VYTAUTAS MAGNUS UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Daiva ŠVEIKAUSKIENĖ

AUTOMATIC SYNTACTIC ANALYSIS OF LITHUANIAN SIMPLE SENTENCES

Summary of doctoral dissertation

Physical Sciences (P 000)
Informatics (09 P)
Artificial Intelligence (P 176)

Vilnius, 2009

This research was accomplished in the period from 2000 to 2009 at the Institute of Mathematics and Informatics.
The right for the doctoral studies in informatics was granted to Vytautas Magnus University together with Institute of Mathematics and Informatics by Government of the Republic of Lithuania, Decree No. 1285, issued on the 13th of December 2004.
Dissertation is defended entrance.

Scientific Consultant

Prof Dr Habil Laimutis TELKSNYS (Institute of Mathematics and Informatics, Physical Sciences, Informatics, 09 P).

The dissertation is being defended at the Council of Scientific Field of Informatics at Vytautas Magnus University:

Chairman

Prof Dr Habil Vytautas KAMINSKAS (Vytautas Magnus University, Physical Sciences, Informatics, 09 P).

Members:

Prof Dr Dalė DZEMYDIENĖ (Mykolas Romeris University, Physical Sciences, Informatics, 09 P),

Prof Dr Habil Genadijus KULVIETIS (Vilnius Gediminas Technical University, Physical Sciences, Informatics, 09 P),

Doc Dr Gailius RAŠKINIS (Vytautas Magnus University, Physical Sciences, Informatics, 09 P),

Prof Dr Habil Rimantas ŠEINAUSKAS (Kaunas University of Technology, Technological Sciences, Informatics Engineering, 07 T).

Opponents:

Doc Dr Regina KULVIETIENĖ (Vilnius Gediminas Technical University, Physical Sciences, Informatics, 09 P),

Doc Dr Andrius UTKA (Vytautas Magnus University, Humanitarian Sciences, Philology, 04 H).

The dissertation will be defended at the public meeting of the Scientific Council in the field of Informatics in the auditorium 203 of the Institute of Mathematics and Informatics at 1 p.m. on February 23, 2010.

Address: Akademijos street 4, LT-08663, Vilnius, Lithuania.

The summary of the dissertation was sent-out in January, 22, 2010.

The dissertation is available at M. Mažvydas National Library of Lithuania, the Library of the Institute of Mathematics and Informatics, the Library of Vytautas Magnus University.

VYTAUTO DIDŽIOJO UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Daiva ŠVEIKAUSKIENĖ

LIETUVIŲ KALBOS VIENTISINIŲ SAKINIŲ AUTOMATINĖ SINTAKSINĖ ANALIZĖ

Daktaro disertacijos santrauka

Fiziniai mokslai (P 000)
Informatika (09 P)
Dirbtinis intelektas (P 176)

Vilnius, 2009

Disertacija rengta 2000-2009 metais Matematikos ir informatikos institute.
Doktorantūros teisė suteikta kartu su Vytauto Didžiojo universitetu 2004 m. gruodžio mėn. 13 d. Lietuvos Respublikos Vyriausybės nutarimu Nr. 1285.
Disertacija ginama eksternu.

Mokslinis konsultantas:

prof. habil. dr. Laimutis TELKSNYS (Matematikos ir informatikos institutas, fiziniai mokslai, informatika, 09 P).

Disertacija ginama Vytauto Didžiojo universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

prof. habil. dr. Vytautas KAMINSKAS (Vytauto Didžiojo universitetas, fiziniai mokslai, informatika, 09 P).

Nariai:

prof. dr. Dalė DZEMYDIENĖ (Mykolo Romerio universitetas, fiziniai mokslai, informatika, 09 P),

prof. habil. dr. Genadijus KULVIETIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, informatika, 09 P),

doc. dr. Gailius RAŠKINIS (Vytauto Didžiojo universitetas, fiziniai mokslai, informatika, 09 P),

prof. habil. dr. Rimantas ŠEINAUSKAS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija, 07 T).

Oponentai:

doc. dr. Regina KULVIETIENĖ (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, informatika, 09 P),

doc. dr. Andrius UTKA (Vytauto Didžiojo universitetas, socialiniai mokslai, filologija, 04 H).

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2010 m. vasario mėn. 23 d. 13 val. Matematikos ir informatikos institute, 203 auditorijoje.

Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2010 m. sausio mėn. 22 d.

Su disertacija galima susipažinti Martyno Mažvydo nacionalinėje bibliotekoje, Matematikos ir informatikos instituto bei Vytauto Didžiojo universiteto bibliotekose.

The Object of the Investigation:

The present research analyses the automation of syntactical analysis of a Lithuanian sentence. The present work analyses the possibilities of the syntactical structure of a Lithuanian simple sentence to be arranged by computer.

The Actuality of the Theme:

Many systems of automatic syntactic analysis have already been created to satisfy the needs of a great number of world languages. The Lithuanian language has not got a similar system yet, and the major reason of this backwardness can be explained by the fact that the Lithuanian language has not been sufficiently formalized and prepared to processing by computer.

The Aim of the Work:

The aim of the work is to create the system for the performing of the automatic syntactic analysis of the Lithuanian sentences, i.e., to prepare the method and the software for the purpose of the arranging of the syntactic structure of a Lithuanian simple sentence by computer.

Tasks:

1. To analyze the already existent systems of the automatic syntactic analyses and to research into the possibilities of their application for the automatic syntactic analysis of the Lithuanian language.
2. To find out the specific futures of the Lithuanian language and to create the new method for the automatic syntactic analysis of the Lithuanian simple sentences.
3. To create the formal grammar for the purpose of the description of the syntax of the Lithuanian language and to present it in the form of BNF (Backus and Naur form).
4. To create the algorithm and to prepare the software, determining the syntactic structure of a Lithuanian simple sentence.
5. To test the work of the created system while using the set of simple sentences of the Lithuanian language.

The Methods of the Analysis

In the course of the theoretical investigation, the methods and achievements in computer linguistics, automatic syntactic analysis, in the Lithuanian linguistics, the theory of programming languages as well as in the knowledge and methods of programming have been made avail of.

In the course of the experimental research, the system of the morphological analysis of the Lithuanian language, created by Vytautas Zinkevičius, as well as the Visual Basic'6, and the corpus of the Lithuanian language have been used.

The Newness of Research

Bearing in mind the huge inflexion of the Lithuanian language, the syntactic functions of words are differentiated in accordance with their morphological categories. Such an approach has never been known in any of the already created systems of automatic syntactic analysis. Even the syntactical analysis of the Russian language does not link parts of a sentence with their morphological attributes though the Russian language is very close to the Lithuanian language in the aspect of their inflexion. All the already created systems of the syntactic analyses tend to improve their results with the

help of the semantic information. Never was used the morphological data regarding the word for similar purposes.

Considering another typically Lithuanian feature of the language, namely, the free word order in a sentence, the formal parameter GIJA (THREAD) has been introduced. THREAD indicates the links among the words and their positioning. The description of THREAD in BNF consists of three parts: in the first and third position words, or rather their syntactical functions are indicated. The syntactical links between the words are being sought for. In the middle a non-terminal symbol INTARPAS (INSERTION) is put in. INSERTION reflects the information about what could have intervened between the words of THREAD. A similar principle is not known to any of the already created systems of automatic syntactic analysis.

The syntactical structure, embracing simple Lithuanian sentences, has been formed in this work. The structure defines all the possible cases of the simple Lithuanian sentences. The analogical structures current in other languages are not presented.

The dependency tree has been modified: new arcs are added to present the syntactic relation of the predicative attribute. The present research proves that part of information would be lost if a traditional dependency tree were used to demonstrate the syntax of the Lithuanian language.

Publications:

The main statements and results of the research presented in this dissertation have been published in three publications, of which the first was published in the ISI indexed journal "Informatica".

1. Šveikauskienė, D. *Graph Representation of the Syntactic Structure of the Lithuanian Sentence.* Informatica. 2005,Vol. 16, Nr. 3, p. 407-418.
2. Šveikauskienė, D. *A System for Automatic Syntactic Analysis of Lithuanian Simple Sentences*. Information Technologies and Control. 2007 Vol. 36, Nr. 2, p.221-237.
3. Šveikauskienė, D. *Formal Description of the Syntax of the Lithuanian Language*. Information Technologies and Control. 2005 Vol. 34, Nr. 3, p.245-256.

For the purposes of the defense the following materials are presented:

- The formal description of the syntactic rules of the Lithuanian language.
- The method of the determination of the syntactical functions of the words in a simple Lithuanian sentence. The specific features of the Lithuanian language, namely, a great inflexion of words and their free order in a sentence are taken into account.
- The software enabling to perform the syntactical analysis of the simple Lithuanian sentences by computer.

The Results Achieved

The precision of the software was tested when using 8 samples from different parts of Lithuanian corpus. The following results have been achieved: the analysis of 92% sentences was correct. The correctness their syntactical structure was approbate by the Lithuanian linguist, Doctor E. Valiulytė. The syntactical structures of 8% sentences were structured wrongly. The mistakes of the analysis can be divided into three types:

1. Mistakes, which occurred due to the lack of semantic information. For example, an adverb was wrongly taken to be an adverbial modifier of an adjective. Such mistakes can be avoided when we create an automatic semantic analysis of the Lithuanian

language, i.e., when we automatically receive the information that the adverbial modifier of place cannot accompany an adjective possessing the features of time.

2. The mistakes, which occur because of the incorrectness of the initial data, when, during the stage of the morphological analysis, the lemma of a very rarely used word is given in the first place. For example, the noun *dienas* (diene) is given as the first alternative of the lemma of the word *dieną* (in the day).
3. The mistake caused by the inefficiency of the syntactical analysis. Such mistakes are likely to occur when we use one of the homographs, i.e., the words written the same way but differing in their morphological forms. This can happen when, for example, the forms of the Genitive case of the feminine nouns and adjectives in the singular and the Nominative case of the feminine nouns and adjectives in the plural coincide. We read the sentence: *Ateities istorikų laukia nelengvos mūsų Lietuvos studijos* (The uneasy studies of our Lithuania await the historians of the future). The word *nelengvos* (uneasy) was attached to the word *Lietuva* (Lithuania). If, for the purposes of the analysis, the sentence were presented orally via the microphone or telephone and not as the text containing the letters the mistake would be eliminated, because the oral stress would help unequivocally to determine the form of the word. The other way helping the researcher to avoid similar mistakes remains the semantic analysis of the text. If we have the information that the studies can be easy or uneasy but not Lithuania itself, similar mistakes would also be eliminated.

The mistakes can be avoided by perfecting the morphological analysis and by creating the automatic semantic analysis of the Lithuanian language.

1. INTRODUCTION

If we choose to remember the already completed works dedicated to the task of the formalization of the Lithuanian language, the lemmatizing (automatic morphologic analysis) created by V. Zinkevičius should be the first to be mentioned [Zin00].

The automatic syntactic analysis of the Lithuanian language has not been prepared yet. That is why this work attempts to present the automatic syntactical analysis of a Lithuanian simple sentence.

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 automatic syntactic analysis systems already, are too big.

This work attempts to evaluate the specific features of the Lithuanian language – its great inflexion and the free word order in a sentence. The work also aspires to create the method enabling a good quality automatic syntactical analysis of Lithuanian simple 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. At least the author of this work is not familiar with any literary source, describing the morphological methodology of the syntactic analysis. 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.

2. 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, which is called dependency tree.

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 1).

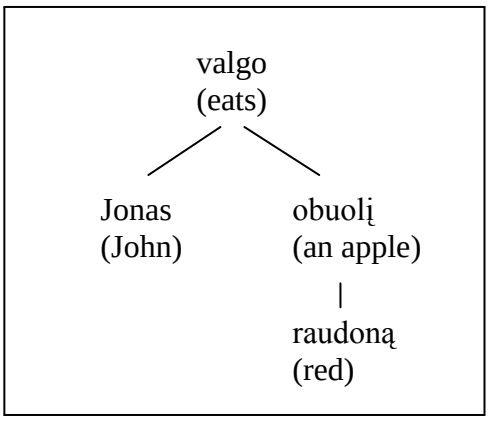


Figure 1 The dependency tree of the sentence *Jonas valgo raudoną obuolį (John eats a red apple)*

A generalized structure of the node in the dependency tree is shown in Figure 2. Every node of the dependency tree is occupied by a word, which can have one or more subordinated words and only one superordinated word [Hel02].

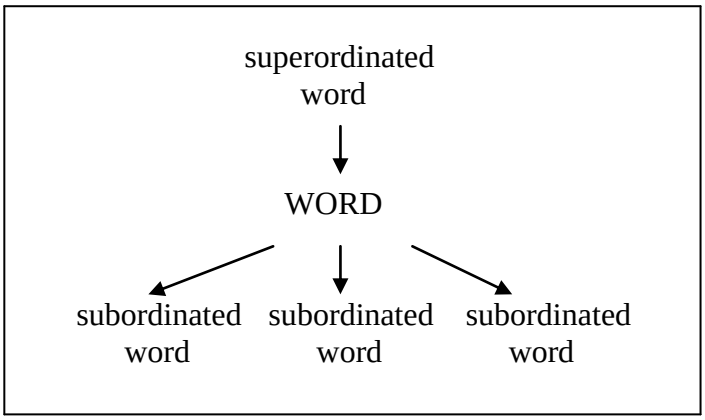


Figure 2 The links of the node of the dependency tree with adjacent nodes

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. SOME SPECIFIC FUTURES 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 [Lab02]. 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. 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.

4. THE BNF DESCRIPTION OF THE LITHUANIAN SYNTAX

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 language, which are able to perform these functions. Later every part of the language is divided into morphological categories specific to this part of language. 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 dainą — to sin the song*), but the accusative case indicating the time performs the function of the adverbial modifier of time (*dainuoti naktį — to sing at night*). 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;

etc.
```

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>;
```

5. 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 relationship. 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 relationship of a subject and the first non-agreeing attribute. For example, the collocation *mano namas* (*my house*) will admit only an agreeing attribute, such as *senas* (*old*), which will not affect the initial relationship: *mano senas namas* (*my old house*) will remain *mano namas* (*my house*), anyway (Figure 3). 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 4).

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 5).

Consequently, the BNF description 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 6).

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>;

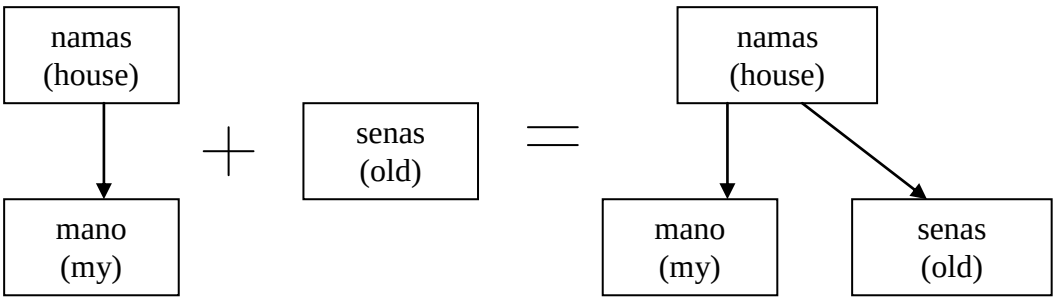


Figure 3 The interference of the agreeing attribute into the word collocation *mano namas* (*my house*)

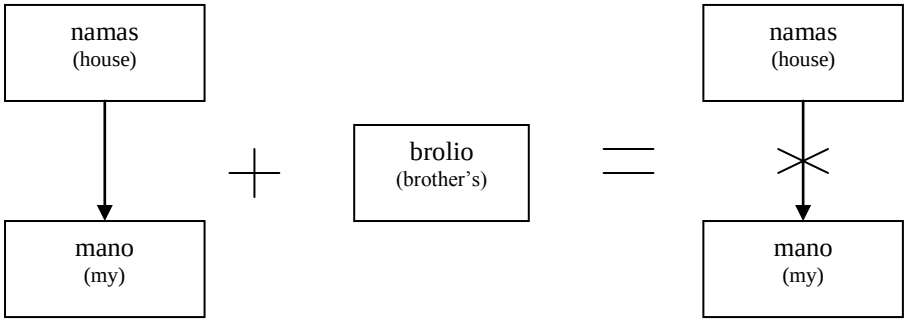


Figure 4 The interference of the non-agreeing attribute into the word collocation *mano namas* (*my house*)

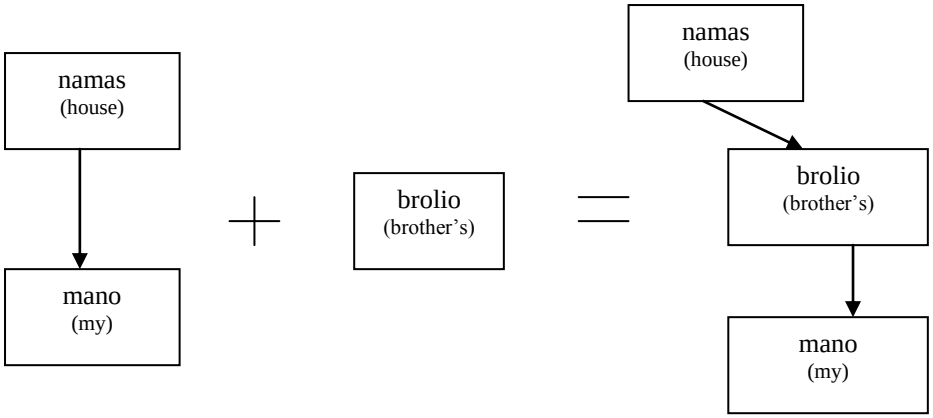


Figure 5 The formation of new word collocations

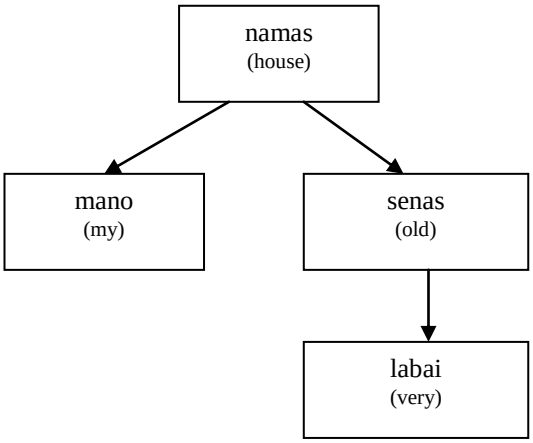


Figure 6 Insertion expressed by a THREAD of an agreeing attribute *labai senas* (*very old*)

6. EXEMPLES OF THE ANALYSIS OF THE SENTENCES

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: *Tamsūs pušų sakai blizgėjo saulėje* (*The dark resin of the pine trees was glistening in the sun*). The syntactic structure of this sentence should look as in Figure 7.

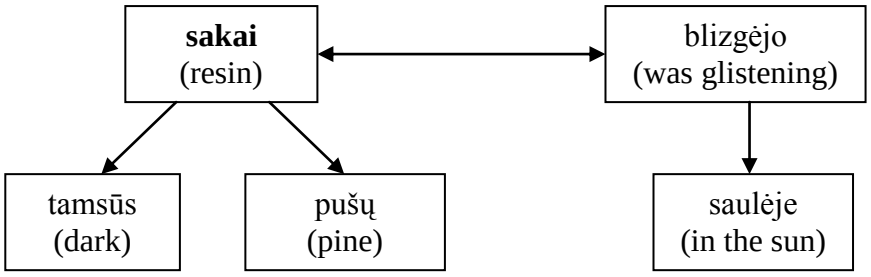


Figure 7 The syntactic structure of the sentence *Tamsūs pušų sakai blizgėjo saulėje* (*The dark resin of the pine trees was glistening in the sun*)

The arcs connecting the nods of the graph, that is, the syntactical relationships 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 8.

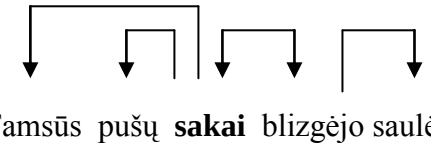


Figure 8 Syntactical relationships among words shown in the linear structure of the sentence *Tamsūs pušų sakai blizgėjo saulėje* (*The dark resin of the pine trees was glistening in the sun*)

The non-terminal symbol THREAD in BNF description corresponds to the arrows placed over the words of the sentence in Figure 8.

The syntactic 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 9). 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 9.

The syntactical alternatives of the words *pušų* (*pine trees*), *sakai* (*utter*) and *blizgėjo* (*was glistening*), which are given in Figure 9, 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.

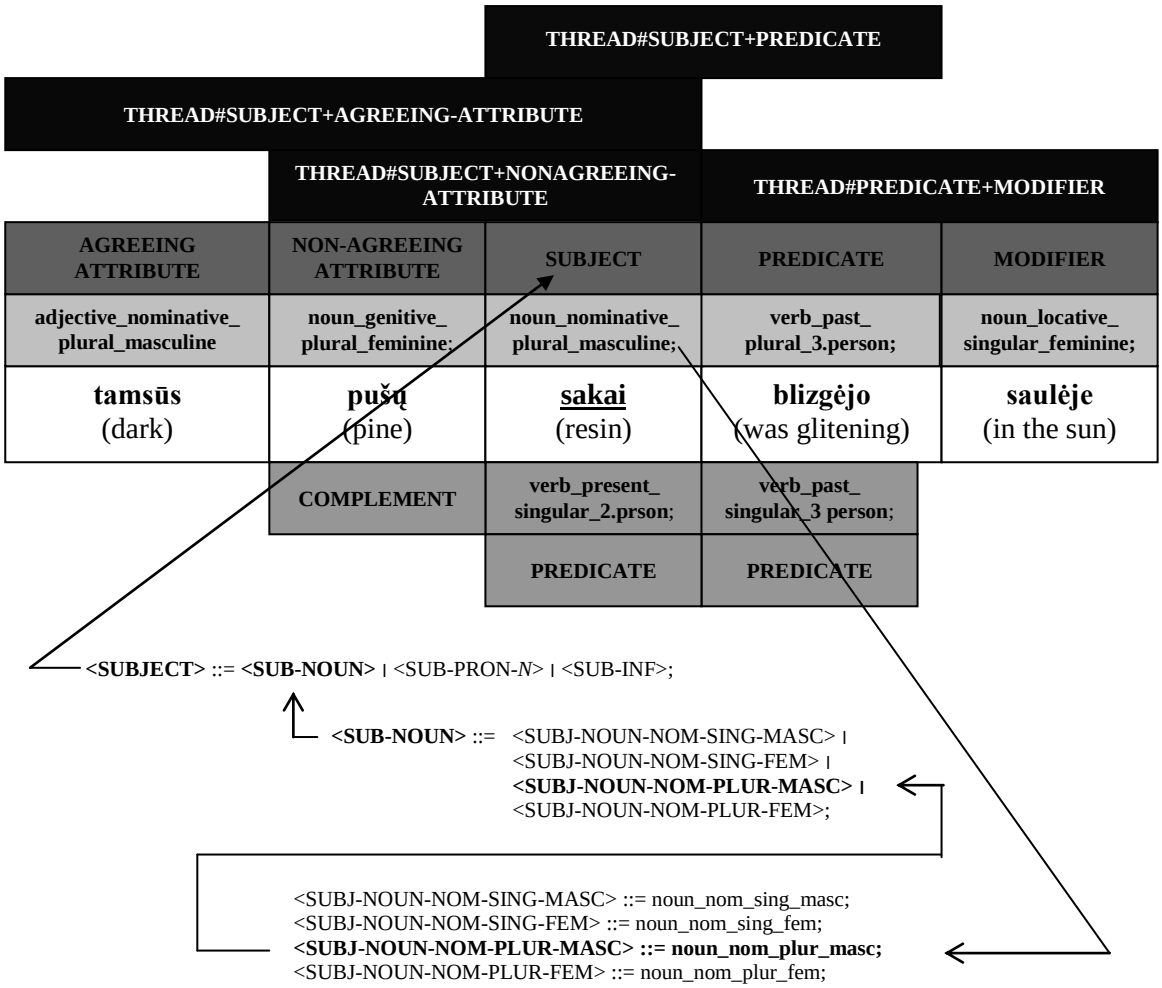


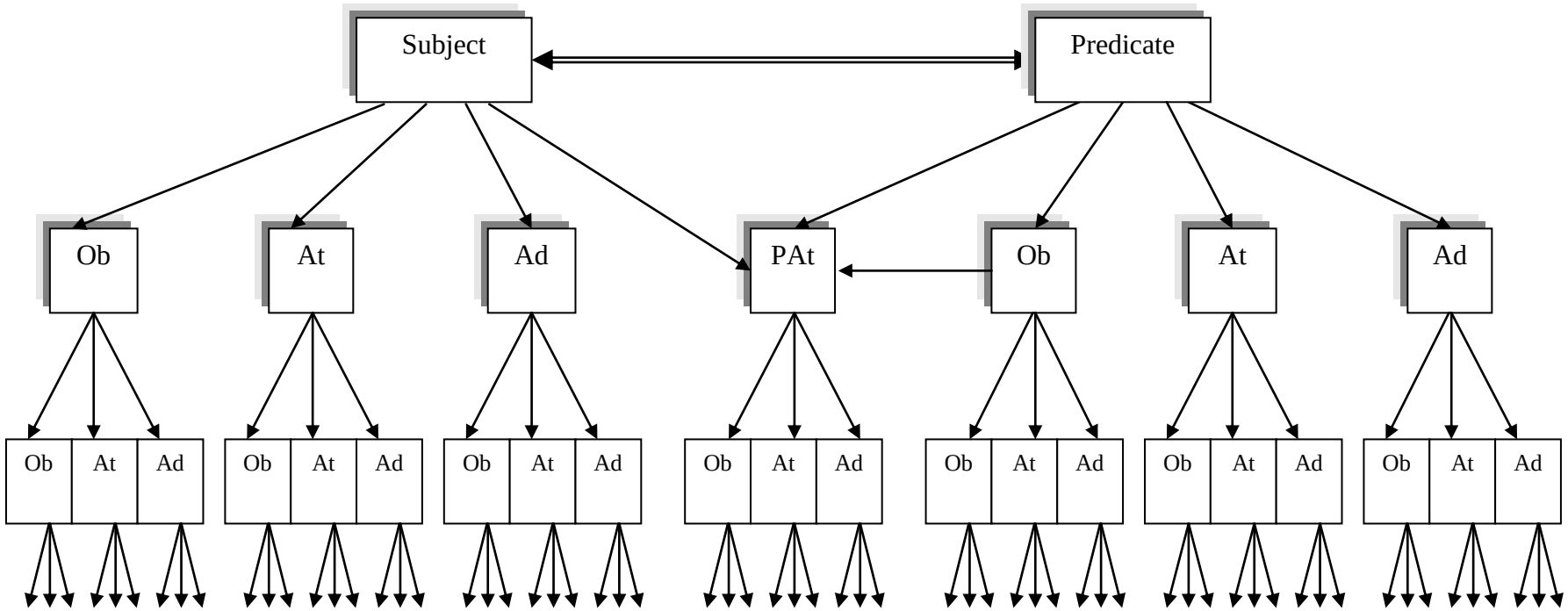
Figure 9 The way of finding the subject in the sentence *Tamsūs pušų sakai blizgėjo saulėje*

7. GRAPH REPRESENTATION OF THE LITHUANIAN SENTENCE

While performing the computerized syntactic analysis of the Lithuanian language, it would be preferable to have a generalized scheme, which would embrace any Lithuanian sentence. The scheme should be common for all the simple sentences of the Lithuanian language. Every particular sentence should activate one path in the scheme. The generalized scheme of the Lithuanian sentence is shown in Figure 10. All the five parts of a sentence – subject, predicate, object, attribute and adverbial modifier – can be extended by the additional usage of attribute, object and adverbial modifier. None of them can be extended through the additional usage of subject or predicate, though. The scheme reflects the mentioned statements. The shadows on the blocks, corresponding to the parts of the sentence, denote possible homogeneous parts of a sentence.

The syntactic structure of the Lithuanian sentence is presented in accordance with the rules indicated in the newest grammar of the Lithuanian language. The latest " Syntax of the Lithuanian Language" by V. Labutis (2002) states that the Lithuanian language contains the two principle parts, which are subject and predicate, and three secondary parts, which are object, attribute and adverbial modifier.

The principle parts of the sentence are placed on the same level at the top of the graph, and they are regarded to be the equal nods of the same range. The secondary parts of the sentence, which extend the principle ones, are placed lower.



Abbreviations: At – attribute, Ob – object, Ad – adverbial modifier, PAt – predicative attribute.

Figure 10 Generalized structure of a simple Lithuanian sentence.

Predicative attribute can be characterized by double syntactic relationships. Formally it is made to agree either with the subject or the object of a sentence and it is also made to adjunct to the verb. Therefore, the scheme presents three arcs leading to the predicative attribute. In a particular sentence only two arcs will be used. The arc between the predicative attribute and the predicate will characterize every sentence, possessing the predicative attribute. The other arc, be it the one leading from the subject or from the object, will be determined by the words of a particular sentence.

It is necessary for the syntactic structure of the Lithuanian sentence to use a graph, because a tree can't reflect all syntactic information, which a Lithuanian sentence contains. The statements mentioned above can be illustrated by the following example: in the German language, which is less inflective than the Lithuanian language (that is, the German language has fewer forms with differing endings than the Lithuanian language has) formally the predicative attribute is identical with the adverbial modifier. The endings of German words do not have to agree with the part of the sentence the predicative attribute indicates. For example, in the following sentences we read:

Der Vater kam gestern verärgert. (*The father returned yesterday angry*)
Die Mutter kam gestern verärgert. (*The mother returned yesterday angry*)
Die Brüder kamen gestern verärgert. (*The brothers returned yesterday angry*)
Die Schwestern kamen gestern verärgert. (*The sisters returned yesterday angry*)

In the examples mentioned above, the form of the predicative attribute remains the same (*verärgert*), irrespective of the gender or number of the subject, which should mean, that the predicative attribute does not change its form depending on the noun. The Lithuanian language is different. In the Lithuanian translation of the sentences mentioned above, the word *verärgert (angry)* will have four correspondences whose forms will correlate with the subject:

Tėvas vakar grįžo piktas.
Motina vakar grįžo pikta.
Broliai vakar grįžo pikti.
Seserys vakar grįžo piktos.

Consequently, when translating these sentences from German into Lithuanian, the syntactic structure of a German sentence, shown in Figure 11, should be changed for the syntactic structure of a Lithuanian sentence, shown in Figure 12.

The lack of information is particularly clear when we consider the structure of the tree of those sentences, which have both the predicative attribute and the object, because the predicative attribute can depend both on the subject and on the object. Consequently, the problem which word the ending of the predicative attribute should correlate with is not clear at all. For example, if we demonstrate the sentence *Die Mutter aß die Mohrrüben roh* (*The mother ate the carrots raw*) the way it is shown in Figure 13, it remains not clear what or who was raw – carrots or the mother. The word *žalias*, i.e., raw, depends on the correlation in a Lithuanian sentence. The possibility is twofold:

1. **Motina morkas valgė žalia* (*The mother was raw, when she ate the carrots*)

like in the sentence

Motina vakar grįžo pikta (*The mother returned yesterday angry – the mother was angry, when she returned yesterday*);

2. *Motina morkas valgė žalias* (*The carrots were raw, when the mother ate them*).

If one wishes correctly and without mistakes to generate the sentence, translated into the Lithuanian language, in the process of the machine translation one has to use the structure of the Lithuanian sentence, indicated in Figure 14.

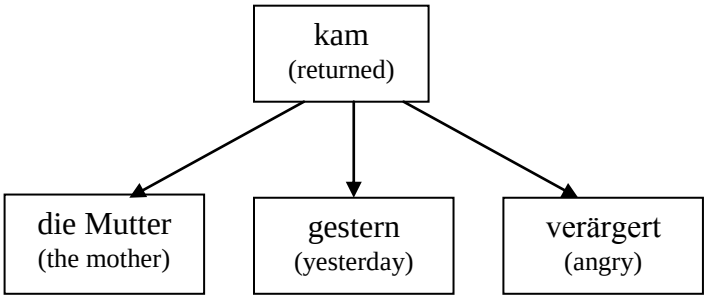


Figure 11 The syntactic structure of the German sentence *Die Mutter kam gestern verärgert*. (*The mother returned yesterday angry*)

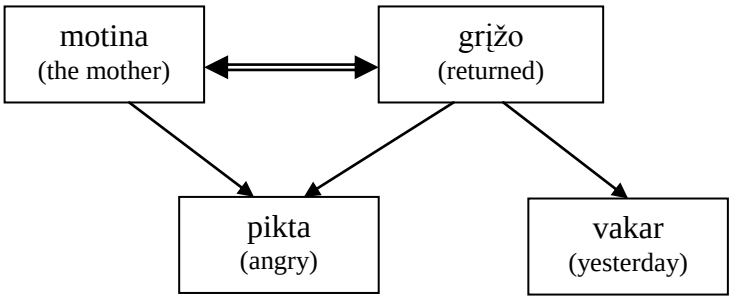


Figure 12 The syntactic structure of the Lithuanian sentence *Motina vakar grįžo pikta*. (*The mother returned yesterday angry*)

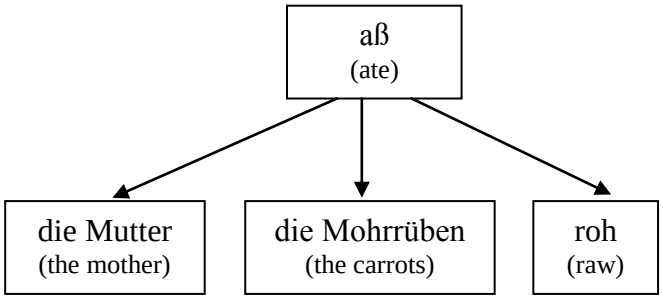


Figure 13 The syntactic structure of the German sentence *Die Mutter aß die Mohrrüben roh*. (*The mother ate the carrots raw*)

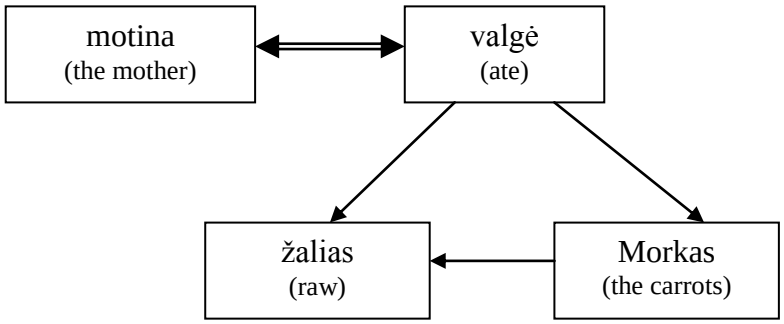


Figure 14 The syntactic structure of the Lithuanian sentence *Motina morkas valgė žalias.* (*The mother ate the carrots raw*)

Those two examples testify that the two German sentence structures presenting the same graphic picture, as they are shown in Figure 11 and Figure 13, (the graph consists of four nodes and three arcs leading from one node to the remaining three ones that have no

interconnection) should be changed by two differing structures of two Lithuanian sentences in Figure 12 and Figure 14. Consequently, a German sentence structure does not present enough information to us to enable us correctly to generate a sentence, translated into the Lithuanian language.

8. EXPERIMENTAL RESEARCH

The program test was performed on the basis of eight samples of sentences, selected from various sections of the Lithuanian language corpus by the expert. Each sample consisted of 50 simple sentences, procured from a coherent text. It was additionally determined which part of the text consisted of simple sentences. It has been concluded that the coherent text of the Lithuanian language contained round 57% of simple sentences.

The results of the test were the following: 368 sentences out of 400 used in the test were analyzed correctly, which means that the precision of the software is 92%. The remaining mistakes could be grouped into three categories:

1. The part of a sentence is incorrectly determined. (Such mistakes were in 5 sentences);
2. The relationships between words were incorrectly ascertained. (Such mistakes figured in 16 sentences.)
3. The sentence structures were not formed. (Such mistakes were noticed in 11 sentences).

The sources of the mistakes could be grouped into three categories:

1. Mistakes occurring because of the lack of semantic information, that is, because of the absence of the automatic semantic analysis of the Lithuanian language.
2. Mistakes made because of the coincidence of the morphological forms of the Lithuanian words.
3. Mistakes occurring because of the morphological data being presented in not optimal way.

For the purpose of the reduction of mistakes, it is necessary to create:

1. The data basis of the morphologic data of the Lithuanian language.
2. The data basis of the semantic data of the Lithuanian language.
3. The word collocations frequency dictionary of Lithuanian language.

Figure 15 is the example of a correctly analyzed sentence.

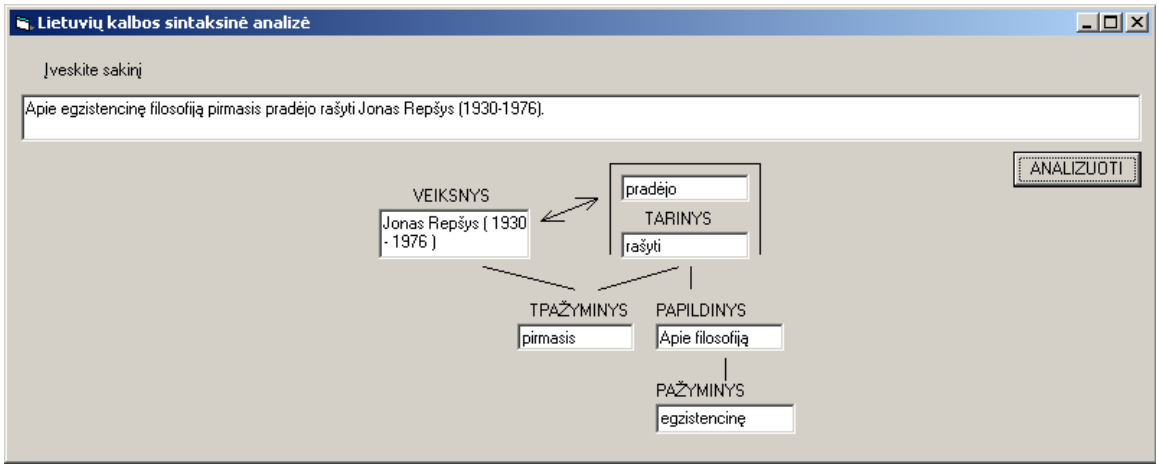


Figure 15 The syntactic structure of the Lithuanian sentence with predicative attribute.

9. APPLICATION OF THE SYNTACTIC ANALYSIS

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* (I go home) 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 16, when translating this sentence into the German language.

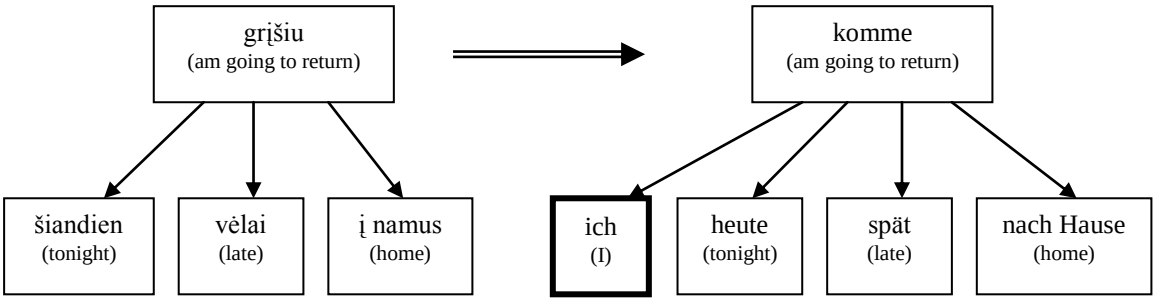


Figure 16 Restoring of the missing subject by translating the sentence *Šiandien grįšiu į namus vėlai* (*I am going to return home late tonight*) into the German

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*.

10. CONCLUSIONS

1. The analytical overview of the automatic syntactical analyses of the English, German and Russian languages has been made and it emerged that the systems created for other languages cannot be used for the purposes of the syntactical analysis of the Lithuanian language.
2. It has been shown that the tree-like syntactic structures, applicable to foreign languages, lack the information enabling us to generate a faultless Lithuanian sentence and it has been proved that the usage of a graph is indispensable for the syntactic structure of a Lithuanian sentence.
3. For the first time the formal grammar (BNF – Backus and Naur form) used for the purposes of the description of the syntax of the Lithuanian language has been created.
4. To perform the syntactical analysis of the Lithuanian language, the new methodic has been created which takes into account the very specific features of the Lithuanian language - great inflexion of the language and the free word order in a sentence.
5. To perform the syntactical analysis of the Lithuanian language algorithm has been created and the software has been written.
6. The work of the system has been tested with 8 samples from different parts of Lithuanian corpus. The precision of the work is 92%.

BIBLIOGRAPHY

[Hel02] **Hellwig, P.** http://www.cl.uni-heidelberg.de/~hellwig/dug-2002.pdf .

[Lab02] **Labutis, V.** 2002. *Lietuvių kalbos sintaksė*. Vilnius: Vilniaus universiteto leidykla

[Win83] **Winograd, T.** 1983. *Language as a Cognitive Process. Volume I: Syntax.* London: Addison- Wesley Publishing Company

[Zin00] **Zinkevičius, V.** 2000. *Lemuoklis - morfologinei analizei - Darbai ir dienos 24,* Kaunas:VDU, p. 245-273

CURRICULUM VITAE

Daiva Šveikauskienė, date of birth:1960-05-31, place of birth: Panevėžys, Lithuania.
Education:

1. Kaunas Technological University (1978-1984), graduated in computer engineering with honor,
2. Vilnius State University (1988-1993), graduated in linguistic,
3. Postgraduate student at the Institute of Mathematics and Informatics Vilnius (2000-2004).

Publications:

1. Šveikauskienė, D. *Graph Representation of the Syntactic Structure of the Lithuanian Sentence.* Informatica. 2005,Vol. 16, Nr. 3, p. 407-418.
2. Šveikauskienė, D. *Formal Description of the Syntax of the Lithuanian Language*. Information Technologies and Control. 2005 Vol. 34, Nr. 3, p.245-256.
3. Šveikauskienė, D. *The Review of Automatic Translation* – Information Technologies ’98. The material of Conference Presentations. Kaunas: Technologija, 1998, p. 191-194.
4. Book: Z. Kudirka, A. Mačernis, K. Paulauskas, J. Riaubūnas, B. Stulpinas, D. Šveikauskienė, R. Valatkaitė. *Iformatics. Lithuanian-English-Russian-German Dictionary.* Vilnius, Institute of mathematics and informatics, 1999.
5. CD-ROM “YEAR OF THE LITHUANIAN BOOK”. Translation from Lithuanian into German by Daiva Šveikauskienė in working group at UNESCO chair “Informatics for the Humanities” at the Institute of Mathematics and Informatics, Vilnius.

Reziume

Darbas priklauso dirbtinio intelekto sričiai, jame nagrinėjamas vienas iš žmogaus protinio darbo automatizavimo uždavinių – lietuvių kalbos automatinės sintaksinės analizės sukūrimas.

Dėl didelių skirtumų tarp lietuvių kalbos ir kitų indoeuropiečių kalbų, kurios turi automatinę sintaksinę analizę, neglima tiesiogiai pasinaudoti jau sukurtomis kitose šalyse programomis ir būtina sudaryti naują savitą metodą, kuris gerai atspindėtų specifinius lietuvių kalbos bruožus – didelį kaitomumą ir laisvą žodžių tvarką sakinyje.

Darbe apžvelgtos trijų kalbų – anglų, vokiečių ir rusų – sintaksinės analizės metodikos. Visos šios kalbos priklauso tai pačiai kalbų grupei (indoeuropiečių), kaip ir lietuvių kalba, ir skiriasi viena nuo kitos kaitomumo laipsniu bei žodžių tvarkos sakinyje laisvumu.

Pagrindinis kriterijus, į kurį atsižvelgiama atliekant anglų ir vokiečių kalbų sintaksinę analizę, yra žodžių tvarka, nes beveik tik nuo jos priklauso šiose kalbose, kokią sintaksinę funkciją atlieka žodis. Lietuvių kalbai neturint griežtos, sugramatintos žodžių tvarkos didžiausias sintaksinės informacijos kiekis sukauptas žodžių formose (jų galūnėse). Anglų bei vokiečių kalboms sukurtose sintaksinės analizės sistemose nenumatytas sintaksinės informacijos paėmimas iš žodžių galūnių. Taigi, reikėjo sukurti iš principo naują, visiškai nesiremiančią žodžio vieta sakinyje sintaksinės analizės sistemą.

Rusų kalba artimesnė lietuvių kalbai kaitomumo požiūriu, tačiau rusų kalbos sintaksinę analizę atlieka grupės algoritmų ir čia nesinaudojama formaliu sintaksės aprašu, kaip yra anglų ir vokiečių kalboms sukurtose sistemose. Visos programavimo kalbos aprašomos formalios nekontekstinės gramatikos taisyklėmis ir, jei sprendžiamą uždavinį pavyksta aprašyti II tipo formalia gramatika (pagal Chomskio klasifikaciją), labai supaprastėja programavimas. Todėl šiame darbe buvo siekiama lietuvių kalbos sintaksę aprašyti nekontekstinės gramatikos taisyklėmis.

Visoms sakinio dalims sudarytas aprašas BNF (Bekaus ir Nauro forma), nurodantis kokios žodžio morfologinės formos gali atlikti kiekvieną sintaksinę funkciją.

Laisvai žodžių tvarkai lietuvių kalboje įvertinti naudojami du formalūs parametrai – Gija ir Intarpas. Gija aprašo žodžių junginius, t.y. tiesioginiu sintaksiniu ryšiu susietus žodžius, o Intarpas parodo, kokie kiti žodžiai, nepriklausantys šiam junginiui, gali būti tarp jų įsiterpę. Žodžių junginiai taip pat buvo aprašyti BNF.

Pagal sudarytą formalų lietuvių kalbos sintaksės taisyklių aprašą paruoštos programinės įrangos pagalba galima gana gerai išnagrinėti lietuvių kalbos sakinius. Programos veikimas patikrintas su 400 eksperto atrinktų rišlaus teksto vientisinių sakinių, kurie sudarė 8 imtis (po 50 sakinių) iš skirtingų Dabartinės lietuvių kalbos tekstyno sričių. Buvo gautas tikslumas 92%. Gautų sintaksinių struktūrų teisingumą aprobavo lituanistė, filologijos mokslų daktarė E. Valiulytė.

Apibendrinant galima būtų pateikti tokią išvadą: sukurtas naujas metodas, įgalinantis gerai atlikti lietuvių kalbos vientisinių sakinių sintaksinę analizę kompiuteriu.

Daiva ŠVEIKAUSKIENĖ

AUTOMATIC SYNTACTIC ANALYSIS OF LITHUANIAN SIMPLE SENTENCES

Summary of doctoral dissertation

Printed: VGTU Publishing House "Technika",
Saulėtekio str. 11, LT-10223, Vilnius.
Run of 60 copies. 2010.01.12