ISSN 1392–124X (print), ISSN 2335–884X (online) INFORMATION TECHNOLOGY AND CONTROL, 2014, T. 43, Nr. 4

Accuracy of the Parsing of Lithuanian Simple Sentences

Daiva Šveikauskienė

Institute of Lithuanian language
Vileišio str. 5, Vilnius, Lithuania
e-mail: daiva.sveikauskiene@lki.lt

Laimutis Telksnys

Institute of Mathematics and informatics, Vilnius University
Goštauto str. 12, Vilnius, Lithuania
e-mail: laimutis.telksnys@mii.vu.lt

crossref http://dx.doi.org/10.5755/j01.itc.43.4.6700

Abstract. The problem of the parsing accuracy of simple sentences is solved. The case of the language with high inflexion is investigated. The task is addressed by the example of the Lithuanian language, which one root can give hundreds or even more than one thousand word forms. The method of estimation of the parsing accuracy of the simple sentences of such language is given, which is based on the usage of knowledge of language consistent patterns. It is taken note, that in the case of language with high inflection and small number of users the usage of the statistical data on language is strongly restricted. The method of the accuracy estimation of the parsing of simple sentences is presented. The algorithm of implementation of the software is described. The validity of the propositions is proved by experiments. The material of the Lithuanian corpus was used for the experiments. The recommendations are given for the increasing of the accuracy of the parsing of simple sentences for the languages with high inflection.

Keywords: natural language processing; parsing; rule-based method of syntactic parsing.

1. Introduction

The newest investigations of natural languages define a conception that language is a multi-dimensional variable. Its computer features are the N-th dimension, which we do not know and this hampers our progress in natural language processing [18].

Specific features of the Lithuanian language also hamper the progress in the computerization of it. They prevent the use of the experience of other languages. The structure of the Lithuanian language differs essentially from the English language which is best computerized out of all languages. „It is naive to think: The methods, which are successfully applied for the English language, are suitable for the other languages too.“ [5]. That is the reason because here emerges the question as to endangered languages. According to the data of the META-NET project the Lithuanian language belongs to the Group of the worst computerized languages in Europe [24]. English language was the first one for which the automatic syntactic parsing has been created [11]. However, in the English language, word order in a sentence almost always determines the syntactic function of a word. Exceptions, that is, when the deciding factor is endings, are very rare, for example *I know Danny, and Toni knows me* versus *I know Danny and Toni know me* [26].

In the Lithuanian language word order in a sentence has hardly syntactic information at all and the syntactic function of a word is determined by its ending [15]. In the English language the extraction of syntactic information from word endings is absolutely unforeseen in parsing. Therefore a new distinctive method was needed for the Lithuanian language.

In parsing of the English language, the structure of a sentence is represented by a tree and the English syntax is described by context-free grammar rules, type II according to Chomsky [4]. Allen [2] describes the parsing method, which well reflects the features of the English language - a strict word order in a sentence, based on which the parsing is performed. The English language has grammatically forced word order in a sentence [15]. Subject should necessarily be in the first position, predicate in the second one, next followed by objects and adverbial modifiers. In the German language, word order is not so restricted,

though a strong requirement for the predicate to be at the second position still remains, however, the subject can exchange places with an object, i.e., the object can be in the first position, while the subject in the third one, after the predicate. Therefore another way was chosen for representing the syntactic structure of German sentences proposed by Tesniere [23]. It is more suitable for languages with a free word order, because the syntactic structure of a sentence is not related with the arrangement of words in a sentence [13]. This idea is sustained by Gomez-Rodriguez, Carroll, and Weir [7]: „dependency parser is important when parsing free word order languages".

Russian is a morphologically rich language with a rather free word order. Russian parsers are mostly based on the dependency tree representation [6].

What is of interest here, in the Russian syntax, a tradition to form a dependency tree by a binary principle prevailed. The linking relation between words are treated in a particular way: even a conjunction, for example *and* or *in*, in a binary tree is supposed to be equivalent to other words that can be part of sentence, that is, can have syntactic function [9].

The main problem in phrase based grammar is non-projective trees. Non-projectivity is a common case for the languages with the free word-order, such as Lithuanian. Dependency parsing methods are able to deal with non-projective trees. Thus the dependency grammar tree was chosen for the Lithuanian language, because it better reflects the features of languages with a free word order. However, the dependency tree was been modified, because some of the Lithuanian sentences can't be represented as a tree without loss of the syntactic information. In the dependency tree each word may have several dependent words, but it itself can depend only on one word [10].

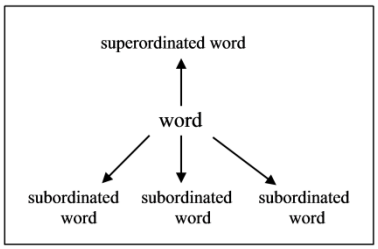

Figure 1. Structure of the word relationships in dependency grammar (Hellwig [10])

In the Lithuanian language some parts of the sentence, for example, predicative attribute, compound object, and others, formally depend on two words, therefore a tree sometimes can't reflect all the syntactic information, which the Lithuanian sentence contains. When giving the structure of a Lithuanian sentence, only the principle of representing the syntactic relationships was taken from the dependency grammar, i.e. direct relationships between words in a sentence are indicated. However, the structure itself is formed on the basis of traditional grammar, treating the subject and predicate as the main parts of a sentence (more in details Šveikauskienė [21]).

The newest method of parsing is statistical. Prins [19] describes in his dissertation how to make a natural language parsing by the probability method, using finite-state automaton. In the case of the natural language model this means that the states correspond to words". The model was tested on 7000 sentences from Dutch newspapers and the accuracy obtained for tri-gram models was 94.96% [19]. However this isn't the parsing of the full sentence, but only word combinations, consisting of 3 words, are defined.

By studying on the current state of the parsing for Russian only syntactic pair detection was assessed too. "The main assumption of expertise was the following: There is no 'correct' parsing algorithm" [6].

2. Problems by usage of statistical methods applied for the Lithuanian language

Hwa [12] notes: "...statistical parsing relies on using large quantities of annotated text as training examples." The Lithuanian language has very little syntactic annotated corpus: only 1500 sentences [14].

Statistical parsing of languages with high inflexion and free word order is more labor-intensive and needs larger amount of annotated texts than uninflected languages. Because the large number of endings the amount of the different word forms increases considerably.

One root gives in the English language 4-5 words with different endings: *-s* for the third person in present simple tense singular or for the noun plural; *-ed* for the past simple tense; *-ing* for the present continuous tense or for the noun; *-er* for the comparative form; *-ly* for the adverb. One root in the Lithuanian language gives from a few hundred to more than a thousand word forms with different endings. As an example we can provide the word *to write* - *rašyti*, whose root is *raš*. It gives 1265 different word forms: 285 forms for the noun, 240 forms for the adjective, 94 forms for the verb and 646 forms for the participle. English word *write* gives *writes, writer, writers, written*, *writing, writ* - 7 word forms with different endings.

When using the statistical method all these word forms must be in the corpus, that is, we need to have 1,000 times larger amount of the corpus than English. Currently we have 1,000 times less corpus than English language [17], thus we need to enlarge our corpus nearly 1 million times and it is hard to realize. There is little promise to have sufficient parallel corpus for Lithuanian language, because of the small number of users of the Lithuanian language, that is, users who produce them. There are no possibilities to create sufficient database for the statistical parsing of Lithuanian language. Whereas rule based parsing do not need any corpus and gives good results already now.

Rule-based parsing is necessary for the Lithuanian language both as part of the rule-based machine translation system and for the syntax based statistical machine translation, because the string-to-string machine translation systems do not give satisfactory results, for example Google. One can judge on its performance only by work of the translation system itself. A user pool showed the following results: doctors, who have not studied English, translated articles on medical topics with Google and argued that it was impossible to understand what the article was about. On the day of birthday of the author of the number *pi*, information delivered by Google on him was translated into Lithuanian. The first 7 sentences have been proposed to read to a Lithuanian with a general education and he was asked what the paper was about. His answer was: "Somewhat date". That much information was given from 7 sentences, translated into Lithuanian by Google statistical translation system, for a man whose native language was Lithuanian.

TILDE IT uses the statistical machine translation too and as the main problem points out the sparing parallel texts [1].

Yamada [27] presents syntax based statistical translation model tree-to-string. Wang [25] describes the syntax based statistical methods of machine translation: string-to-tree and tree to tree.

Thus we created the rule-based parsing producing tree structures of Lithuanian sentences without corpus. The rule-based parsing uses the method which takes into account specific features of the Lithuanian language – high inflexion (for more details see Šveikauskienė [21]) and free word order: in principle, any part of a sentence in Lithuanian can be either at the beginning of a sentence or in middle, or at the end of it (more details in Šveikauskienė [22]). Characteristic feature for the Lithuanian language is following: the position of the word in the sentence almost does not have any syntactic information – it is cumulate in the endings of words.

Statistical machine translation does not give as good results for the highly inflected languages, as for the English-French languages the case is. English and French languages are similar in their structure. Word order has the main syntactical information. The word is usually given in lemma form. During the translation this lemma form is transmitted. Word order determines the meaning of the sentence. The sentence *The dog sees the cat* denotes, that the cat is observed by the dog. If we need the opposite meaning, we must change the word order in the English language: *The cat sees the dog*. In the Lithuanian language we do not need to change the word order. We can make this with the help of word endings. The Lithuanian sentences *Šuo mato katę (The dog sees the cat)* and *Šunį mato katė (The cat sees the dog)* correspond to the English sentences with different word order. Lithuanian sentences *Šunį mato katė* and *Katė mato šunį* have no differences in the meaning [5]. They only differ in stylistic or rhetoric approach. The users of the Lithuanian language try to find the agreeing endings of words in the sentence. If it fails, the sentence remains incomprehensible. The sequence of words in lemma form does not give any meaning, any thought for Lithuanian. It is very difficult to receive the endings of Lithuanian words from the English sentence, because one form of the English word can correspond to many forms in the Lithuanian language.

There are three basic approaches to parsing for Russian: systems, which are manually enriched with expert linguistic knowledge, automata-based systems, and machine-learning systems. The manually enriched with rules systems have shown the best results [6]. Thus the statistical methods do not give the best results for Russian too.

3. The method of parsing

The hypothesis method is used for the parsing of the Russian language [16] which is close to the Lithuanian language in its structure. However, if we succeed to create a formal description of the problem to be solved, programming is simplified a great deal. Therefore it has been decided to prepare a formal description of Lithuanian syntax and in accordance with it to seek direct syntactic relations between words. The main steps of the parsing are following (Fig. 2):

Step 1: Decompose syntactic functions according to the parts of speech by which they can be expressed, for example, attribute expressed by an adjective, attribute expressed by a participle, attribute expressed by a numeral and so on.

Step 2: Continue the decomposition according to morphological categories typical of each part of speech: case, number, gender, tense and the like.

Step 3: Find out which semantic features can determine the assignment of a morphological form to a certain syntactic function. For example, the accusative case of a noun usually indicates an object: *dainuoti dainą - to sing the song*, but the accusative case of the noun, which has the feature of time, performs the function of the adverbial modifier of time: *dainuoti naktį - to sing at night*.

Step 4: Write the obtained information (Step 1 to Step 3) in BNF (Backus and Naur Form). A formal description of the attribute can be given as an example. The results of step 1 in BNF are as follows:

```
<ATTRIBUTE> :: = <ATTRIB-AFJECTIVE>|
                 <ATTRIB-PARTICIPLE>|
                 <ATTRIB-NUMERAL>;
```

Step 5: Direct syntactic relationships are looked for according to the coincidence of morphological categories, using the data obtained in the Step 4. At the same time it is taken in account what can be interpolated between two words, directly related by the syntactic link.

Step 6: If the syntactic relations have been established for all the words, we state that the sentence

structure is correct. If after processing all words of the sentence, at least one of them remains without link to another word, such a sentence is supposed to be impossible. In the case of a negative result, iterations are used. Another alternative of morphological data is taken for the word without the syntactic relationship, and the parsing of the sentence starts again.

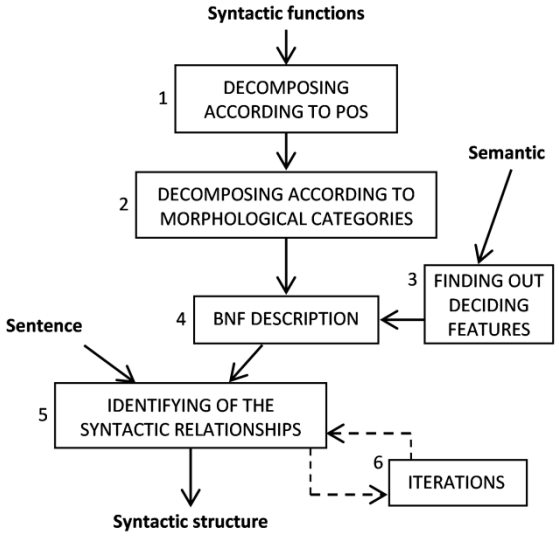

Figure 2. Steps of the parsing method

3.1. Usage of iterations

An algorithm for parsing of simple Lithuanian sentences is described in Šveikauskienė [22]. In order to improve the results of parsing, iterations are used. After the negative result of the parsing, that is, when there remain not parsed words in the sentence, the syntactic analysis tries to correct its errors. It looks for a possible morphological alternative of those words, which the syntactic relations were not found for, and analyses the sentence once more. As example of a detailed parsing with iterations, we can give the analysis of the sentence *Dėvėtais drabužiais prekiaujančios įmonės tvirtina užsiimančios ne vien prekyba - Companies, which sell second-hand clothes, state they don't only deal with market*. The negative result after the first iteration is shown in Fig. 3. The wrong structure was produced because of the following reason: The morphological analysis presents the morphological categories in the same order as they are located in the grammar: numbers - first the singular, then the plural; cases - first the nominative, next genitive, and so on. In this sentence coincide singular nominative and singular instrumental for the word *prekyba - market* and as the first alternative was taken nominative. The genitive singular and nominative plural coincide for the word *įmonės - companies* and as the first alternative was taken the variant of singular genitive. Therefore the syntactic function of the subject-VEIKSNYS was assigned to the word *prekyba - market* and the structure of the sentence was obtained as shown in Fig. 3.

However, the word *įmonės - companies* remains unparsed. Then the sentence is transmitted to the iteration procedure and another morphological alternative of this word is used for the parsing – plural nominative. Thus the word *įmonės - companies* takes the place of the subject, and another variant of the word *prekyba - market* is chosen, namely, the instrumental case. Since the predicate *užsiimti - to deal* has a strong control of the instrumental case in Lithuanian, the word *prekyba - market* is assumed as the object-PAPILDINYS (see Fig. 4).

Further, the parsing of the sentence is resumed again, that is, the search for the words that extend the main parts (subject and predicate) of the sentence. In such a way, the attribute-PAŽYMINYS of the new subject is found later followed by its extension. As a result, the correct syntactic structure is obtained (see Fig. 5).

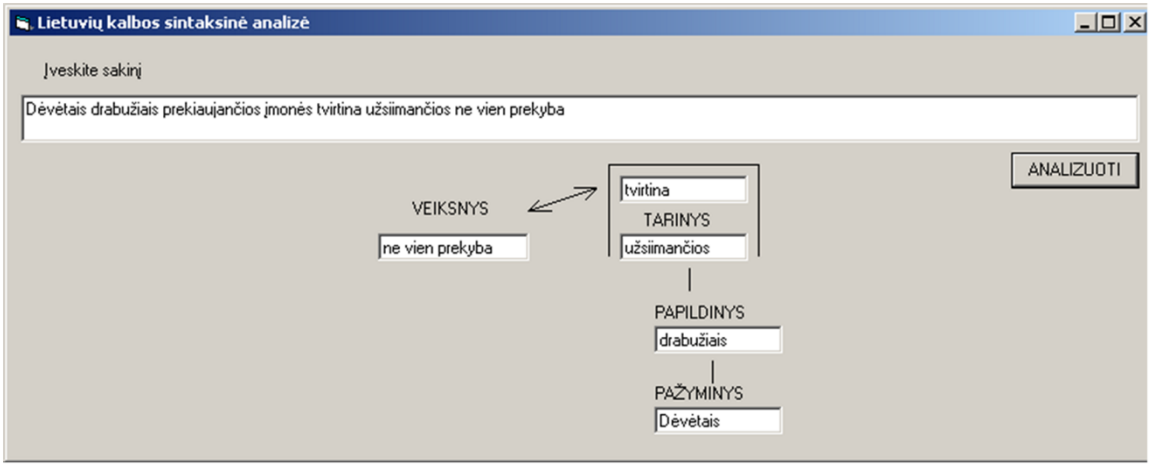

Figure 3. Structure of the sentence *Dėvėtais drabužiais prekiaujančios įmonės tvirtina užsiimančios ne vien prekyba (Companies, which sell second-hand clothes, state, they do not only deal with market)* obtained using the first alternative of morphological data

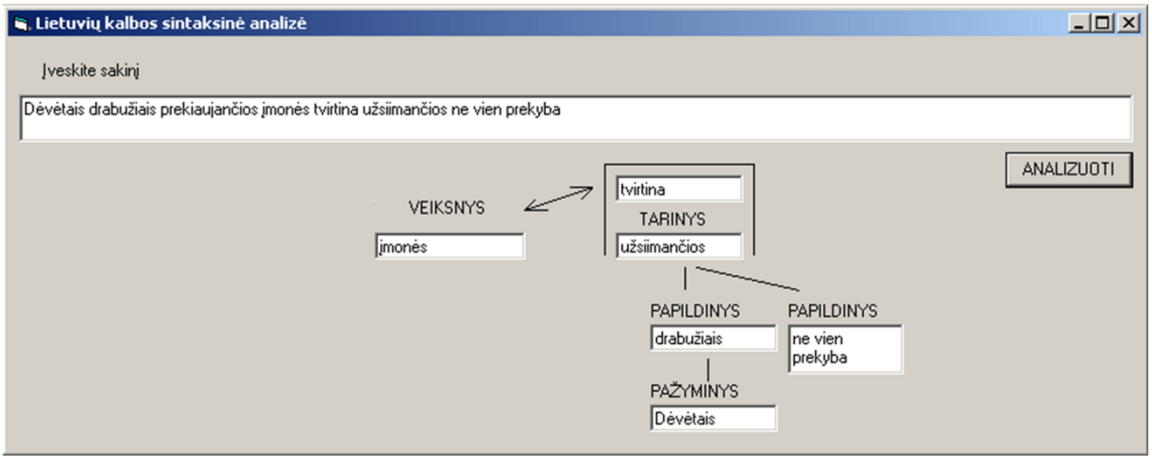

Figure 4. Structure of the sentence after the re-establishing the right subject

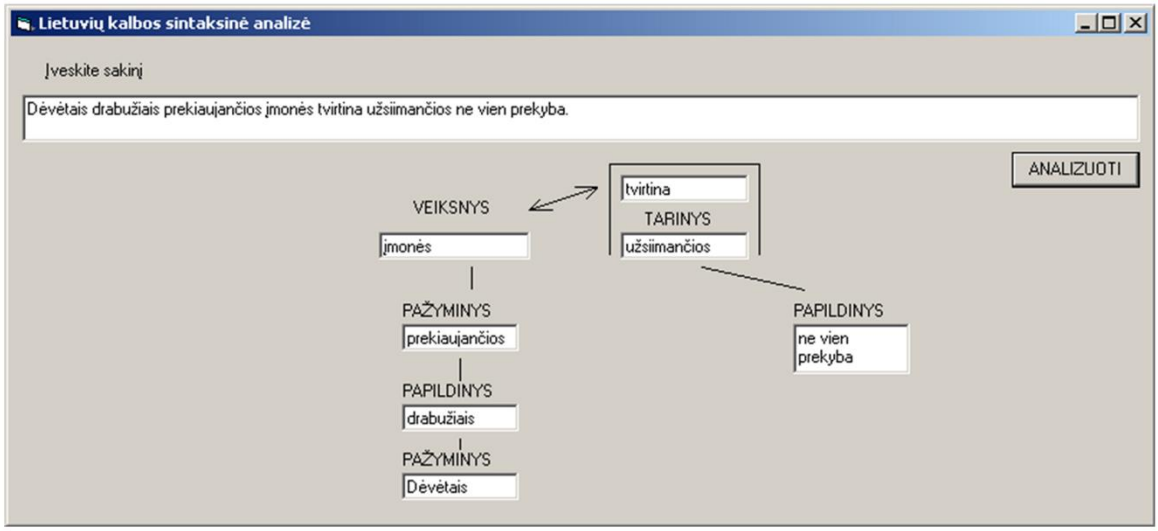

Figure 5. The final structure of the correctly parsed sentence
Dėvėtais drabužiais prekiaujančios įmonės tvirtina užsiimančios ne vien prekyba
(Companies, which sell second-hand clothes, state, they do not only deal with market)
TARINYS - predicate (*tvirtina užsiimančios - state deal*),
PAPILDINYS - object (*ne vien prekyba – not only with the market*), PAŽYMINYS - attribute (*prekiaujančios - selling*),
PAPILDINYS - object (*drabužiais – clothes*), PAŽYMINYS - attribute (*dėvėtais - second-hand*)

3.2. Examples of the patterns

Sentences of quite complex structure were used for the experiment There were used the complex and homogeneous parts of the sentence; words, which do not belong to the parts of the sentence (parenthesis, address); words in quotes; as well as the sentences reflecting the specific features of the Lithuanian language: absence of main part of the sentence (subject or predicate); predicative attribute and so on.

The cases were observed, which are not exactly determined in the printed grammars of the Lithuanian language, that is, specifying words, similes and others.

Fig. 6 shows the example of the sentence with complex parts of the sentence. The predicate in the sentence *Vidinė dorovinė šeimos darna ar destrukcija nėra „gryna" dorovės principo išraiška ar stoka – The internal moral harmony or destruction of the family isn't the "pure" expression or lack of moral principles* consist of the copula *nėra-isn't* and two homogeneous predicative *išraiška-expression* and *stoka-lack*. This sentence else has two homogeneous subjects *darna-harmony* and *destrukcija-destruction*. So this sentence is complicated because it has the word in quotes *„gryna"-"pure"*. However the software successfully managed these difficulties and presented the correct structure of the sentence.

Another example is presented in the Fig. 7. The sentence *Anot jo, nereikia daryti jokių fundamentalių pakeitimų, tačiau išsiaiškinti kiekvieną smulkmeną – According to him, it is not needed to make any changes but to clear every detail* has two homogeneous predicates of the different structure: complex one (*nereikia daryti – it is not needed to make*) and simple one (*išsiaiškinti – to clear*). This sentence has the specific feature of the Lithuanian language – the absence of the subject. In the English language such

structure is not possible, that is, the sentence with the set of these words is grammatically incorrect. .One more property of this sentence is that it contains parenthesis *anot jo – according to him*.

One more example of the sentence with the specific feature of the Lithuanian language is shown in the Fig. 8. The sentence *Tėvas Dausprungas rastas klastingai nužudytas – The Father Dausprung was found craftily killed* has predicative attribute, that is, the part of the sentence which is absent in the English language [8]. This sentence illustrates the proposition: It is impossible to represent the syntactic structure of some Lithuanian sentences by a tree without loss of the syntactic information, more in details in Šveikauskienė [22]. The syntactic structure of this sentence is represented by a graph which does not satisfy the definition of the tree (tree is a connected acyclic graph [20]). This sentence does not satisfy the requirement of the dependency grammar: the word can depend on only one word expanded by it [10]. The word *nužudytas – killed* has agreeing ending with the subject, but it is not an attribute of the subject because of its positions in respect of the subject – the predicate *rastas – found* stands between them. It can't be a predicative either, because the word *rasti – find* is not copula.

The sentence indicating the insufficient theoretical research of the Lithuanian language is shown in Fig. 9. In the sentence Taigi susitikimas nėra paprastas dialogas – dviejų žmonių pasišnekėjimas – Therefore the meeting isn't the ordinary dialogue – chat of two people the words after the hyphen specify the predicative dialogas-dialogue. Specifying words are not homogeneous parts of the sentence with the specified word, because they expand it. So the specifying words are neither controlled by the

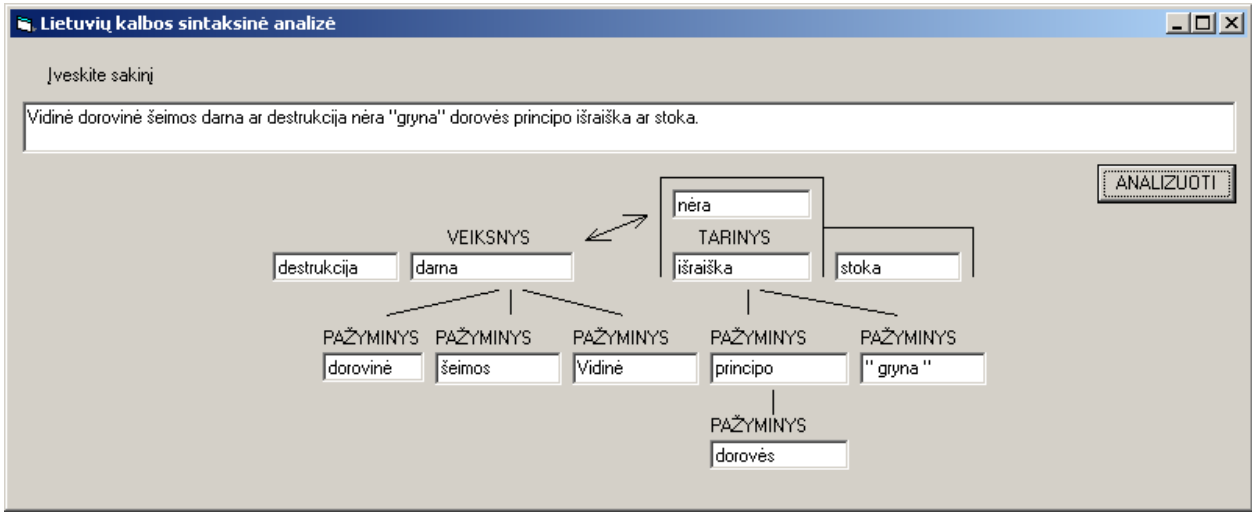

Figure 6. Syntactic structure of the sentence with homogeneous parts of sentence
Vidinė dorovinė šeimos darna ar destrukcija nėra „gryna" dorovės principo išraiška ar stoka
The internal moral harmony or destruction of the family isn't the "pure" expression or lack of moral principles

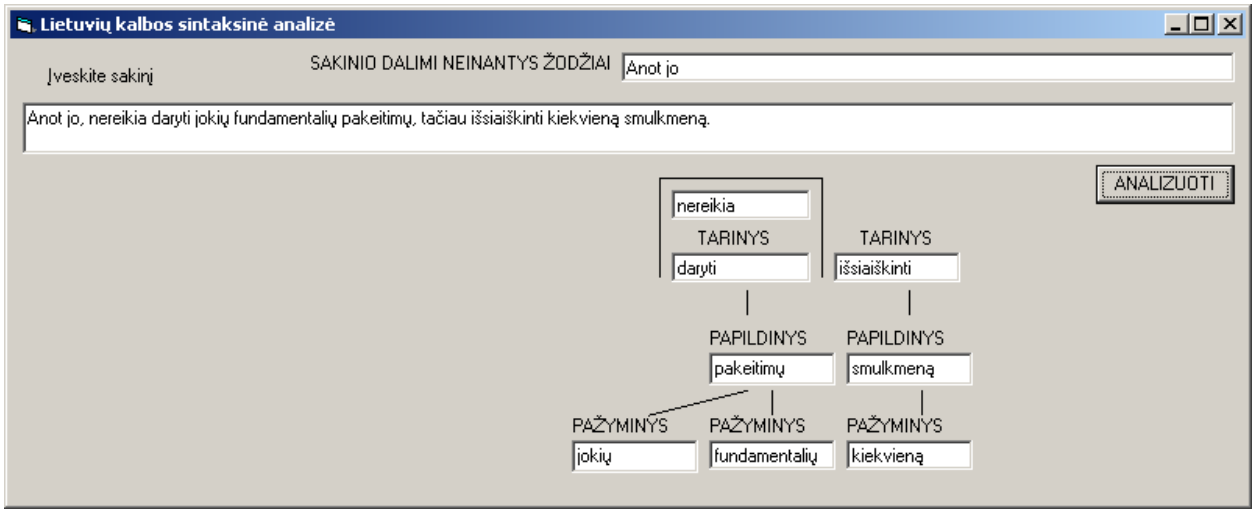

Figure 7. Syntactic structure of the sentence with the omitted subject and parenthesis
Anot jo, nereikia daryti jokių fundamentalių pakeitimų, tačiau išsiaiškinti kiekvieną smulkmeną
According to him, it is not needed to make any changes but to clear every detail

specified word nor agreed with it. Thus the specifying words can't be an attribute or object. Therefore this relationship is not exactly defined jet and requires new theoretical research of the Lithuanian language.

4. Experiment

The assessment of the parsing accuracy of Lithuanian was performed using two experiments. The sentences for the test of the first experiment were selected at random from a variety of printed publications as well as from the corpus. Certain constraints were imposed on the test sentences: prepositions, abbreviations, compound parts of sentence, and the like, were not allowed. Six samples each having 100 sentences were made. The test for the second experiment used only sentences of coherent text without omitting. Eight samples each having 50 sentences from different parts of Lithuanian Text Corpus were taken. Such a verification of the software reflects its accuracy well.

4.1. Experimental basis

The parsing of the Lithuanian language can process only simple sentences so far. With a view to show its capability when parsing Lithuanian texts, we have estimated what part of the text consists of simple sentences. To this end, we used 8 samples each having 50 sentences and 8 samples each having 100 sentences. The sentences were taken from the coherent text. The texts were chosen by an expert, using different topics of the corpus fields, therefore

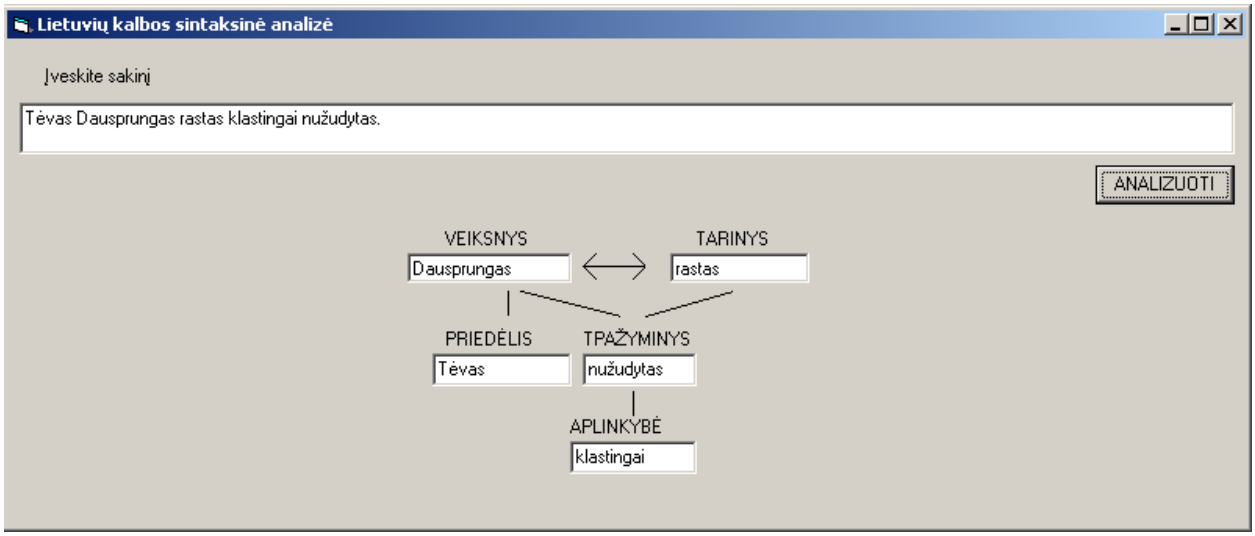

Figure 8. Syntactic structure of the sentence with the predicative attribute
Tėvas Dausprungas rastas klastingai nužudytas – The Father Dausprung was found craftily killed

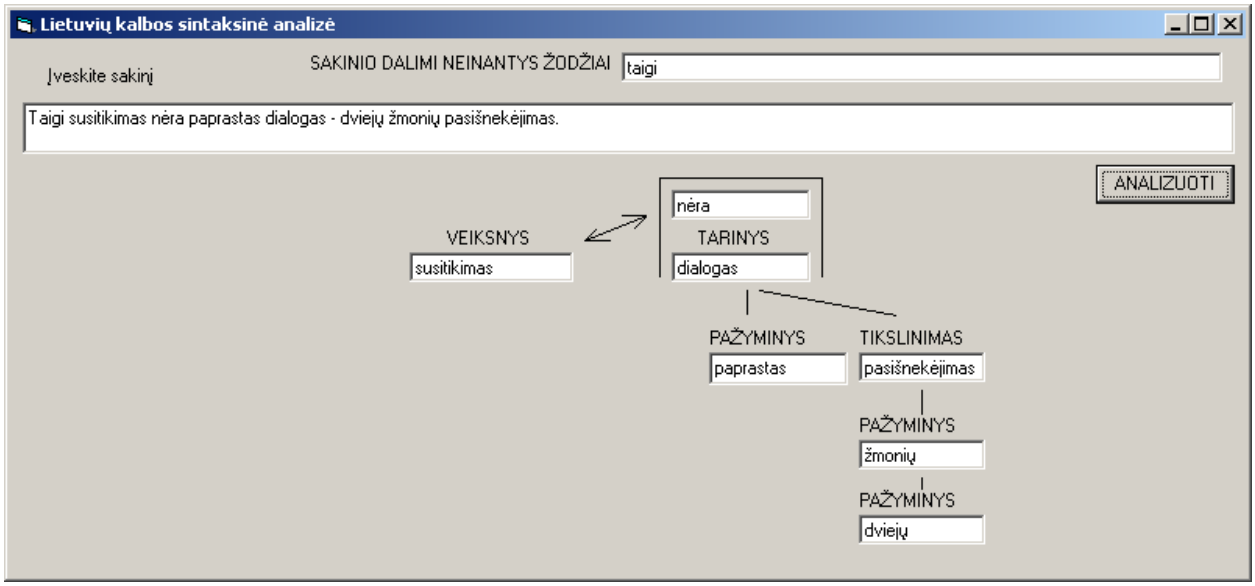

Figure 9. Syntactic structure of the sentence with the specifying words
Taigi susitikimas nėra paprastas dialogas – dviejų žmonių pasišnekėjimas
Therefore the meeting isn't the ordinary dialogue – chat of two people

these samples well represent the specifics of the Lithuanian language. Calculations were performed in the following way (see Fig. 10):

- All the simple sentences of the coherent text are consecutively enumerated up to 50 and marked (white fill in the Fig. 10).
- All the compound sentences, met upon before the 50-th simple sentence, are also consecutively numbered and marked (dark fill in the Fig. 10).
- The incoherent text (that has no solid thought or solid content, for example, authors' names, headings, and so on) is no marked (liny fill in the Fig. 10).
- The remaining part of the corpus, after the 50-th simple sentence, has not been explored (black fill in the Fig. 10).

Figure 10. An example of corpus processing for estimating what part of the text in Lithuanian consists of simple sentences

After counting compound sentences of each sample up to the 50-th simple sentences in the coherent text it has been determined: Simple sentences range from 48.0% to 72.5% (see Table 1). The above mentioned parts of corpus have also been processed analogously up to the 100-th simple sentence. In the samples of 100 each, simple sentences range from 47.2% to 73.0% (see Table 2).

Table 1. Part of simple sentences in the coherent text of the Lithuanian language, using samples of 50 sentences each

Sample Nr.	Topic of text corpus	Till 50-th simple sentence		Percentage of simple sentences till 50-th simple sentence
		Compound	All	
1	Republic periodic	54	104	48.0%
2	Local periodic	25	75	66,7%
3	Common periodic	29	79	63.2%
4	Scientific periodic	43	93	54.3%
5	Fiction	50	100	50.0%
6	Polite literature	19	69	72.5%
7	Delegated legislation	23	73	68.5%
8	Philosophic translations	52	102	49.0%
	In total:	**294**	**694**	**57.6%**

Table 2. Part of simple sentences in the coherent text of the Lithuanian language, using samples of 100 sentences each

Sample Nr.	Topic of text corpus	Till 100-th simple sentence		Percentage of simple sentences till 100-th simple sentence
		Compound	All	
1	Republic periodic	95	195	51,3
2	Local periodic	71	171	58,5
3	Common periodic	63	163	61,3
4	Scientific periodic	88	188	53.5
5	Fiction	116	216	46,3
6	Polite literature	40	140	71.4
7	Delegated legislation	37	137	73.0
8	Philosophic translations	112	212	47.2
	In total:	**622**	**1422**	**56.3**

The results obtained were slightly different. When calculating the part of simple sentences in the samples of 50 sentences each, the result was 57.6%, while in the samples of 100 sentences each, the simple sentences made up 56.3% (see Table 1 and Table 2). So, in summary, we can say that simple sentences comprise a little more than a half of all the Lithuanian texts. When calculating the part of simple sentences in the coherent text, there arose a problematic situation as to the determination of boundaries of a simple sentence. It is not clear how to treat a sentence in the case where there are several sentences within a simple sentence, for example, the case of indirect speech, or by listing something by items. Before the items we put a colon which, according to the grammar rules, is not the end of a sentence, after colon we list the items that may contain several sentences, both simple and compound ones. Where are then boundaries of a simple sentence? If we regard the entire structure as a compound sentence and do not parse it, how can we meet then the requirement 'analyse all the simple sentences without omitting any of them', if several simple sentences remain within that structure. This fact shows that the Lithuanian language has not yet been studied enough theoretically and some issues lack information.

4.2. Experimental investigations

Two experiments were carried out to determine the accuracy of the parsing of Lithuanian simple sentences method. For the first experiment, sentences taken at random from various texts were chosen, whereas during the second experiment sentences from the coherent texts were used. The first experiment was described in detail in Šveikauskienė [22]. Certain constraints on the test sentences have been introduced in it, for example, parts of the sentence can be only simple; the sentences may not contain any prepositions or words with non-Lithuanian spelling, and the like. The sentences for this test were chosen at random from various kinds of texts: printed periodicals, books, corpus, grammars, using for the test only those sentences that satisfy the imposed constrains. Table 3 demonstrates the results of the first experiment.

In the second experiment all the constraints are removed, and the software undertakes to process any simple sentence of the Lithuanian language. In view of the fact that single sentences, even taken from very different kinds of texts, may not represent in full the quality of program functionality in the coherent text, the second experiment has been carried out. It uses for the test the sentences selected by the expert, doc. G. Raškinis, from 8 different areas of the 'Contemporary Lithuanian Language Corpus', following the requirement to take all in turn simple sentences of the coherent text, without omitting any of them. It was taken by 50 first simple sentences out of all the 8 parts of the corpus. The Table 3 shows that the largest amount of errors occur due to morphology that is used as the initial data for syntactic parsing together with the Lithuanian sentence to be parsed. These errors can't characterize the performance of the syntactic parsing software because they predetermine the

Table 3. Results of the first experiment

Type of test	Type of the sample	Size of sample in sentences	Correct parsed sentences	Mistakes				Accuracy %
				Due to:			In total	
				Morphology	Semantic	Syntax		
I	All sorts of texts	100	92	5	2	1	8	92
II	1.periodic in corpus	100	92	3	0	5	8	92
	2. books in corpus	100	90	8	1	1	10	90
	3. printed periodic	100	89	5	4	2	11	89
	4. printed books	100	96	3	1	0	4	96
III	Pattern of sentences	170	170	0	0	0	0	100
In total:		**670**	**629**	**24**	**8**	**9**	**41**	**93,88**

wrong sentence structure before the beginning of the syntactic processing of the sentence. Therefore, in the second experiment, it was decided to eliminate all the errors resulting due to the inaccuracy of morphological information. This helps to better reflect the work of the syntactic parsing itself. We can distinguish two directions of improvement of morphological data for the words in the test sentences:

- Additional information is given which is not produced by the software of morphological analysis. These data are about:
  - that Lithuanian word forms on which exhaustive information is not submitted. These are words with shorted endings, some Lithuanian surnames and others. For example, if we submit a word *turėjome* - *we have had*, which has a shorted ending - *turėjom* (which is permissible in the Lithuanian language, that is, this form is quite a regular Lithuanian word), the software of morphological analysis will report: "Most probably this is the form of a word with shorted ending. Grammanal.dll cannot say anything more on the grammatical data".
  - that Lithuanian words, of which the morphological analysis software cannot recognize as a Lithuanian word at all, e.g., some Lithuanian surnames, that is, in cases when the following message received was: "The form of a proper noun. Grammanal.dll cannot say anything more about its grammatical data."
  - non-Lithuanian words, for example, names of other nations, surnames, acronyms and the like. In such cases the message was: "Submitted string of characters is not recognized by Grammanal.dll as possible form of a Lithuanian word."
- Information, submitted by the morphological analysis software, was rearranged so that the data were represented in the optimal way. The morphological analysis software of the Lithuanian language provides information on a word referring to its initial form and to all its possible variants. The variants are arranged in the order used in grammars: a noun at the beginning, next an adjective and other parts of speech; at first in singular, then in plural; at the beginning the nominative case, then genitive, and other cases. Therefore, sometimes very rarely used word forms fall into the first place, e.g., as the first variant for the word *stovi* - *stands* (*to stand* - present tense, the 3rd person), the vocative case of the noun *stovis* - *status* is submitted which formally may possibly be a Lithuanian word, however, its use is hardly imaginable. At this stage of the experiment, the variants were interchanged so that the most commonly used case was the first one. The use of iterations was helpful in the case, when both forms are equally often used, for example, the forms of feminine gender adjectives in singular genitive case and nominative case in plural coincide in Lithuanian language.

One of the reasons for errors in the parsing is the lack of semantic information. In the case where one morphological form can perform several syntactic functions, only with the help of semantic features we can unambiguously identify which syntactic function the word has. For example, a noun in the accusative case is usually an object when the verb strongly controls the accusative case (*skaito knygą* - *he reads the book*), but if the noun has a time feature, it is an adverbial modifier even under a strong control of the accusative case (*skaito naktį* - *he reads at night*). And if both such accusatives occur in one sentence (*Visą naktį ji skaitė šią knygą* - *Whole night she read this book* and *Visą knygą ji perskaitė šią naktį* - *She has read the whole book this night*), only the semantic features determine which syntactic function can be assigned to the word. The semantic features were provided for the words in the test sentences.

These kinds of error could be avoided after the automatic semantic analysis of the Lithuanian languages has been created. To this end, the works of Lithuanian linguists should be computerised. Since the automatic semantic analysis is not created for Lithuanian language yet, the file of the semantic features which has partial semantic information about the Lithuanian words, which the test sentences contain, is used for parsing. However, this file includes only those features which serve for the needs of syntax and that is not a complete computerized semantic of the Lithuanian language. The file of semantic features provides only following information, for example:

- The verb control, that is, which case the predicate requires as an object.
- Time feature.
- Such features as institution, functions and the like are indicated for nouns.
- Such features as name, surname, place-name, state and the like are indicated for proper nouns and so on.

When making the parsing, semantic data are added to the morphological data: for each lemma of a word, obtained during the morphological analysis, semantic features are looked for in the file of semantic features and the data obtained are then used to decide what syntactic function this word can have. The file of the semantic features can be edited independently of the program therefore always one may include a new word and a new feature, thus the possibilities for expansion of the system are provided.

Table 4. Accuracy of the parsing after the second experiment

Sample Nr.	Topic of text corpus	Correct analyzed sentences	Mistakes	Accuracy %
1	Republic periodic	46	4	92
2	Local periodic	43	7	86
3	Common periodic	44	6	88
4	Scientific periodic	44	6	88
5	Fiction	49	1	98
6	Polite literature	50	0	100
7	Delegated legislation	47	3	94
8	Philosophic translations	45	5	90
	In total:	**369**	**32**	**92.00**

5. Error analysis

Table 4 illustrates the accuracy of the parser, after the second experiment. The following results have been obtained: out of 400 sentences, used for the test, 368 sentences have been parsed correctly. 32 structures of sentences were wrong, which makes up 8% of all sentences. The mistakes made can be grouped into these types:

- Syntactic function of a word was identified incorrectly (such mistakes were in 5 sentences).
- The relationships between words were established wrong (in 16 sentences).
- The structure of a sentence was not formed at all (in 11 sentences).

The amount of mistakes according types in each corpus is presented in Table 5. Most errors were made in periodical texts and the least number was in fiction. This could be explained by the fact that most regular sentences are in fiction, while in periodicals more irregular sentences occur.

Table 5. Types of errors in the second experiment

Sample Nr.	Topic of text corpus	Mistakes	Structure was not formed	Wrong	
				Relationship	Syntactic function
1	Republic periodic	4	0	2	2
2	Local periodic	7	1	5	1
3	Common periodic	6	2	3	1
4	Scientific periodic	6	5	1	0
5	Fiction	1	1	0	0
6	Polite literature	0	0	0	0
7	Delegated legislation	3	1	2	0
8	Philosophic translations	5	1	3	1
	In total:	**32**	**11**	**16**	**5**

The results of both experiments are generalized in Table 6. The difference of accuracy is very little, when the situations were various. This shows the stability the parsing of Lithuanian simple sentences method.

Table 6. Generalized characteristic of two experiments

THE FIRST EXPERIMENT	THE SECOND EXPERIMENT
Sentences were selected at random from a variety of sources.	Only sentences of coherent text from corpus without omitting.
Certain constraints were imposed on the test sentences.	No constraints on the test sentences.
Six samples. In total 670 sentences.	Eight samples, each having 50 sentences. In total 400 sentences.
Accuracy – **93,88%**	Accuracy – **92%**

6. The possibilities to enlarge the accuracy of automatic syntactic analysis

Estimating the experimental results of both stages, in general, we can say that the syntactic parsing errors are caused by:

- Absence of semantic information in electronic form.
- Insufficiency of morphological information in electronic form.
- Homonyms, homographs.

In order that the software could analyse all the sentences of the Lithuanian language, it is necessary to have:

– Semantic information in electronic form about each Lithuanian word, that is, it is necessary to develop a semantic database of Lithuanian words.

– Morphological information in electronic form about all the forms of all Lithuanian words, that is, it is necessary to develop a morphological database of Lithuanian words. A part of the needed information can be achieved by using the software of Lithuanian morphology created by Zinkevičius [28]. However, its data should be supplemented with those forms which the software does not recognize as the Lithuanian word form or fails to provide exhaustive information about the given word form.
– Frequency word dictionary of word combinations that contains information about the fact that a phrase of the same word with one word is more likely than with another.

7. Conclusions

1. The problem of accuracy estimation of rule based parsing of simple sentences of high inflection and free word order Lithuanian language was discussed.
2. A plenty of word forms in the high inflection Lithuanian language, which exceeds hundreds of times the number of word forms in the English language was shown.
3. Strongly limited usage possibilities of the statistical methods for high inflection languages with small number of users were accented.
4. It was shown, that in the case of high inflection languages with small number of users the reasonable accuracy of parsing of simple sentences can be achieved by invoking the rule based method.
5. Method, algorithm and software of the parsing of Lithuanian simple sentences, which gives accuracy up to 92%, was described.
6. Stability of the accuracy of the software for the parsing of Lithuanian simple sentences for two different situations was presented.
7. The method, which was developed for Lithuanian language, can be applied for other languages with high inflexion and free word order, only the specific rules for every language must be created.

References

[1] **T. Albrektas.** About the statistical machine translation *(*In Lithuanian: *Apie statistinį mašininį vertimą).* http://blog.lituanika.lt/2010/02/apie-statistini-masinini-vertima.html.

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

[3] **V. Ambrazas (ed.).** Grammar of the contemporary Lithuanian Language (In Lithuanian: *Dabartinės lietuvių kalbos gramatika*). *Mokslo ir enciklopedijų leidybos institutas, Vilnius*, 1997.

[4] **N. Chomsky.** Systems of Syntactic Analysis. *The Journal of Symbolic Logic*, 1953, Vol. 18, No. 3, 242-256.

[5] **V. Daudaravičius.** Collocation Segmentation for Text Chunking (In Lithuanian: *Teksto skaidymas pastoviųjų junginių segmentais). Summary of doctoral dissertation. Vytauto Didžiojo universitetas, Kaunas*, 2012.

[6] **A. Gareyshina, M. Ionov, O. Lyashevskaya, D. Privoznov, E. Sokolova, S. Toldova.** RU-EVAL-2012: Evaluating dependency parsers for Russian. In: *Proceedings of COLING 2012: Parsers.* Mumbai, December 2012, 349-360.

[7] **C. Gomez-Rodriguez, J. Carrroll, D. Weir.** Dependency Parsing Schemata and Mildly Non-Projective Dependency Parsing. *Computational linguistics*, 2011, Vol. 37, No. 3, 541-586.

[8] **S. Greenbaum.** Oxford English Grammar. *Oxford: Oxford University Press*, 1996.

[9] **T. Grjaznuchina.** Syntactic analysis of scientific texts (In Russian: *Sintaksicheskij analiz nauchnogo teksta na EVM). Naukova dumka, Kiev*, 1999.

[10] **P. Hellwig.** *DUG–Dependency unification grammar*. http://www.cl.uni-heidelberg.de/~hellwig/dug-2003.pdf.

[11] **J. W. Hutchins, H. L. Sommers.** An Introduction to Machine Translation. *London, Academic Press*, 1992.

[12] **R. Hwa.** Sample Selection for Statistical Parsing. *Computational linguistics*, 2004, Vol. 30, No. 3, 253-276.

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

[14] **J. Kapočiūtė-Dzikienė, A. Krupavičius, J. Nivre.** Lithuanian Dependency Parsing with Rich Morphological Features. In: *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages,* Seattle, Washington, USA, 2013, 18 October, pp. 12-21.

[15] **V. Labutis.** Syntax of the Lithuanian language *(*In Lithuanian: *Lietuvių kalbos sintaksė). Vilniaus universiteto leidykla*, 2002.

[16] **A. Leontyeva, I. Kagirov.** Automatic syntactic analysis of the Russian Texts (In Russian: *Avtomatitsheskij sintaksitsheskij analiz russkich tekstov*). In: *Trudy X vserosijskoj konferencii RCDL'2008, Dubna*, pp. 397-400.

[17] **L. Leščinskas.** The machines will not translate Maironis jet, but... (In Lithuanian*: Maironio mašinos dar nevers, bet…*). In*: Verslo klasė*, June 2012, 28-32.

[18] **P. Mondal.** Exploring the N-th Dimension of Language. In: *A. Gelbukh, (ed.) Natural Language Processing and its Applications, Vol. 46, Instituto Politecnico Nacional, Centro de Investigationen Computacion, Mexico*, 55-66.

[19] **R. Prins.** A Finite-State Pre-Processing for Natural Languages Analysis, Ph.D. thesis. *Rijksuniversiteit Groningen, Groningen*, 2005.

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

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

[22] **D. Šveikauskienė.** A System for Automatic Syntactic Analysis of Lithuanian Simple Sentences. *Information Technologies and Control*, 2007, Vol. 36, No. 2, 221-237.

[23] **L. Tesniere.** Elements de syntaxe structural. *Klincksieck*, Paris, 1959.

[24] **D. Vaišnienė, J. Zabarskaitė.** The Lithuanian Language in the digital age. In: *G. Rehm and H. Uzhkoreit (eds.) META-NET White paper series, Springer-Verlag, Berlin-Heidelberg*, 2012.

[25] **W. Wang, J. May, K. Knight, D. Marcu.** Re-structuring, Re-labeling, and Re-aligning for Syntax-Based Machine Translation. *Computational linguistics*, 2010, Vol. 36, No. 2, 247-279.

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

[27] **K. Yamada, K. Knight.** A syntax-based statistical translation model. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, Toulouse, France, 2001, pp. 523-530.

[28] **V. Zinkevičius.** Lemmatiser for the morphologic analysis (In Lithuanian: *Lemuoklis morfologinei analizei*). In: *L. Gudaitis (ed.), Darbai ir dienos, 24 Vytauto Didžiojo universitetas, Kaunas*, pp. 245-274.

Received March 2014.