patent covered only the machine-performed operations of the second stage.

When in 1942 the first computer MARK I [10] was created in Harvard University, there arose a new possibility of automatically performing the functions, which used to be performed by man exclusively, namely, translations. With the appearance of computers it was recognized that they could work not only with numbers but with other symbols, such as letters, too. Consequently, computers could work with languages. A number of linguists started creating descriptions of natural languages with the formality and precision needed for computer implementation [14]. While enumerating the ways in which the computers could demonstrate their "intelligence" Alan Turing mentioned language translation as its third possibility [12]. In 1947, it was Warren Weaver who was the first to raise the idea of using computers for the purposes of translation: "I have wondered if it were unthinkable to design a computer which would translate [4]. It was round 1985 when the common consensus was reached regarding the absence of any possibility of having the constantly growing number of texts translated unless the translation work were performed automatically [10].

2. Machine translation systems

In modern systems of machine translation the work is divided into three phases: namely, analysis, transfer and synthesis [5].

The block scheme of such a system is presented in Figure 1.

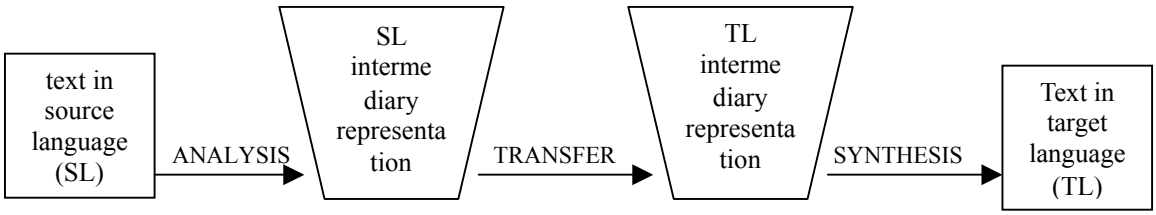

Figure 1. Machine translation system

In the course of the analysis the text of the source language gets transformed into an intermediary representation in three stages:

a) morphological analysis,
b) syntactic analysis,
c) semantic analysis.

All the stages of machine translation are presented in Figure 2.

In the course of the morphological analysis, every word is given in its initial form (Infinitive, Nominative case, etc.). The morphological information about every word in a sentence, such as its gender, number, case, person etc., is also presented. The morphological analysis of the Lithuanian language can be performed by usage of the lemmatizer, created by V. Zinkevičius.

While performing the syntactic analysis, the syntactic functions of the words are determined and their links are indicated. No system of the syntactic analysis of the Lithuanian language has been created yet.

ANALYSIS	Morphological Syntactic Semantic
TRANSFER	Word Syntactic structure
SYNTHESIS	Semantic Syntactic Morphological

Figure 2. Stages of machine translation

The semantic analysis is used to improve the results of the syntactic analysis. In the course of the semantic analysis, certain signs of the meaning of the words are indicated. In certain cases, with the help of these signs, the ambiguity of the syntactic structure of a sentence can be destroyed. Besides, the signs of meaning are useful by choose the words in the target language [9]. For example, for the purpose of determining the best word among the many possible choices, the following signs are important:

a) It is important to ascertain whether the object is alive or not. This feature can be very useful when translating relative and interrogative pronouns into Russian (*что, кто*). Consequently, in the case mentioned above all the nouns of the source language should be divided into two semantic groups – live or not live objects [18].
b) The fact whether the object is a countable or non countable noun can be important when translating the Lithuanian word *daug* into English — *much time*, *many books*, etc.

The system of the semantic analysis of the Lithuanian language has not been created yet either. When formalizing the syntax of the Lithuanian language, certain semantic features are used as the constituent parts of the syntactic analysis.

The second stage of the machine translation aims at linking the two languages, i.e., translating one into the other. In the course of the translation, the intermediary representation of the source language is changed into the corresponding intermediary representation of the target language. The transfer is performed along the two levels:

1. The words of one language are changed into the words of the other language.
2. The syntactic structure of a sentence in one language gets changed into the syntactic structure of the sentence in the other language. Very often the syntactic structures in both the languages are very different, that is why without the syntactic analysis of the sentence one cannot expect to get a good quality translation. For example, while translating the sentence *He likes this book* into the Lithuanian language the structural scheme of the verb *like* would be transformed in the manner as shown in Figure 3.

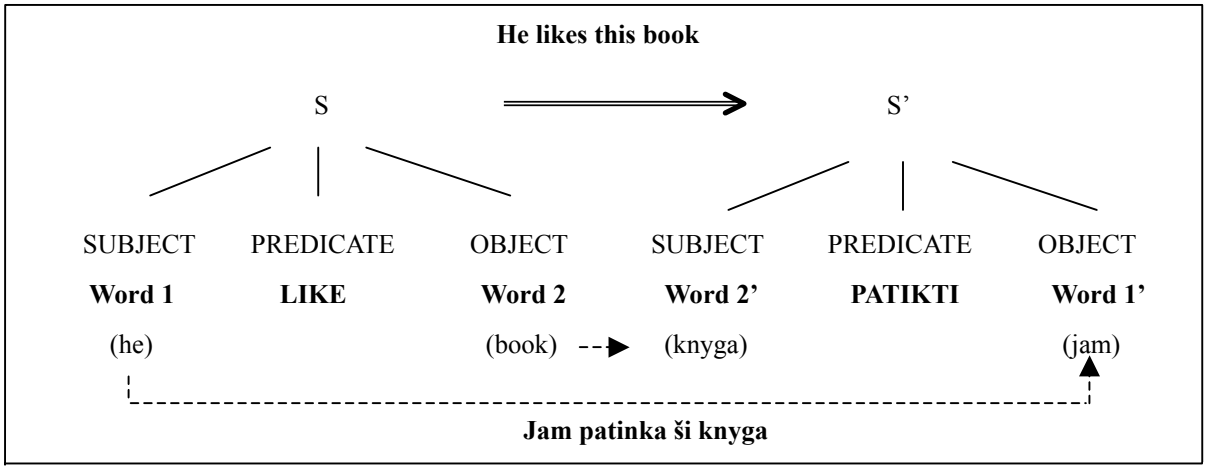

Figure 3. The change of the structural scheme of the word *like* by the structural scheme of the word *patikti*

3. Syntactic structure of a sentence

The syntactical structure of a sentence demonstrates that words are interconnected. The widest spread method of demonstrating the structure of a sentence is a graph or, to be more precise, a tree [1]. Trees, presented by linguists, are usually drawn with their tops downwards; i.e., their roots are on the top, and their leaves are at the bottom [2]. Linguists are familiar with two very different ways of drawing trees. They are the phrase-based method and the method of dependency. The phrase-based method is more applicable to the languages, which can be characterized by a strict word order in a sentence. As an example, the English language can be indicated. The tree of dependency is

more convenient in demonstrating sentences of those languages, which can be characterized by having a free word order. When dealing with the Lithuanian language, the latter method is usually chosen because the word order of a Lithuanian sentence is free; that is, the Lithuanian language cannot be characterized as having any part of a sentence require a fixed position in respect of the beginning or the end of a sentence. In contrast, we could mention the German language, where the predicate in a direct sentence requires the second place in a sentence; in questions, the German predicate should occupy the first place, and in subordinate clauses the German predicate should be placed at the very end of the sentence. By contrast, in the Lithuanian language any part of a sentence can be placed either at the beginning or in the middle or at the end of a sentence.

The finite verb is placed at the root of the dependency tree. The words modifying the meaning of the verb are placed below. For example, the tree of dependency of the sentence *Jonas valgo raudoną obuolį* (*John eats a red apple*) would be drawn in the manner shown in (Figure 4):

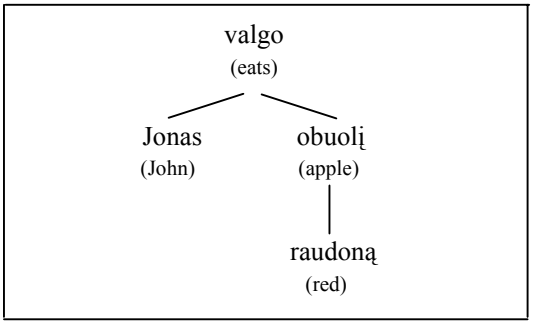

Figure 4. The dependency tree of the sentence *Jonas valgo raudoną obuolį (John eats a red apple)*

The above drawn structure is not linked with any word order in a sentence. This structure is useful in the transfer phase of machine translation systems. While translating, the original word order could be ignored. Basing the results of the translation on the links of the words in the dependency tree, one can form the translated sentence in accordance with the word order characteristic of the target language [6].

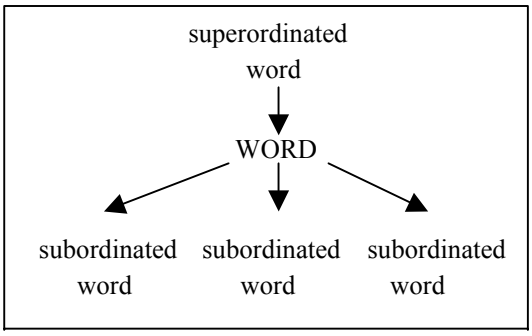

Figure 5. The links of the node of the dependency tree with adjacent nodes

A generalized structure of the node in the dependency tree is shown in Figure 5. Every node of the dependency tree is occupied by a word, which can have one or more subordinated words and only one superordinated word.

The task of the syntactic analysis is to find for the every word of the sentence all the subordinated words and the superordinated word [3]. Bearing in mind that the tree cannot reflect all the syntactical information that a Lithuanian sentence possesses, it is usually the graph which can best demonstrate the structure of Lithuanian sentences [13].

4. Syntactic analysis

Judging by the peculiarities of the syntactic analysis stage, the machine translation systems can be divided into three types [17]:

1. Systems, which have no syntactic analysis.
2. Systems, which single out the stage of the syntactic analysis, usually preceded by the morphological analysis, closely followed up by the semantic operations aiming at improving the results of the syntactic analysis.
3. Systems, where both the syntactic and semantic analyses are united in a complex procedure, and where the syntactic and semantic interaction assures a fast and good quality analysis.

The direct translation systems belong to the first type. In cases of direct translation, the text of the source language is directly reflected in the translated text whose quality is usually very bad. For example, the Russian sentence *Вчера мы целый час катались на лодке* gets translated into English as *Yesterday we the entire hour rolled themselves on a boat*, whereas the correct translation should be *Yesterday we went out boating for a whole hour* [5].

The quality of the systems of the second type guarantees much better results, but even in these cases we can get too many superfluous variants, i.e., there remains too much of the ambiguity of the syntactic structures. For example, syntactic structures of a sentence often do not have one meaning because of the usage of the multiple constructions of prepositions. Concrete examples can be furnished by literature [7]. The following sentences *The coastguard observed the yacht in the harbor with binoculars* and *The gold watch was sold by the jeweler to a man with a beard* can be characterized by the ambiguity. Syntactical means cannot determine which word the proposition *with* is linked with. In the first sentence, the phrase *with binoculars* is linked with the verb *observe*, whereas in the second sentence the phrase *with a beard* denotes a noun *man*. This interdependency can be determined only with the help of the semantic information, which assign to link the word *binoculars*, possessing the quality of an 'instrument', with the verbs indicating the activity and perception of a man, such as the verb *observe* connotes. In conjunction with

the same kind of semantic information, the word *beard* cannot be taken to be an object completing the verb *sell*. Consequently, in the newest systems of automatic translation syntax and semantics form a unity.

While performing the syntactic analysis of the Lithuanian language, all the three stages indicated as the second point are joined into one. Before the performance of the syntactic analysis which is united with the morphology, some of the semantic features of a sentence are also indicated.

5. Some specific features of the syntactic analysis of the Lithuanian language

The syntactic analysis of the Lithuanian language should be performed while bearing in mind the specific characteristics of the Lithuanian language, which are a great inflexion and a free word order in a sentence. While determining the parts of a sentence in the English language, the morphology, i.e., word flexions will play no role in this quest. The main factor, helping the researcher to determine the parts of an English sentence, is the word order. In the Lithuanian language, though, syntactical links among the words are mostly indicated by the flexions of the words [8]. Consequently, when performing the syntactic analysis of the Lithuanian language, one cannot rely on the word order. The main weight of the syntactical information is usually born by multiple flexions of the words in a sentence, and all the manifold information should be evaluated. That is why in the course of the formal description of the syntax of the Lithuanian language, all the parts of the sentence are differentiated in accordance with the morphological categories of the words which can carry out the above mentioned syntactical functions. For example, it would not be sufficient to indicate, that a subject is expressed with a noun. The case, number and gender of that noun should also be registered. In consequence, the example of the description of a subject BNF might be the following:

<SUBJECT-NOUN-NOMINATIV-SINGULAR-FEMININUM>::=
noun_nominative_singular_feminine;

This description would indicate a subject, expressed by a noun in the nominative case, singular, and feminine in gender. Then the agreeing attribute, which agrees with subject, mentioned above, should also be found in accordance with all the requisite morphological categories. The attribute will also be described in the same manner, indicating all the morphological categories of an adjective or a participle: nominative case, feminine in gender and singular in form

<AGREEING-ATRIBUT-ADJECT-NOMINAT-SING- FEMIN> ::=
adjectiv_nominative_singular_feminine;

The method given above differs greatly from the strategy of the systems of the automatic syntactic analyses, which have been already created. In the already created systems only semantics (the meaning of the word) was used for the purpose of overcoming the syntactical ambiguity, whereas morphology remained unheeded.

6. The description of the syntactic rules of the Lithuanian language in BNF

The formal description of the rules of the syntax of the Lithuanian language given in BNF consists of two parts. The first part offers the description of the correspondence of the syntactical functions and morphological categories, that is, every syntactical function bears an indication of the morphological categories, which can perform that function. In the structure of a sentence, that correspondence would be reflected at the nodes of a graph. The second part denotes syntactical links, that is, the arcs in the graph, which connect those nodes. It is here that the free word order of Lithuanian sentences gets evaluated.

While describing the nods of a graph, first all the syntactical functions are made dependent on the parts of the speech, which are able to perform these functions. Later every part of the speech is divided into categories depending on its morphological functions. For example, the description of the subject bears an indication the subject may be expressed by a noun, by a pronoun, or by an infinitive form of a verb. Later, the subject expressed by a noun is divided into the following categories: a subject expressed by a noun in the nominative case, masculine in gender and singular in form or a subject expressed by a noun in the nominative case, feminine in gender and singular in form, etc. The subject, which is expressed by the infinitive form of a verb, is defined by the valence of the verb, that is, the infinitive which does not require any noun in any case, the infinitive which has to be accompanied by a noun in the genitive case, the infinitive which requires a noun expressed in the dative case, accusative case, and so on and so forth. The cases, demanded by a verb are marked in an inclined print, and they are considered to be notional features, similar to the semantic features, such as time feature for nouns. Depending on the semantic features of the words, one can decide which of the syntactical functions morphological forms can be alluded to. For example, the accusative case of a noun usually indicates an object (*dainuoti <u>dainą</u> — to sin the <u>song</u>*), but the accusative case indicating the time performs the function of the adverbial modifier of time (*dainuoti <u>naktį</u> — to sing <u>at night</u>*). The adjectival pronouns and the pronouns, which can be used instead of a noun are marked as *A* and *N* in formal description. This information belongs to the semantic features too.

Morphological categories are presented as terminal symbols in the formal description. The description of a subject in the BNF acquires the following form:

<SUBJECT> ::= <SUB-NOUN> | <SUB-PRON-*N*> | <SUB-INF>;

<SUB-NOUN> ::= <SUBJ-NOUN-NOM-SING-MASC> |
<SUBJ-NOUN-NOM-SING-FEM> |
<SUBJ-NOUN-NOM-PLUR-MASC> |

```
                    <SUBJ-NOUN-NOM-PLUR-FEM>;
<SUB-PRON-N> ::=    <SUB-PRON-NOM-SING-MASC-N> |
                    <SUB-PRON-NOM-SING-FEM-N> |
                    <SUB-PRON-NOM-PLUR-MASC-N> |
                    <SUB-PRON-NOM-PLUR-FEM-N> |
                    <SUB-PRON-NEUTR>;
<SUB-INF> ::=       <SUB-INFINITIVE> |
                    < SUB-INFINITIVE-GENIT> |
                    < SUB-INFINITIVE-DAT> |
                    < SUB-INFINITIVE-ACC> |
                    < SUB-INFINITIVE-INSTR> |
                    < SUB-INFINITIVE-LOC> |
<SUBJ-NOUN-NOM-SING-MASC> ::= noun_nom_sing_masc;
<SUBJ-NOUN-NOM-SING-FEM> ::= noun_nom_sing_fem;
<SUBJ-NOUN-NOM-PLUR-MASC> ::= noun_nom_plur_masc;
<SUBJ-NOUN-NOM-PLUR-FEM> ::= noun_nom_plur_fem;
<SUB-PRON-NOM-SING-MASC-N> ::= pron_nom_sing_masc_n;
<SUB-PRON-NOM-SING-FEM-N> ::= pron_nom_sing_fem_n;
<SUB-PRON-NOM-PLUR-MASC-N> ::= pron_nom_plur_masc_n;
<SUB-PRON-NOM-PLUR-FEM-N> ::= pron_nom_sing_fem_n;
<SUB-PRON-NEUTR> ::= pron_neutr;
<SUB-INFINITIVE> ::= inf;
< SUB-INFINITIVE-GENIT> ::= inf_genit;
< SUB-INFINITIVE-DAT> ::= inf_dat;
< SUB-INFINITIVE-ACC> ::= inf_acc;
< SUB-INFINITIVE-INSTR> ::= inf-instr;
< SUB-INFINITIVE-LOC> ::= inf_loc;
```

While describing the arcs of a graph, that is, the syntactical links among words, a formal parameter, named THREAD, is used. This THREAD should be able to take care of the free word order in the Lithuanian language, that is, it should be able to link the tree of dependency with the linear arrangement of words in a sentence. The description of THREAD in the right-hand side of the BNF has three positions. In the first and the third positions are placed the parts of the sentence among which the syntactical link is being sought for. The middle position is the non-terminal symbol, which is called the INSERTION between the parts of the sentence, which are being described.

```
<THREAD#SUBJECT+AGREEING-ATTRIBUTE> ::=
            <AGREEING-ATTRIBUTE>
 [{<INSERTION-BETWEEN-SUB-&-AGREEING-ATTR>}]
            <SUBJECT> |
            <SUBJECT>
 [{<INSERTION-BETWEEN-SUB-&-AGREEING-ATTR>}]
            <AGREEING-ATTRIBUTE>;
```

7. Word order in a Lithuanian sentence

The insertion should evaluate the free word order in a Lithuanian sentence, id est. it should indicate which differing parts of the speech might enter the space between the two words linked into a direct syntactical union. The word order in the Lithuanian language is free only in a sentence. Word collocations might be governed by certain rules, which might not have been discussed by Lithuanian linguists. For example, a non-agreeing attribute cannot occupy a position in between a subject and another non- agreeing attribute, because in this manner the second non-agreeing attribute would destroy the union of a subject and the first non-agreeing attribute. For example, the union *mano namas* (*my house*) will admit only an agreeing attribute, such as *senas* (*old*), which will not affect the initial union: *mano senas namas* (*my old house*) will remain *mano namas* (*my house*), anyway (Figure 6). The new collocation *senas namas* (*old house*) does not destroy the first collocation. In a sentence the new collocation stands next to the old, that is, in the sentence instead of the initial first collocation *mano namas* (*my house*) we have two collocations *mano namas* (*my house*) and *senas namas* (*an old house*). Consequently, the initial collocation remains, it only gets complemented by an additional collocation.

If on the other hand, the word *brolio* (*brother's*) intervenes in between the words *mano namas* (*my house*), the first word collocation gets destroyed — the house of my brother is not my house (Figure 7).

When the word *brolio* (*brother's*) intervenes in the first collocation we get two very different collocations instead the initial collocation: *mano brolio* (*my brother's*) and *brolio namas* (*brother's house* / the house of my brother) (Figure 8).

Consequently, the description of BNF should bear an indication that the INSERTION in between a subject and a non-agreeing attribute cannot be another non-agreeing attribute. This INSERTION can only be an agreeing attribute or a THREAD of that attribute, that is an agreeing attribute accompanied with the words which modifies it, for example, *mano labai senas namas* (*my very old house*) (Figure 9).

In the description of BNF the above given information should be reflected in the following manner:

```
<THREAD#SUBJECT+NONAGREEING-ATTRIBUTE> ::=
      <NONAGREEING-ATTRIBUTE>
[{<INSERTION-BETWEEN-SUB-&-NONAGR-ATTR>}]
      <SUBJECT> ;
< INSERTION#BETWEEN-SUB-&-NONAGR-ATTR> ::=
      <AGREEING-ATTRIBUTE-OF-THE -SUBJECT> |
<THREAD# AGREEING-ATTRI-OF-THE-SUBJECT+MODIF>;
```

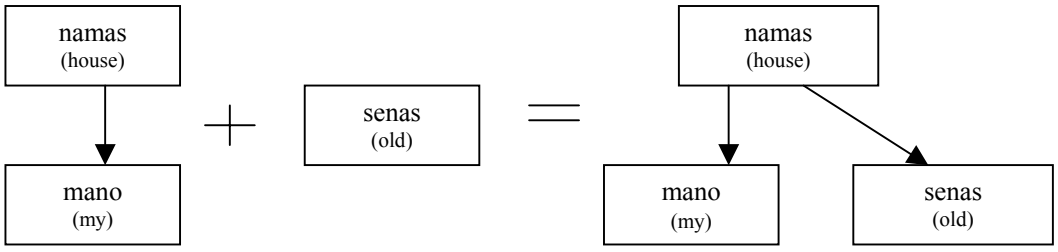

Figure 6. The interference of the agreeing attribute into the word collocation *mano namas* (*my house*)

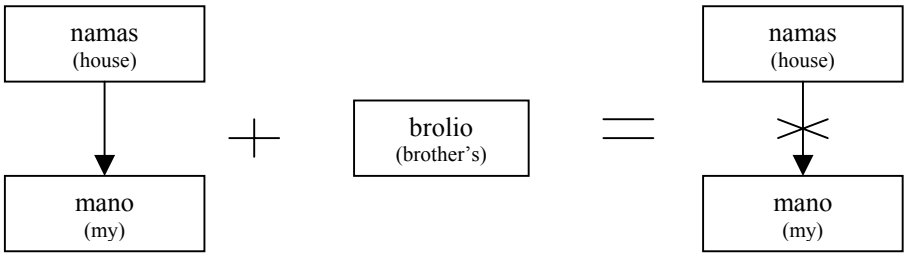

Figure 7. The interference of the non-agreeing attribute into the word collocation *mano namas* (*my house*)

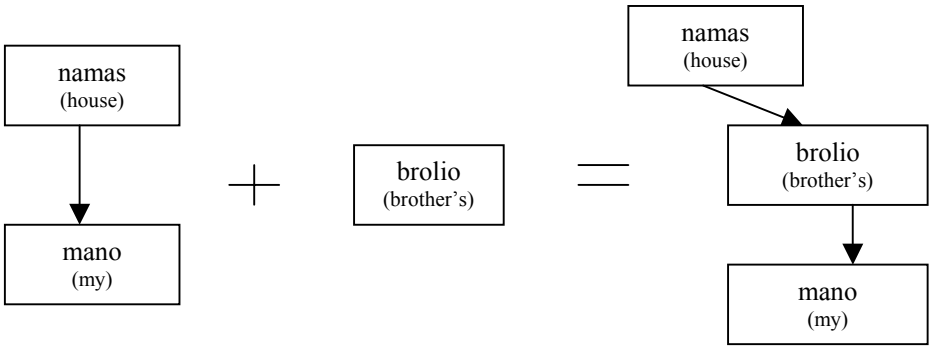

Figure 8. The formation of new word collocations

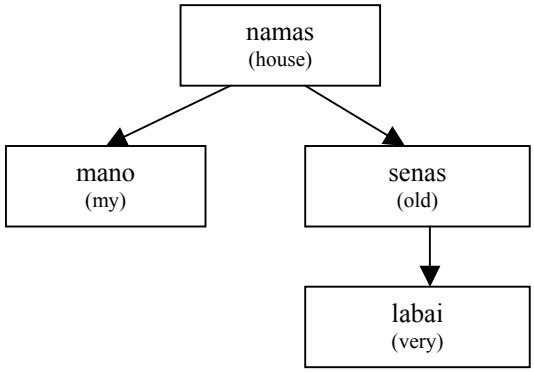

Figure 9. Insertion expressed by a THREAD of an agreeing attribute *labai senas* (*very old*)

<THREAD#SUBJECT-**NOUN-NOM-SING-MASC**+AGREEING-ATTRIBUTE-**ADJ-NOM-SING-MASC**> ::=
<AGREEING-ATTRIBUTE-**ADJ-NOM-SING-MASC**>
[{<INSERTION#BETWEEN-SUB-**NOUN-NOM-SING-MASC**-&-AGREEING-ATTR-**ADJ-NOM-SING-MASC** >}]
<SUBJECT-**NOUN-NOM-SING-MASC** >;

<THREAD#SUBJECT-**NOUN-NOM-SING-FEM**+AGREEING-ATTRIBUTE-**ADJ-NOM-SING-FEM**> ::=
<AGREEING-ATTRIBUTE-**ADJ-NOM-SING-FEM**>
[{<INSERTION#BETWEEN-SUB-**NOUN-NOM-SING-FEM**-&-AGREEING-ATTR-**ADJ-NOM-SING-FEM** >}]
<SUBJECT-**NOUN-NOM-SING-FEM** >;

<THREAD#SUBJECT-**NOUN-NOM-PLUR-MASC**+AGREEING-ATTRIBUTE-**ADJ-NOM-PLUR-MASC**> ::=
<AGREEING-ATTRIBUTE-**ADJ-NOM-PLUR-MASC**>
[{<INSERTION#BETWEEN-SUB-**NOUN-NOM-PLUR-MASC**-&-AGREEING-ATTR-**ADJ-NOM-PLUR-MASC** >}]
<SUBJECT-**NOUN-NOM-PLUR-MASC** >;

<THREAD#SUBJECT-**NOUN-NOM-PLUR-FEM**+AGREEING-ATTRIBUTE-**ADJ-NOM-PLUR-FEM**> ::=
<AGREEING-ATTRIBUTE-**ADJ-NOM-PLUR-FEM**>
[{<INSERTION#BETWEEN-SUB-**NOUN-NOM-PLUR-FEM**-&-AGREEING-ATTR-**ADJ-NOM-PLUR-FEM** >}]
<SUBJECT-**NOUN-NOM-PLUR-FEM** >;

Figure 10. BNF Description of the subject by usage of morphology

The THREAD is also described by usage of morphology. For example, the THREAD between the subject and its agreeing attribute is presented in the manner as shown in Figure 10, when divided in respect to morphological categories.

8. Examples of the analyses of sentences

The syntactic analysis of a sentence signifies the search for word collocations. The end result is bound to be the list of the words linked by direct syntactical contacts. The elements of that list are word pairs. Every word of a pair has its syntactical function defined, and the form of the syntactical link between these two words is also presented. Later the graph of dependency is formed on the basis of the list mentioned above.

The following example will illustrate the very process of the analysis. To make matters clear, a very simple set of BNF rules is made up for the purpose of enabling a researcher to analyse only a very small group of sentences (Figure 11). The initial symbol is taken to be the symbol *S*, which means a sentence. This sentence gets divided into the subject group *V* and the predicate group *T*. Sometimes a sentence may consist only of one group, that is, of a subject group or of a predicate group. The subject group *V* is expressed by the subject *v* or by the subject with modifying words, i.e. by a THREAD *N,* which consist of a non-agreeing attribute *n*, followed by a subject *v*. Sometimes an INSERTION *B* may stand in between them. This INSERTION *B* signifies an agreeing attribute *d* or the THREAD *D* of an agreeing attribute. The THREAD *D* is expressed by the agreeing attribute *d* preceded by an adverbial modifier *a*, or by the THREAD *A* of an adverbial modifier. The THREAD *A* can consist only of two adverbial modifiers *a*, which follow each other. The predicate group *T* is expressed by the predicate *t* or by the THREAD *C,* which denotes the links between the predicate and modifiers. All possible cases of word order are indicated.

In accordance with the collection of rules in Figure 11 we can analyse the following sentence: *Mano labai seniai statytas namas ir šiandien atrodo visai gražiai* (*My house built long ago looks very nice even now*) The schema of analyses is presented in Figure 12.

```
S ::= V[T] | T[V];          *V – group of the subject, T – group of the predicate*
V ::= v | N;                *N – THREAD# subject + non-agreeing attribute*
N ::= nv | nBv;             *B – INSERTION# between subject & attribute*
B ::= d | D;                *D – THREAD# attribute +modifier*
D ::= ad | Ad;              *A – THREAD# modifier + modifier*
A ::= aa;
T ::= t | C                 *C – THREAD# predicate + modifier*
C ::= atA | AtA | Ata | ata
```

Figure 11. An example of the set of BNF rules

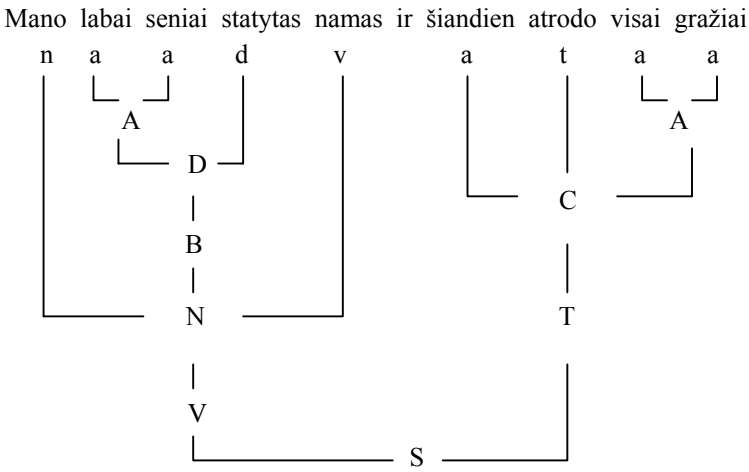

Figure 12. An example of a syntactically analysed sentence Mano labai seniai statytas namas ir šiandien atrodo visai gražiai (My house built long ago looks very nice even now)

Having chosen a morphologically ambiguous word, for example *sakai* (*utter*, singular, second person; and *resin*), we can observe how in the course of the syntactical analysis the ambiguity of a word gets destroyed. We can choose separate sentences to illustrate different meanings of the same word: *Tamsūs pušų sakai blizgėjo saulėje* (*The dark resin of the pine trees was glistening in the sun*), and *Per tyliai sakai tuos reikšmingus žodžius* (*You are uttering these important words in too low a voice*). The syntactic structures of those sentences should look as in Figure 13 and in Figure 14.

The arcs connecting the nods of the graph, that is, the syntactical links among the words, can also be demonstrated in the linear structure of the sentence, i.e., in the very same sentence which we see written, in the manner as shown in Figure 15 and Figure 16.

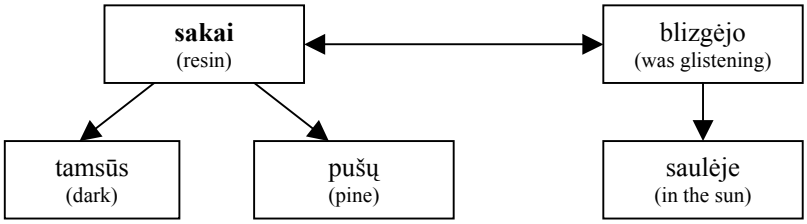

Figure 13 The syntactic structure of the sentence *Tamsūs pušų sakai blizgėjo saulėje* (*The dark resin of the pine trees was glistening in the sun*)

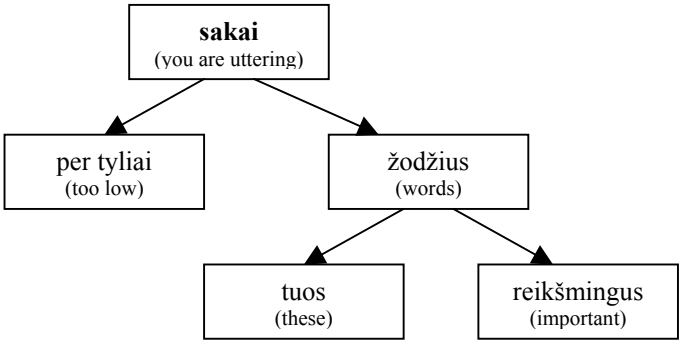

Figure 14 The syntactic structure of the sentence *Per tyliai sakai tuos reikšmingus žodžius* (*You are uttering these important words in too low a voice*)

The arcs connecting the nods of the graph, that is, the syntactical links among the words, can also be demonstrated in the linear structure of the sentence, i.e., in the very same sentence which we see written, in the manner as shown in Figure 15 and Figure 16.

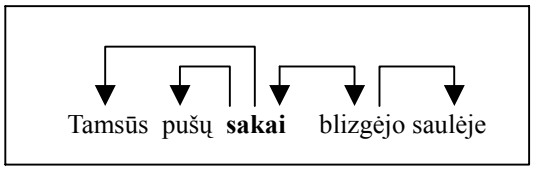

Figure 15. Syntactical links among words shown in the linear structure of a sentence *Tamsūs pušų sakai blizgėjo saulėje* (*The dark resin of the pine trees was glistening in the sun*)

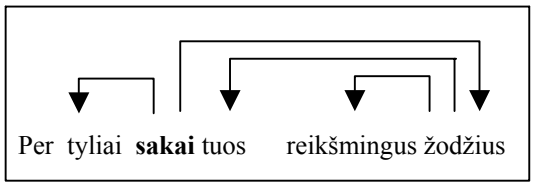

Figure 16. Syntactical links among words shown in the linear structure of a sentence *Per tyliai sakai tuos reikšmingus žodžius* (*You are uttering these important words in too low a voice*)

The non-terminal symbol THREAD in BNF description corresponds to the arrows placed over the words of the sentences in Figure 15 and Figure 16.

The syntactical analysis of the sentences mentioned above starts with the morphological information given about every word in a sentence (shown in the first line over the words of the sentence, Figure 17). The morphological analysis will be performed with the help of the lemmatizing program, created by V. Zinkevičius. The arrows point out the way, how syntactical categories follow the morphological ones. The allotting of the function to a word starts from bottom, i.e., from the morphological categories of a word (from terminal symbols in the BNF description). The subject in the sentence *Tamsūs pušų sakai blizgėjo saulėje* (*The dark resin of the pine trees was glistening in the sun*), is determined as shown in Figure 17.

The THREAD between the subject *sakai* (*resin*) and the agreeing attribute *tamsūs* (*dark*) occupies the positions of three words. The non-agreeing attribute is the INSERTION. The information is reflected in the BNF description of this INSERTION: the non-agreeing attribute can be placed in between the subject and the agreeing attribute.

<INSERTION#BETWEEN-SUB-&-AGREEING-ATTR> ::=
<AGREEING-ATTRIBUTE> |
<THREAD#AGREEING-ATTRIBUTE + MODIFIER> |
<**NONAGREEING-ATTRIBUTE**>;

The syntactical alternatives of the words *pušų* (*pine trees*), *sakai* (*utter*) and *blizgėjo* (*was glistening*), which are given in Figure 17, are rejected because in this sentence the syntactical alternatives do not form THREADS. The verb *blizgėjo* (*was glistening*) has no subject for the third person singular, which would be expressed by a noun in the nominative, singular, the predicate *sakai* (*utter*-2 person, singular) contains its unrealized valence: the verb *sakyti (to utter*) requires the accusative case which is absent in the sentence; the word *pušų* (*pine trees*) cannot act as an object, because the predicate *blizgėjo* (was

glistening) does not require the genitive case. This means that the verb acting as a predicate in this sentence does not have any semantic features, which point out that this verb must have a complement in genitive. The analysis of the second sentence with word *sakai* will look as shown in Figure 18.

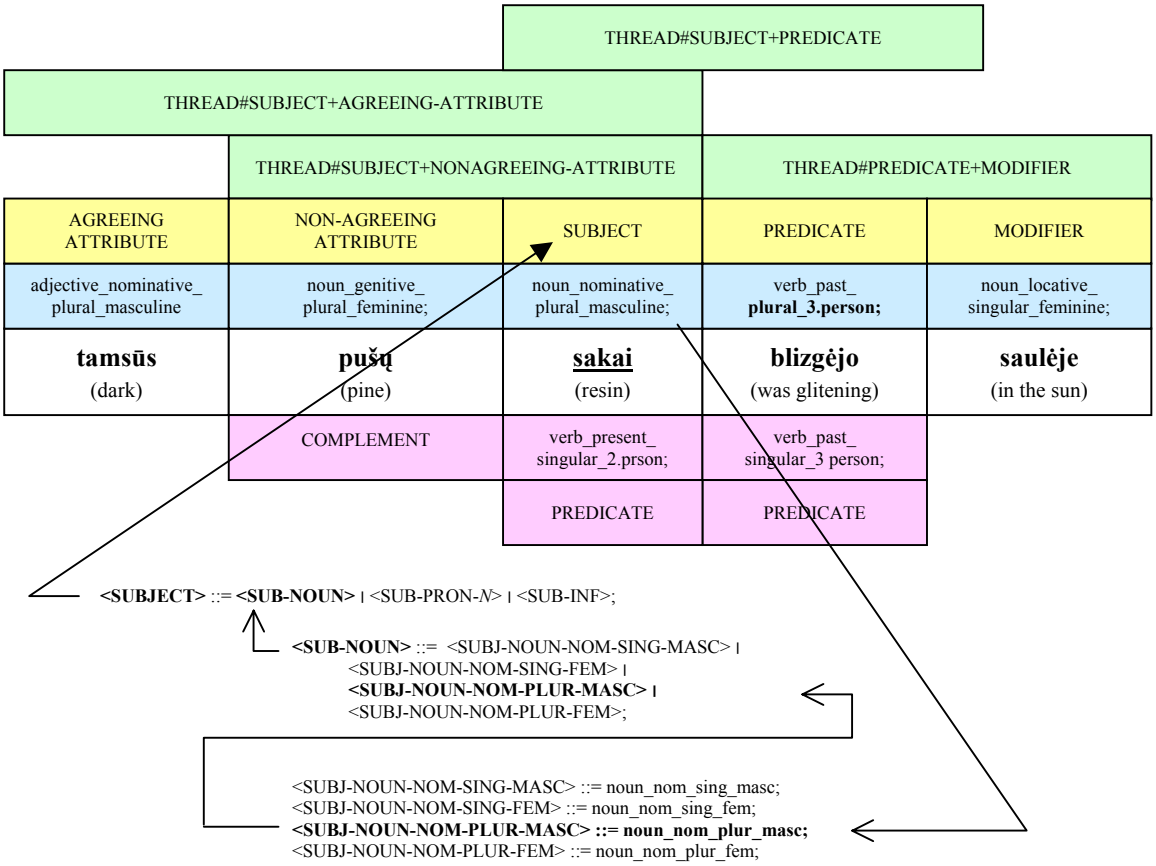

Figure 17. The way of finding the subject in the sentence *Tamsūs pušų sakai blizgėjo saulėje*

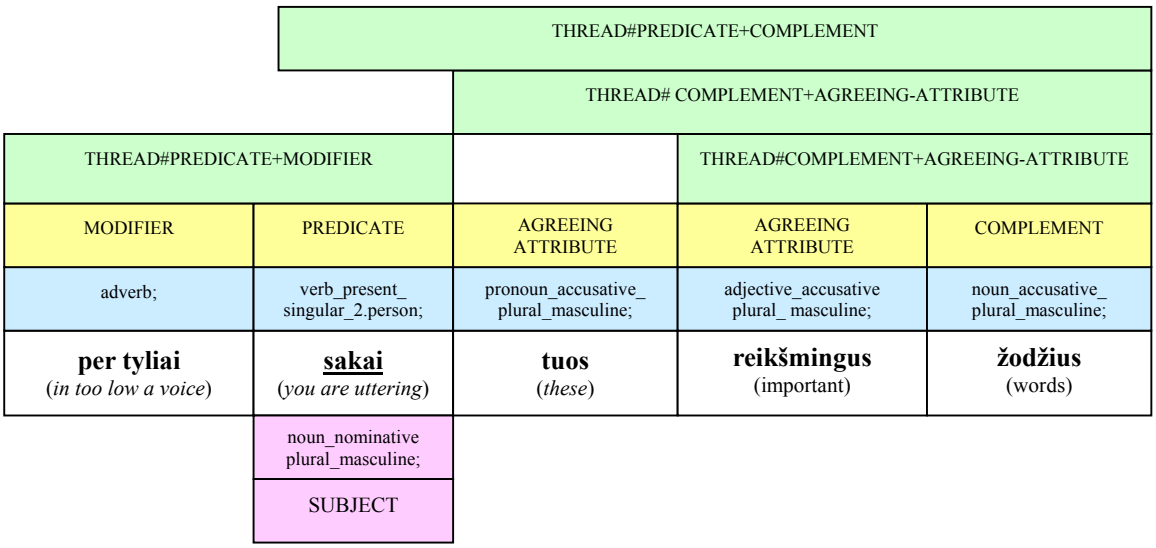

Figure 18. The syntactic analysis of the sentence *Per tyliai sakai tuos reikšmingus žodžius*

9. The possibilities of the application of the results of the syntactic analysis

The examples analysed in the previous section show how the morphological ambiguity of words can be destroyed with the help of the syntactic analysis. Consequently, with the help of the syntactical analysis, the results of the morphological analysis of the Lithuanian language can be improved. The process mentioned above is not the principal aim of the syntax formalization, though. The aim of the syntactic analysis is to prepare a Lithuanian sentence for the machine

translation, that is, to prepare such a structure of a sentence, which could be changed for the corresponding structure in a different language. One cannot translate verbally because the results of similar attempts would be grammatically incorrect sentences in different languages. For example, the Lithuanian sentence *Einu namo*, if translated in verbatim into the German language **Gehe nach Hause* would be grammatically incorrect, and the spellers in the German language would indicate the syntactical mistakes immediately. Sometimes the results of verbal translations can be wrong. The verbatim translation of *Einu namo* into the English language *Go home* is a sentence in the imperative mood, which would sound *Eik namo* in the Lithuanian language. That is why during the stage of the transfer all the Lithuanian sentences where the personal pronouns of the first or the second person are omitted (*aš – I; mes – we; tu, jūs – you*), the subject should be restored in the adequate form. In the Lithuanian language the personal pronouns tend to be omitted for the purposes of style, in an attempt to avoid the superfluity of information. We can guess those pronouns from the flexions of the verbs. For example, the structure of the sentence *Šiandien grįšiu į namus vėlai* (*I am going to return home late tonight*) should be changed in the manner shown in Figure 19, when translating this sentence into the German language.

There are many similar cases to be encountered in the Lithuanian language. The copula of the Present (*yra – is, are*) usually gets omitted in the Lithuanian sentence. This copula should be restored when translating texts into the English or German languages, because the Germanic languages do not tolerate sentences without verbs. In Lithuanian, for example, the sentence *Jis geras mokytojas* is quite correct. In English or German the verbatim translations are not correct: **He a good teacher, *Er ein guter Lehrer*.

Transfer is not the task of the syntactic analysis, though. While creating the systems of machine translation, the structures of a sentence received during the syntactical analysis are used as initial data during the stage of the transfer. All the work in transfer phase is performed in accordance with the specially prepared programs.

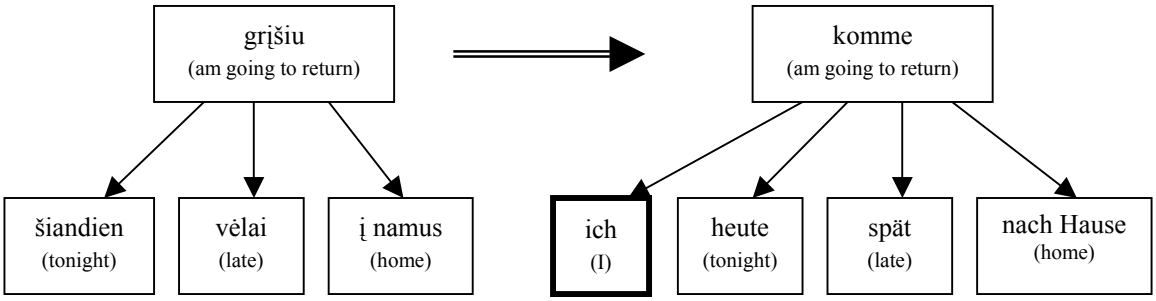

Figure 19. Restoring of the missing subject by translating the sentence *Šiandien grįšiu į namus vėlai* (*I am going to return home late tonight*) into the German

10. Conclusions

The new method of the syntactic analysis is presented in this article. Specific features of the Lithuanian language i.e. a great inflexion of the language and the free word order in a sentence are evaluated in this work.

The results will prove useful when creating the systems of the machine translation of the Lithuanian language.

References

[1] **J. Allen.** Natural Language Understanding. *Amsterdam: The Benjamin/Cummings Publishing Company*, 1987.

[2] **I.S. Batori, W. LendersW.** Computerlinguistik: Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendung. *Berlin: Walter de Gruyter,* 1989.

[3] **P. Hellwig.** *http://www.cl.uni-heidelberg.de /~hellwig/dug-2002.pdf.*

[4] **W.J. Hutchins.** Machine Translation: Past, Present, Future. *Chichester: Ellis Harwood Limited*, 1986.

[5] **W.J. Hutchins, H.L. Sommers.** An Introduction to Machine Translation. *London: Academic Press*, 1982.

[6] **M. Kay, J.M. Gawron, P. Norvig.** Verbmobil: A Translation System for Face-to-Face Dialog. *Stanford: CSLI*, 1994.

[7] **J.D.K. Kelly**. Progress in Machine Translation. *Wilmslow, (UK): Sigma Press*, 1989.

[8] **V.Labutis.** Lietuvių kalbos sintaksė. *Vilnius: Vilniaus universiteto leidykla,* 2002.

[9] **S. Langer.** Selektionsklassen und Hyponymie im Lexikon: semantische Klassifizierung von Nomina. *Dissertation. München: Universität München*, 1996.

[10] **M. Schwanke.** Maschinelle Übesetzung: Ein Überblick über Theorie und Praxis. *Berlin: Springer-Verlag*, 1991.

[11] **J. Slocum.** A Survey of Machine Translation: Its History, Current Status, and Future Prospects. *Computational linguistics, Vol*.11, *No*.1, *January-March*, 1985, 1-17.

[12] **D. Šveikauskienė.** Graph Representation of the Syntactic Structure of the Lithuanian Sentence. *Informatica, Vol*.16, *No*.3, 2005, 407-418.

[13] **D. Šveikauskienė.** Automatinio vertimo apžvalga. *Informacinės technologijos*'98. *Konferencijos pranešimų medžiaga. Kaunas: Technologija*, 1998, 191-194.

[14] **T. Winograd.** Language as a Cognitive Process. *Vol*. I: *Syntax. London: Addison-Wesley Publishing Company*, 1983.

[15] **V. Zinkevičius.** Lemuoklis - morfologinei analizei. *Darbai ir dienos* 24, *Kaunas: VDU*, 2000, 245-273.

[16] **Т.М. Гарина.** Программирование синтаксических преобразований. *Международный семинар по машинному переводу. Москва: ВЦП*, 1979, 43-44.

[17] **Ю. Кунце**. Введене семантических критериев в синтаксические правила. *Научно-технические исследования* (6), 1981, 30-34.

[18] **Ю.Н. Марчук.** Опыт машинного семантико-синтаксического анализа текста для перевода. *Семантика текста и проблемы перевода. Москва: Институт языкознания*, 1984.

DOI: 10.5755/j01.itc.34.3.12018