Graph Representation of the Syntactic Structure of the Lithuanian Sentence

Daiva ŠVEIKAUSKIENĖ

Institute of Mathematics and Informatics
Goštauto 12, LT-01108 Vilnius, Lithuania
e-mail: daivasv@ktl.mii.lt

Abstract. The paper offers a new way of presenting the structure of a sentence. None of the two widely known methods of representation the syntactic structure of a sentence can be of any avail when applied to the Lithuanian language. Neither the tree, based on the phrase structure principle, nor the tree, suggested by the dependency grammar, do reflect all the syntactic information, which a Lithuanian sentence contains.

The paper points out the differences between the Lithuanian language and other languages as well as presents the reasons why a Lithuanian sentence should be represented by a graph.

The paper presents a generalized structure of a simple sentence in the Lithuanian language, namely, such a structure, which would embrace all the possible instances of a Lithuanian simple sentence. Every sentence of the text would have to activate only one path in the generalized structure.

Key words: computer linguistics, natural language processing, machine translation, syntactic analysis, syntax, dependency grammar, graph theory application.

1. Introduction

Many systems of the machine translation, covering the needs of numerous world languages, have already been developed. For example, TAUM-METEO usually translates weather forecasts from the English language into the French language and the other way round, transmitted by the Canadian radio (Schwanke, 1991). SPANAM and ENGSPAN systems usually translate the texts from English into Spanish and the reverse (Vasconselos and Leon, 1988). The Textile Institute in France has adopted the system TITUS IV, which provides the translations from and into the four languages: French, German, English and Spanish (Schwanke, 1991). EUROTRA has been started for the purpose of catering for the needs of the European Community, and it envisages the working out of the possible means of translations among nine European languages (King and Perschke, 1987). The system GETA (ARIANE-78), which has been created by the university of Grenoble, embraces the Russian, German, French and Japanese languages (Vauquois and Boitet, 1988). The system VERBMOBIL, created in Germany, is busy translating conversations (Kay *et al.*, 1994). A great number of other systems could be mentioned as well.

The Lithuanian language has no machine translation system so far. 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.

Machine translation can be characterized by three stages. The first stage is the analysis stage, when the text in the source language gets transformed into its intermediate representation. During the second stage, the intermediate representation of the source language gets transformed into the intermediate representation of the target language. It is only during the third stage, which is called the generation stage, when the translated text gets generated.

Analysis is usually divided into three parts: morphological analysis, syntactic analysis and semantic analysis (Hutchins and Somers, 1992). In the course of the morphological analysis, the initial form of every word is presented – be it the infinitive of a verb or a declension case. The form used by the dictionary is usually indicated, and the form used in the text is also designated.

In the course of the syntactic analysis, the syntactic structure of a sentence is presented. During the process of translation the source language structure gets changed into the structure of the target language because the syntactical functions of the same words differ in various languages. The same word can play the role of the subject in one language and the role of the object in a different language – for example, *He* (subject) *likes this film* (object). *Ihm* (object) *gefällt dieser Film* (subject).

The semantic analysis helps to destroy the ambiguity of the initial words and to select the adequate words in the target language (Langer, 1996).

The translated text gets generated also in the course of the three stages, which are the following: the semantic generation, the syntactic generation and the morphological generation.

Among the formalizing works carried out in the Lithuanian language, the lemmatizing of the Lithuanian language, worked out by V. Zinkevičius (2000), could be used during the first stage of the machine translation, i.e., during the morphological analysis of the Lithuanian text, and also during the morphological generation of the Lithuanian sentence. There are no publications, which could direct our attention to the second stage of the analysis, i.e., to the syntactic analysis of a sentence. That is why the creation of the automatic syntactic analysis of the Lithuanian language and syntactic generation of the Lithuanian sentences should be considered to be the main task in this sphere of research at the present time.

The possibilities of adopting the systems of the syntactic analysis and generation, which have been created in other languages, are very limited because there should be noted very great differences existing among the Lithuanian language and the other Indo-European languages, which have already developed similar systems.

Nearly all machine translation systems use some form of tree to represent the structure of the sentence being translated. There are two basically different types of tree representation: phrase structure and dependency. Phrase structure based method originated from the American structuralism, and the method of the dependency grammar comes from European tradition through such linguist as Tesnièrre (Batori *et al.*, 1989). The differences

between both methods become obvious when in both cases we analyze the structures of the same sentence. For example: *The boy gave the girl an apple* (Kay *et al.*, 1994). The phrase structure of the sentence is shown in the Fig. 1. The sentence (S) consists of a noun phrase (NP) and a verb phrase (VP). The verb phrase consists of a verb (V) and the noun phrases (NP).

The main drawback of the phrase structure method is that the analysis is based on the word order, and in the sentence *A man walked by wearing earrings*, there is no way to attach *wearing earrings* direct to *a man* (Winograd, 1983). Such a case in the Lithuanian language, which enjoys free word order, is very frequent.

The other grammar theory, that of the dependency grammar, emphasizes that the job of syntax is not so much to group words in to the phrases, as to establish direct relationships among the words themselves. The structure of a sentence has the form of a tree with words at the nodes and with the finite verb at the root (Kay *et al.*, 1994). The dependency tree differs from the phrase structure tree. The above mentioned sentence *The boy gave the girl an apple* could be presented in the following manner, if presented in accordance with the demands of the dependency grammar (Fig. 2).

The basic difference between those two trees is the following. The phrase structure tree has words of a sentence at the leaf nodes and syntactic categories (sentence, noun phrase etc.) at all other nodes. The dependency tree has words of a sentence at each node (Winograd, 1983). The structure suggested by the dependency grammar is not linked to the word order in a sentence. That is why the dependency grammar has appealed most to languages with relatively free word order (Kay *et al.*, 1994).

While attempting to make the syntactic analysis and generation of the Lithuanian language automatic, one cannot make use of the model of the tree of the dependency
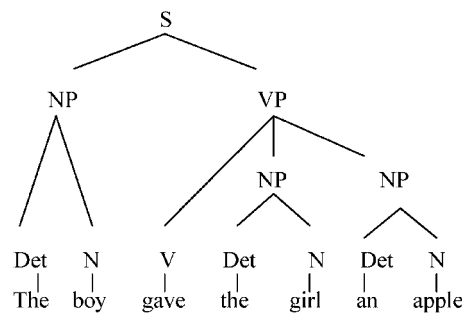

Fig. 1. Syntactic structure of the sentence *The boy gave the girl an apple* (the phrase based method).
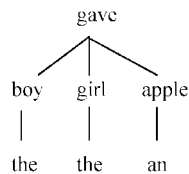

Fig. 2. Syntactic structure of the sentence *The boy gave the girl an apple* (the dependency method).

grammar either, because a tree is unable to reflect all the syntactical information that a Lithuanian sentence contains; i.e., the tree is unable to demonstrate all the relationships which exist among the words of a Lithuanian sentence. The statement, made by the initiator of the dependency grammar Tesnièrre, emphasizing that a word can be dependent on one word only, whereas it can have several words dependent on itself (Glück, 1993), is not applicable to the analysis and generation of the Lithuanian language either. A predicative attribute in the Lithuanian language is formally dependent both on the predicate and on the subject or the object.

2. Peculiarities of the Lithuanian Language

The principal feature of the Lithuanian language is the fact that the language has very many forms. Nearly all the inflectional parts of the language have 24–28 forms. For example, the English word *two* has five forms in the nominative case alone. The phrases: *two boys, two girls, two years*, *two doors* and *two horses* respectively get translated into *du berniukai, dvi mergaitės, dveji metai, dvejos durys* and *dvejetas arklių* . Each of these forms has six cases. All of them mould 30 forms of the numeral *two* alone.

The quantity of the forms of the words in the Lithuanian language has determined their role as the major means of expressing syntactic relationships in a sentence. Every language has its own grammatical means, which indicate the meaningful links among the words in a sentence. Mostly those are the forms of words, auxiliary words, such as prepositions and conjunctions, as well as the word order in a sentence. Separate languages give preeminence to different means, able to express the relationships among the words. For example, the English language accords preference to the word order, which is seconded by auxiliary words, whereas the forms of words are given the last role to play in the syntactical signification of a sentence. In the Lithuanian language, on the other hand, the relationships among words are firstly indicated by the forms of the words, wherein the system of cases comes first in the gradation of signification of their roles. The auxiliary words follow suit, and the last place is accorded to the word order of the Lithuanian sentence (Labutis, 2002). Thus the word forms express nearly all the syntactic relationships of a sentence in the Lithuanian language.

3. Problems Connected with the Predicative Attribute

Predicative attribute indicates that quality of a subject or an object, which is prominent or actual in the course of the action the sentence describes. For example, in the sentence *Tėvas grįžo namo piktas (The father returned home angry / When the father returned home, he was angry)*, the adjective *piktas (angry)* indicates that the father was angry when he returned, but it does not specify that he was an angry father as such. This is the way the predicative attribute differs from the ordinary attribute. For example in the sentence *Piktas vilkas atsėlino prie trobelės (An angry wolf came up to the hut)*, the adjective *piktas (angry)* indicates the constant quality of the wolf, i.e., the wolf is always

angry. In the case of an object: *pamačiau motulę pareinančią (I saw my mother coming home)* the word *motulė (mother)* is qualified by the participle *pareinanti (coming)* only during the time of my seeing her.

Not all the languages have the grammatical category of the predicative attribute. For example, considering the possibility of only very slight inflectional changes within the English language and the demand of the strict word order in English, we should not be surprised that the category of the predicative attribute does not exist in the English language at all (Greenbaum, 1996).

Those are only the inflected languages, which can be characterized by relatively free word order that can have the predicative attribute as such. For example, the predicative attribute is indicated in the grammar of the German language (Helbig and Buscha, 1989), where the relationships of the predicative attribute both with the predicate, and the subject or the object, get also enlarged upon. When the syntactic analysis gets formalized, only the relationship between the predicative attribute and the predicate gets indicated. U. Engel, who adopted the dependency grammar created by Tesnièrre to suit the needs of the German language, does not discuss the predicative attribute in his book "Syntax der Deutschen Gegenwartssprache" (Engel, 1994). He does not present double dependences in his tree either. This question does not seem to be topical for the German language. Though the ordinary attribute is usually made to agree with the indicated word – for example, *schöner Tag, schöne Katze, schönes Mädchen*, but the predicative attribute does not agree with the subject or the object. The German grammar states that the predicative attribute is formally coincident with the adverbial modifier ("...das prädikative Attribut mit der Adverbialbestimmung... in der äusseren Form und Position im konkreten Satz übereinstimmt") (Helbig and Buscha, 1989). The difference of its relationships with the subject from that with the object remains on the notional level exclusively. P. Hellwig calls the representation of both the relationships of the predicative attribute in a sentence "the intuitive description of a sentence" ("eine intuitive Illustration") (Hellwig, 2003).

In the Lithuanian language it would be impossible to show only one of the both relationships of the predicative attribute without losing the syntactical information because both the relationships are formally expressed. That is why the syntactic structure of the Lithuanian sentence cannot be represented in the form of a tree either. Once the relationships between the predicative attribute with the words that the predicative attribute depends on, such as the predicate, and the subject or the object, are indicated, we get a graph, which does not correspond to the definition of a tree (tree is a connected cycle free graph (Swamy and Thulasiraman, 1981)).

4. Necessity to Represent the Syntactic Structure of the Lithuanian Sentence by Using a Graph

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 <u>verärgert</u>. (*The father returned yesterday angry*)
Die Mutter kam gestern <u>verärgert</u>. (*The mother returned yesterday angry*)
Die Brüder kamen gestern <u>verärgert</u>. (*The brothers returned yesterday angry*)
Die Schwestern kamen gestern <u>verärgert</u>. (*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* will have four correspondences whose forms will correlate with the subject:

<u>Tėvas</u> vakar grįžo pikt<u>as</u>.
<u>Motina</u> vakar grįžo pikt<u>a</u>.
<u>Broliai</u> vakar grįžo pikt<u>i</u>.
<u>Seserys</u> vakar grįžo pikt<u>os</u>.

Consequently, when translating sentences from German into Lithuanian, the syntactical structure of a German sentence, shown in Fig. 3, should be changed for the syntactical structure of a Lithuanian sentence, shown in Fig. 4.

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
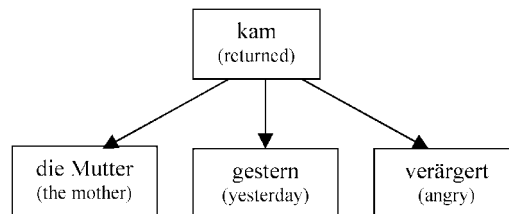

Fig. 3. The syntactic structure of the German sentence *Die Mutter kam gestern verärgert.* (*The mother returned yesterday angry.*)
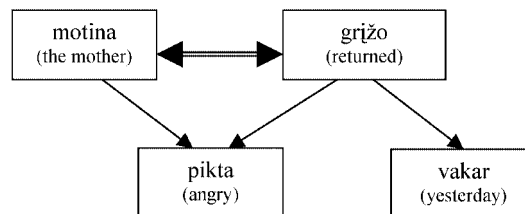

Fig. 4. The syntactic structure of the Lithuanian sentence *Motina vakar grįžo pikta.* (*The mother returned yesterday angry.*)

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 Fig. 5, 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 Fig. 6.

Those two examples testify that the two German sentence structures presenting the same graphic picture, as they are shown in Fig. 3 and Fig. 5, (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 Fig. 4 and Fig. 6. 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.

If we decide to retain the graphic representation of a tree by all means (since it is easier to process a tree than a graph by computer) and to show only the relationship, where the ending agreement takes place, i.e., only the relationship between the predicative attribute
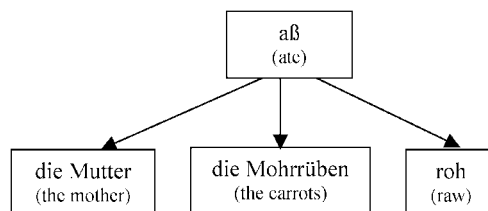
Fig. 5. The syntactic structure of the German sentence *Die Mutter aß die Mohrrüben roh.* (*The mother ate the carrots raw.*)
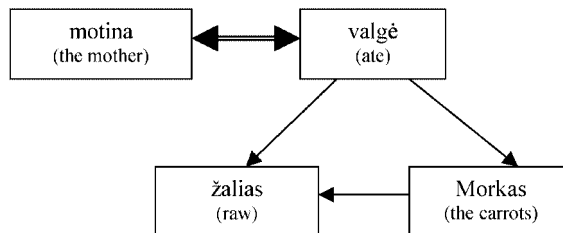
Fig. 6. The syntactic structure of the Lithuanian sentence *Motina morkas valgė žalias.* (*The mother ate the carrots raw.*)

and the subject or the object, and to disregard their dependence on the predicate, some other kind of mistakes will emerge during the generation of the Lithuanian sentence, i.e., a wrong word order.

The predicative attribute presents one of the few cases in the Lithuanian language when the word order plays an important part in determining parts of a sentence. Predicative attribute should be separated from the subject by the predicate. Here the principle of the free word order in the Lithuanian language does not get violated. Predicative attribute *(piktas)* can be placed at the beginning of a sentence, in its middle as well as at the end of a sentence. For example: *Piktas grįžo tėvas į namus (The father returned home angry). Tėvas grįžo piktas į namus. Tėvas grįžo į namus piktas.* It cannot be placed close to the subject *(tėvas),* though, because in that manner it would become an ordinary attribute. The sentence *Piktas tėvas grįžo į namus* indicates the permanent property of the subject. The predicative attribute, which is related to the object, should always follow the object *Parėjusi namo radau motulę susikrimtusią. (On coming home I discovered my mother was very sad).* If the word order is different, the predicative attribute can become an ordinary attribute *Susikrimtusią motulę ištisai lydėjo nesėkmės. (The sad mother was constantly accompanied by misfortunes).*

In case the relationships between the predicative attribute and the subject or the object are not indicated, the information regarding the flexion of the predicative attribute will be lacking during the morphological generation, i.e. it will be not possible to decide, which of words – the subject or the object – determines the flexion of the predicative attribute. In case the relationships between the predicative attribute and the predicate are missing during the syntactic generation, the place of the predicative attribute in the sentence will be impossible to determine.

Out of all the Indo-European languages the Lithuanian language was least of all changed in the course of time. It is very close to Sanskrit "the morphology of the Lithuanian language is close to Sanskrit" (Mironas, 1990). During the processing of the Lithuanian language, the antique quality of the Lithuanian language demands more of the expenditure of the software than other Indo-European languages do.

5. The Choice of the Structure of a Sentence

Since the Lithuanian language is characterized by the free word order the syntactic structure of a Lithuanian sentence could preferably be represented by using the method developed by the dependency grammar. The tree nevertheless cannot embrace all the information that a Lithuanian sentence contains. That is why a graph is suggested as the best means to represent the structure of the Lithuanian sentence. The graph is formed using the principles developed by the dependency grammar.

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. The presentation of the sentence in a way which is based on the rules of traditional grammar is chosen for the purpose of avoiding the problems inherent in the grammar of dependency. The above mentioned problems arise in the sentences containing several homogenuous predicates.

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 Fig. 7. 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 arcs outgoing of the subject and directed at the three secondary parts of the sentence may strike one as contradicting the rules of the Lithuanian grammar, because the grammar states that the object can be attached to a verb, whereas the attribute only to a noun. A typical form of a subject is that of the nominative case of a noun. In that case the noun will be extended by the attribute – for example, *mažas vaikas (a little child)*. When the subject is expressed by the infinitive, it can get extended both by the object and by the adverbial modifier. For example, *Versti tekstus automatiškai yra nelengva (To translate the texts automatically is not easy)* The generalized scheme presents all the cases. In each separate case a different arrow corresponding to the structure of a sentence will be used.

Analogously one can explain the extension of the predicate by an attribute. When the predicate is made up of a verb alone, it cannot be extended with the help of an attribute, but when the predicate consist both of a copula and a predicative, expressed by a noun, then the noun can be accompanied by an attribute. In the Lithuanian verbs, used in the Present Tense the copula is usually omitted. Consequently, the predicate often consist only of a noun, which can have outgoing arc directed at an attribute.

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.

For example, the graph of the syntactic structure of the sentence *Maža mergaitė labai nori raudono obuolio (A little girl badly wants the red apple)* (Fig. 8) or of the sentence *Vakar aš mačiau savo motulę verkiančią (Yesterday I saw my mother crying)* (Fig. 9) – will be a part of the generalized sentence structure.
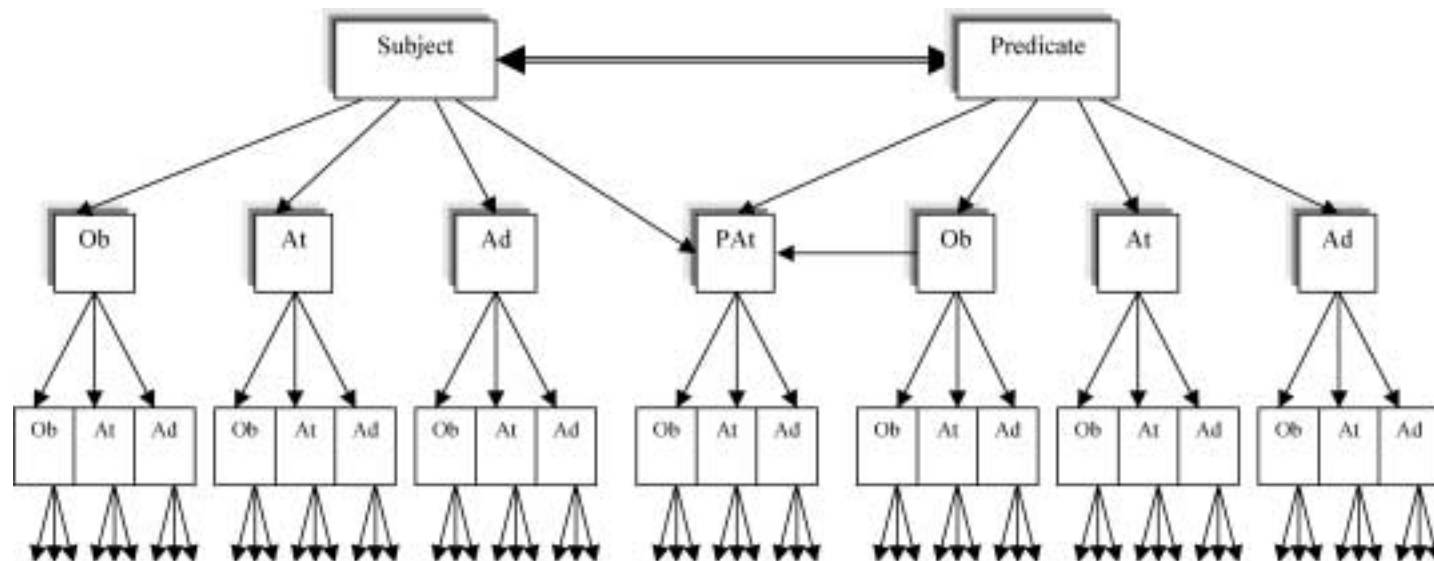
Abbreviations: At – attribute, Ob – object, Ad – adverbial modifier, PAt – predicative attribute.

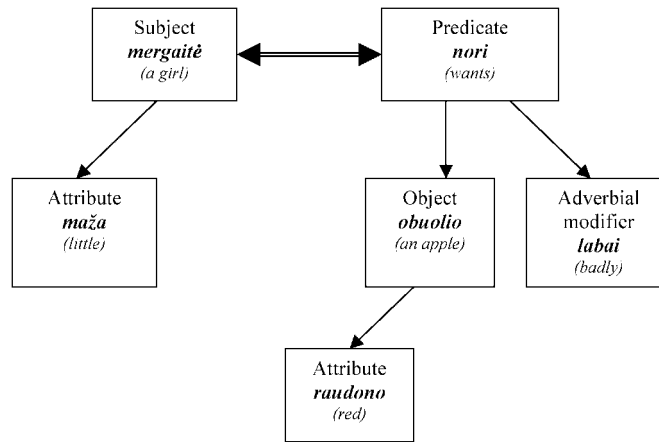Fig. 7. Generalized structure of a simple Lithuanian sentence.

Fig. 8. One path, activated in the generalized structure by the sentence *Maža mergaitė labai nori raudono obuolio.* (*A little girl badly wants a red apple.*)
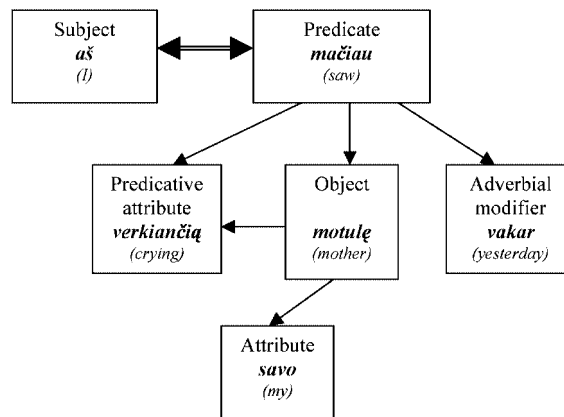
Fig. 9. One path, activated in the generalized structure by the sentence *Vakar aš mačiau savo motulę verkiančią.* (*Yesterday I saw my mother crying.*)

6. Conclusions

1. The problem of the generation of a Lithuanian sentence has been considered.
2. It has been shown that the tree-like syntactic structures, applicable to other languages, lack the information enabling us to generate a faultless Lithuanian sentence.
3. It has been proved that the usage of a graph is indispensable for the syntactic structure of a Lithuanian sentence.
4. The obtained results can be used when creating the machine translation system from other languages into the Lithuanian language.

References

Batori, I.S., W. Lenders and W. Putschke (1989). *Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications.* Walter de Gruyter Verlag, New York.

Engel, U. (1994). *Syntax der deutschen Gegenwartssprache.* 3rd ed. Erich Schmidt Verlag, Berlin (in German).

Glück, H. (1993). *Metzler-Lexikon Sprache.* Verlag J.B. Metzler, Stuttgart (in German).

Greenbaum, S. (1996). *Oxford English Grammar.* Oxford University Press, Oxford.

Helbig, G., and J. Buscha (1989). *Deutsche Grammatik. Ein Handbuch für Ausländer.* VEB Verlag Enzyklopädie, Leipzig (in German).

Hellwig, P. (2003). Manuscript (in German).

Hutchins, W.J., and H.L. Somers (1992). *An Introduction to Machine Translation.* Academic Press, London.

Kay, M., J.M. Gawron and P. Norwig (1994). *Verbmobil. A Translation System for Face-to-Face Dialog.* Centre for the Study of Language and Information, Stanford.

King, M., and S. Perschke (1987). EUROTRA. In S. Michaelson and Y. Wilks (Eds.), *Machine Translation Today*: *the State of the Art*. Information Technology Series, Edinburgh University Press, Edinburgh. pp. 373–391.

Labutis, V. (2002). *Lietuvių kalbos sintaksė.* (*Syntax of the Lithuanian Language*). 3rd. ed. Vilniaus universiteto leidykla, Vilnius (in Lithuanian).

Langer, S. (1996). *Selektionsklassen und Hyponymie im Lexikon*: *semantische Klassifizierung von Nomina.* Dissertation, Universität München (in German).

Mironas, R. (1990). *Pančatantra* (*Pancatantra*). In D. Navickaite (Ed.) (Translated from Sanskrit). *Raštijos paminklai* (*Monuments of Writings*). Mintis, Vilnius (in Lithuanian).

Schwanke, M. (1991). *Maschinelle Übesetzung. Ein Überblick über Theorie und Praxis.* Springer-Verlag, Berlin (in German).

Swamy, M.N.S., and K. Thulasiraman (1981). *Graphs, Networks and Algorithms.* John Wiley & Sons, New York.

Vasconcellos, M., and M. Leon (1988). SPANAM and ENGSPAN: machine translation at the Pan American Health Organization. In J. Slocum (Ed.), *Machine Translation Systems.* Cambridge University Press, Cambridge. pp. 187–235.

Vauquois, B., and Ch. Boitet (1988). Automated translation at Grenoble University. In J. Slocum (Ed.), *Machine Translation Systems.* Cambridge University Press, Cambridge. pp. 85–110.

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

Zinkevičius, V. (2000). Lemuoklis – morfologinei analizei. (Lemmatizer for the morphological analysis). *Darbai ir dienos*, **24**, 245–273 (in Lithuanian).

D. Šveikauskienė is a postgraduate student at the Institute of Mathematics and Informatics Vilnius. The trend of the research is machine translation. The topic of her thesis is "Formalization of Syntax of the Lithuanian Language". Her scientific adviser is professor L. Telksnys.

Lietuvių kalbos sakinio sintaksinės struktūros pavaizdavimas grafu

Daiva ŠVEIKAUSKIENĖ

Straipsnyje aprašomas naujas sakinio sintaksinės struktūros pavaizdavimo būdas. Lietuvių kalbos sakiniams netinka medžio tipo sintaksinės struktūros, kurios sėkmingai naudojamos kitų kalbų sakiniams pavaizduoti. Medis iš pricipo negali atspindėti visos sintaksinės informacijos, esančios lietuviškame sakinyje. Todėl, parodžius skirtumus tarp lietuvių kalbos ir kitų kalbų, pagrindžiama, kad lietuvišką sakinį būtina pavaizduoti grafu.

Straipsnyje taip pat pateikiama apibendrinta lietuvių kalbos sakinio struktūrinė schema, kuri apima visus galimus vientisinio sakinio atvejus. Kiekvienas konkretus sakinys turėtų aktyvuoti toje schemoje vieną kelią.