DAIVA ŠVEIKAUSKIENĖ
Institute of Lithuanian Language

Fields of research: computational linguistics, treebanking, layers of annotation.

HIGH QUALITY SYNTACTIC ANNOTATED CORPUS OF LITHUANIAN – VILSINTEKS

Aukštos kokybės sintaksiškai anotuotas lietuvių kalbos tekstynas VILSINTEKS

ANNOTATION

This paper presents a twofold annotation, which is used for the high quality annotation of the Lithuanian corpus. Comprehensive information about a sentence is given in a table and the syntactic structure of a sentence is presented in a picture. The experience of other languages is being used, and specific features of the Lithuanian language are taken into account. The insufficiency of the tree-representation for the syntactic structure of Lithuanian sentences is shown through the statistically annotated examples. The goal of the creation of the annotated corpus bearing exhaustive information is also clearly emphasized. The examples of the annotated sentences are given, which reflect the specific features of the Lithuanian language.

KEYWORDS: syntactic annotated corpus; graph representation of the syntactic structure; layers of annotation in the Lithuanian corpus; insufficiency of tree-representation for Lithuanian sentences; goal of the syntactic annotated corpus.

ANOTACIJA

Straipsnyje aptariamas dviem lygmenimis atliekamas anotavimas, naudojamas aukštos kokybės anotuotam lietuvių kalbos tekstynui sukurti. Išsami informacija apie sakinį nurodoma lentelėje, o sintaksinė sakinio struktūra nubraižoma grafiškai. Naudojamasi kitų kalbų patirtimi, atsižvelgiant į specifinius lietuvių kalbos bruožus. Medžio nepakankamumas vaizduojant lietuvių kalbos sakinių sintaksinę struktūrą parodomas statistiniu

metodu anotuotų sakinių pavyzdžiais. Straipsnyje taip pat aiškiai pabrėžiamas anotuoto tekstyno, kuriame bus sukaupta išsami informacija, kūrimo tikslas. Pateikiami anotuotų sakinių pavyzdžiai, atspindintys specifinius lietuvių kalbos bruožus.

ESMINIAI ŽODŽIAI: sintaksiškai anotuotas tekstynas, sakinio struktūros vaizdavimas grafu, tekstyno anotavimo lygmenys, medžio nepakankamumas vaizduojant lietuvių kalbos sakinius, tekstyno kūrimo tikslas.

1. INTRODUCTION

The first annotation of corpora was a part of speech tagging POST [Church 1988: 136]. Such annotation is found in the Penn Treebank. Later, other sentences, carrying the syntactic information, were introduced. They were represented by using various schemes [Atwell et al. 2000: 12]. In 2012 Köhler mentioned the following: "...there is no general standard as to how corpora should be structured and notated" [Köhler 2012: 32]. Therefore, the Lithuanian corpus VILSINTEKS is annotated, bearing in mind the specific features of the Lithuanian language – a large amount of inflexion and a free word order in a sentence.

Of the two leading types of syntactic structure representation, which are the phrase structure grammar and dependency grammar, the latter was chosen for Lithuanian because "...dependency grammar has appealed most to students of languages with relatively free word order..." [Kay, Gawron and Norvig 1994: 55].

The most common representation of the syntactic structure is a tree [Allen 1987: 41]. The name of the syntactically annotated corpus – *treebank* – is related to this term. The name of the Lithuanian treebank does not contain the word 'tree', because the syntactic structure of some Lithuanian sentences is represented by a graph with a cycle, that is, it does not meet the conditions established by the definition of a tree, as a tree is a connected acyclic graph [Swamy, Thulasiraman 1984: 33]. The acronym VILSINTEKS means VILniaus SINtaksinis TEKStynas – Vilnius syntactic corpus.

2. ANNOTATION OF THE LITHUANIAN CORPUS

Recently, much has been said and written about the low degree of computerization of the Lithuanian language. According to the data of the META-NET

project, Lithuanian belongs to the group of the least computerized European languages [Vaišnienė, Zabarskaitė 2012: 35]. The software created for other languages does not produce satisfactory results when it is applied to the Lithuanian language. It is high time to speak not about the low computerization of the Lithuanian language but about the quality of its computerization. Thus, it is worth looking back at the Lithuanian language and trying to create one's own software for the computerization of the Lithuanian language so as to produce high quality computerization of the Lithuanian language.

The first endeavors in the syntactic annotation of the Lithuanian corpus were made in 2013. The syntactic annotated corpus VILSINTEKS was created at the Institute of the Lithuanian Language in Vilnius. We aim at providing as much data as possible regarding a sentence, especially if they happen to be the data which are needed during the process of translation. Exhaustive information is given in the Prague Dependency Treebank, which has a three-level annotation [Hajič 2000: 103]. More levels of annotation in one picture could not

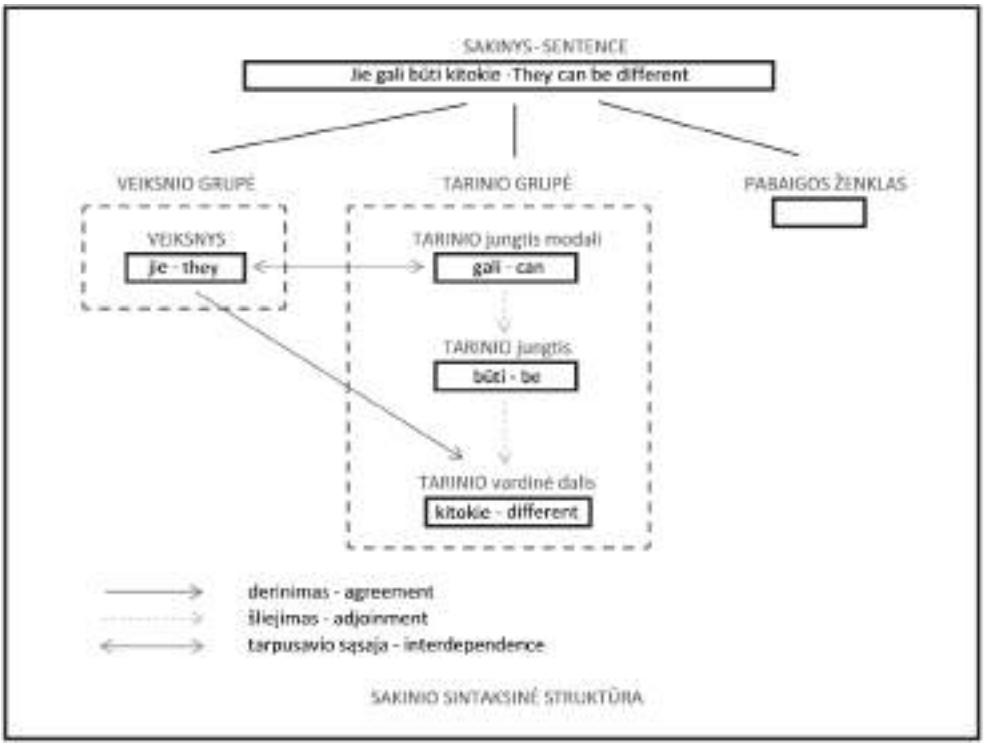

FIGURE 1. Example of the syntactic structure of the Lithuanian sentence *Jie gali būti kitokie – They can be different*

be achieved so we decided to divide the representation of information into two parts when we deal with Lithuanian sentences: a table, which contains the morphological, syntactic and semantic information, while the syntactic structure should be presented in a picture.

A graph with cycles is used for Lithuanian sentences because a tree is not able to reflect the entire syntactic information which the Lithuanian sentence contains. The predicative attribute depends on two parts of the sentence: on the subject or the object, and on the predicate. These relationships are expressed formally and could not be ignored (for more details, see Šveikauskienė 2005: 412).

Figure 1 shows the syntactic structure of a title sentence. It contains a predicative, which has formally expressed relationships with the subject and the predicate, and both relationships must be represented by annotating the corpus.

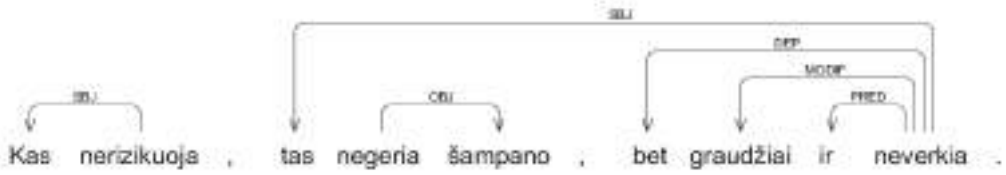

FIGURE 2. Statistically parsed sentence *Kas nerizikuoja, tas negeria šampano, bet graudžiai ir neverkia (Who does not risk, that does not drink champagne but does not cry tearfully either)* [Kapočiūtė, Nivre, Krupavičius 2013: 15]

FIGURE 3. Statistically parsed sentence *Bet štai pro medį, kuriame sėdėjau, praslinko didelis šešėlis (But here through the tree in which I sat passed a small shadow)* [Kapočiūtė, Nivre, Krupavičius 2013: 15]

The predicative must have the ending agreeing with the subject and at the same time it is adjoined to the predicate; that is, it is not an attribute of the subject.

The first attempts to syntactically annotate the Lithuanian corpus demonstrated that the software crated for other languages did not produce good results in the Lithuanian language. Some 1500 sentences were annotated statistically at Kaunas University of Technology [Kapočiūtė, Nivre, Krupavičius 2013: 12]. The presented results showed the insufficiency of the information hidden in the

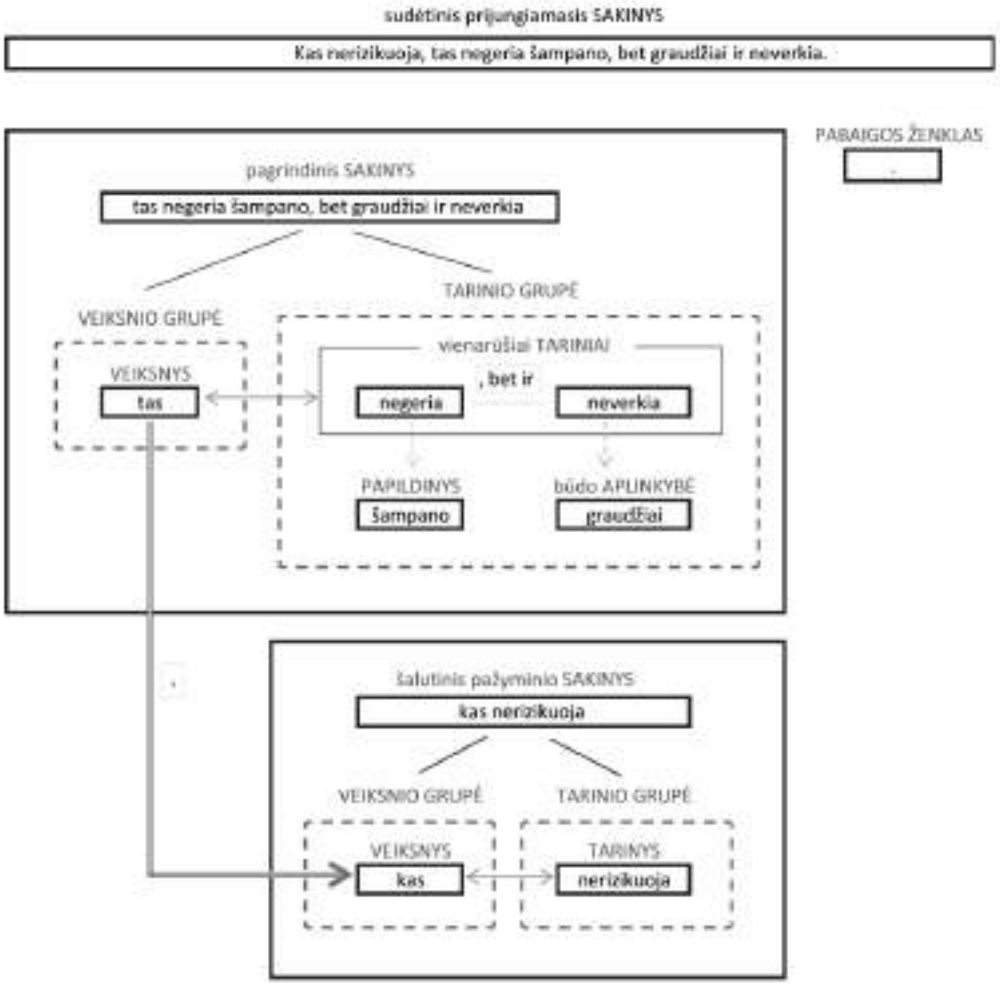

FIGURE 4. VILSINTEKS parsed sentence *Kas nerizikuoja, tas negeria šampano, bet graudžiai ir neverkia*

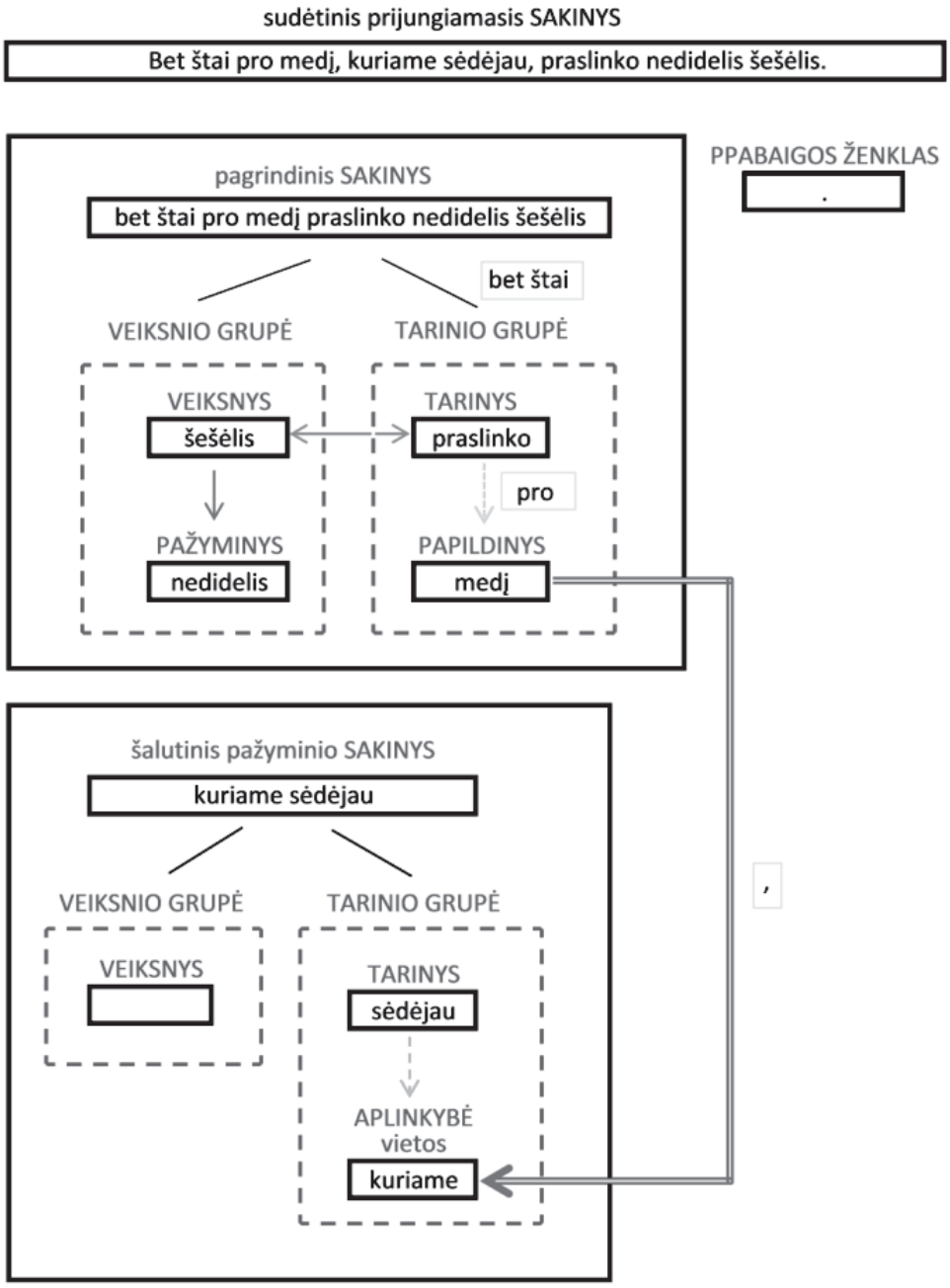

FIGURE 5. VILSINTEKS parsed sentence *Bet štai pro medį, kuriame sėdėjau, praslinko nedidelis šešėlis*

structure of sentences. The two figures illustrate two parsed sentences, both lacking very important information. The first sentence lacks the arrow between the predicate *negeria* and subject *tas* (Figure 2).

In the second sentence the agreement relationship is not shown between two words *medį* and *kuriame*, which have agreeing endings (Figure 3) when the authors in the same article wrote: "an adjective modifying a noun has to agree in GENDER, NUMBER and CASE".

The words *medį* and *kuriame* have to agree in gender and number, and this information is absent in the structure of the sentence. It is difficult to agree that the relations between the words *kuriame sėdėjau – in which I sat* and *pro medį – through the tree* are of the same type.

These sentences annotated according the method used in VILSINTEKS are shown in Figure 4 and Figure 5 respectively.

3. LAYERS OF ANNOTATION IN THE TABLE

The information in the table has three types:

- Information about the whole sentence,
- Information about the words in the sentence,
- Non-grammatical information about the words and the sentence.

3.1. Information about the Whole Sentence

Information regarding the whole sentence consists of its code, that is, its position in the corpus, its type, and its features. The feature of the sentence is its characteristic taking into account its function in the text, that is, whether it is a title, an author, a subtitle or a text sentence. The type of the sentence indicates communicative information, that is whether it is declarative, imperative, interrogative, etc., and structural information: personal, impersonal, elliptical, simple, composite sentence, etc. [Ambrazas 1997: 573].

3.2. Information about the Words in the Sentence

Each word is provided with the data about its number in the sentence, morphological data (tense, case, gender, etc.), lemma, data on its lexical semantics,

syntactic function, direct syntactic relationships with other words in the sentence, and deep cases.

Since the word order in the Lithuanian language usually does not have any syntactic information, the features of lexical semantics are very important for identifying its syntactic function. The noun in the accusative case with the feature of time is an adverbial modifier and without it – an object. If two nouns in the accusative case appear in the sentence the feature of the lexical semantics is sometimes the only criterion, which allows one to decide the syntactic function.

3.3. Non-grammatical Information about the Words and the Sentence

The stylistic information about the word is given in the table. It helps to choose the right equivalent in the other language when translating the word.

The antecedent of the pronoun is necessary. Mille, Wanner and Burga [2012: 5] describe coreferential structure, which links the pronoun with its antecedent in one sentence. The Lithuanian treebank provides the information about the antecedent of the pronoun if it is outside the sentence too. It is very important for translation because the gender of the noun (or pronoun accordingly) may differ in various languages. For example, the Lithuanian sentence *Ji buvo graži* has three translations into German. If it is "a girl", the right translation is *Es war schön*. If it is "a cat" the right translation is *Sie war schön*, and if it is "a day", the right translation is *Er war schön*. All three pronouns in Lithuanian are feminine because all three nouns are feminine, whereas in the case of the German language this pronoun has three different equivalents taking into consideration the noun it replaces.

The table contains the missing words in elliptical sentences and the omitted subject, which is expressed by a personal pronoun. It is very often the case in the Lithuanian language. We can guess it from the ending of the verb. The copula of the predicate in the present tense is usually omitted too, so the table contains these missing words.

Acronyms are represented in full words in the table.

The numbers are additionally represented in the table by numerals and numerals by numbers because the lexical expression of numbers in various languages may differ.

4. TYPES OF INFORMATION IN THE SYNTACTIC STRUCTURE

In the syntactic structure the sentence is divided at the first stage into a noun group, a verb group and the sentence-end character. The Lithuanian language has a free word order and sometimes the sentence-end character is the only means by which to determine the type of the sentence, i.e. whether it is a declarative or interrogative sentence, for example, *Tu šiandien laimėjai prizą. – You have won a prize today.* and *Tu šiandien laimėjai prizą? – Have you won a prize today?*. Figure 6 shows the syntactic structure of the interrogative sentence. The translation of the sentence is chosen according to the sentence-end character. Furthermore, the structure of the sentence is represented using dependency grammar. The head of the subject group is the subject and below are depicted the words that expand it, and the head of the predicate group is the predicate with the subordinated words located below.

Other two fields, which are very important for annotating the Lithuanian corpus, are the type of the syntactic relations between the words and semantically irresolvable word groups.

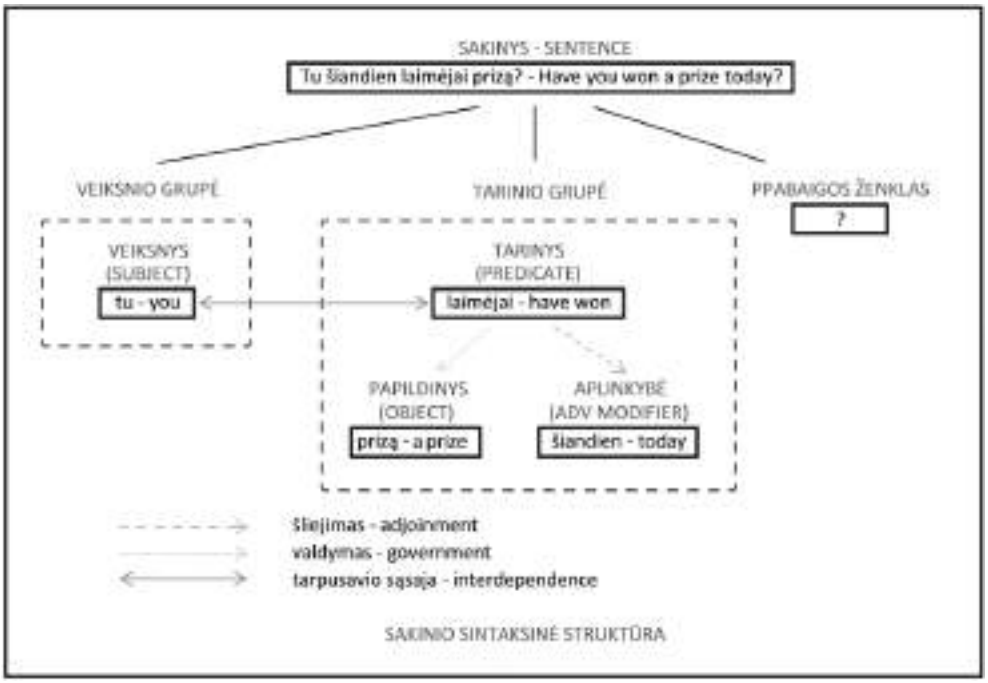

FIGURE 6. Syntactic structure of the interrogative sentence *Tu šiandien laimėjai prizą? – Have you won a prize today?*

4.1. Type of Syntactic Relation

The Lithuanian language has a considerable amount of inflexions. Thus it is very important to represent the information about the type of the syntactic relations because they are often expressed by word endings. The Lithuanian language has three types of syntactic relationships: interdependence between the subject and the predicate, coordination between two or more words of equivalent syntactic status in the sentence, and subordination between two words of which one determines the ending of the other. Subordination is divided into government, agreement and adjunction [Ambrazas 1997: 478]. Each type of syntactic relation is depicted in the structure with a different color. Interdependence is represented by a lilac bi-directional arrow because both words decide the form of one another; coordination – a yellow line without direction because the words have no influence on the form of one another; government – a blue unidirectional small dotted arrow; agreement – a red unidirectional arrow; and adjoinment – a green uni-directional large dotted arrow. Prepositions, conjunctions or punctuation marks are represented as labels of the arrows between the words. Figure 7 shows the example of the sentence with conjunction.

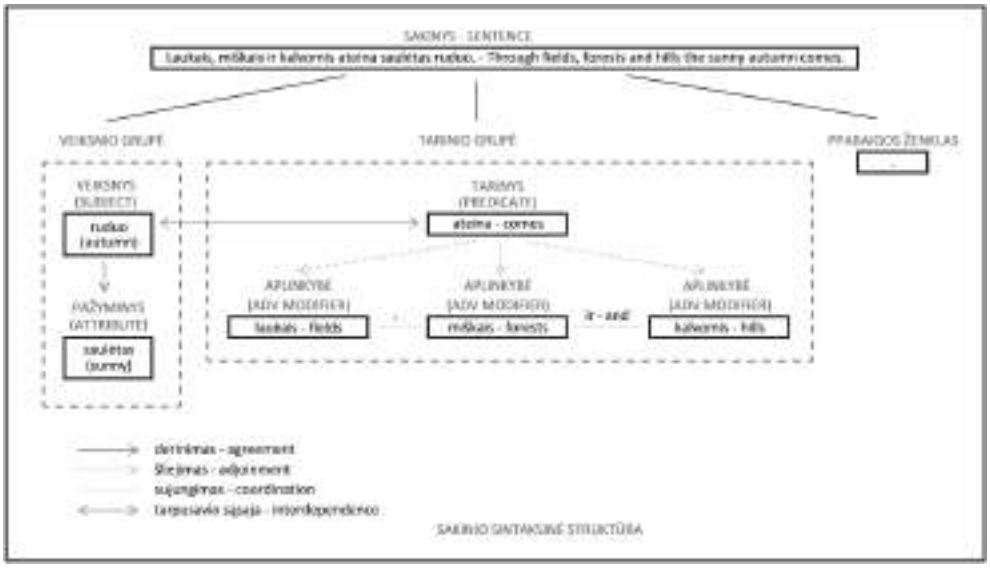

FIGURE 7. Syntactic structure of the sentence with conjunction *Laukais, miškais ir kalvomis ateina saulėtas ruduo. – Through fields, forests and hills the sunny autumn comes.*

4.2. Semantically Irresolvable Word Groups

A semantically irresolvable word group is a word combination in which one word can expand another word only when they are together and the dependent word can expand the whole group only. The relation of one word from that group with the expanding word has no meaning. Mel'chuk describes a similar case observed in the Russian language [Mel'chuk 2006: 37]. An irresolvable word group is placed into one box by representing syntactic relationships within. Figure 8 shows the syntactic structure of the sentence with an irresolvable word group: *Penkis tūkstančius hektarų žemės valdantis ūkininkas pigiai parduoda šviežias daržoves. – A farmer owning five thousand acres of land sells cheap fresh vegetables.* The verb *valdyti – 'to own'* is transitive in the Lithuanian language and requires the accusative, thus the word *tūkstančius-'thousand'* is in the accusative case. Numerals require the genitive case in Lithuanian, and that is why the word *hektarų-'acres'* is in the genitive case. The farmer semantically does not own 'thousand', though the word *thousand* is in the case controlled by the participle *valdantis-'owning'*. Rather, he owns acres, that is, the word which is in the genitive case, i.e. in the case controlled by numeral *thousand*. Consequently, the word group *tūkstančius hektarų '-thousand acres'* is semantically irresolvable and therefore both words are located in one box in the syntactic structure of the sentence. The word *five* is an attribute of the word *thousand,* so it is located in the same box. The non-agreeing attribute *žemės – 'of land'* expands the whole box, that is, the whole word group, which is located in the box, and may be replaced by other non-agreeing attributes, for example: *five thousand acres of forest*. Thus, the semantically irresolvable word group is considered one lexical unit.

Word combinations of numerals (*five* and *thousand*) have dual syntactic relationships: agreement and government. The case of the word 'thousand' (whose ending is determined by the government of the verb) determines the ending of the word 'five' – they must have agreeing endings (*jis valdo penkis tūkstančius hektarų – he owns five thousand acres*, but *penkiems tūkstančiams hektarų jis sunaudojo 10 tonų trašų – he used 10 tons of manure for five thousand acres*) and lexeme 'five' governs the word 'thousand', i.e. requires a certain ending for it. For example, in the word groups *he owns five thousand acres (jis valdo penkis tūkstančius hektarų)* and *he owns twenty thousand acres (jis valdo dvidešimt tūkstančių hektarų)* the word *thousand* has different endings in the Lithuanian language. That is why it is very important for us to depict all types of syntactic relations very precisely.

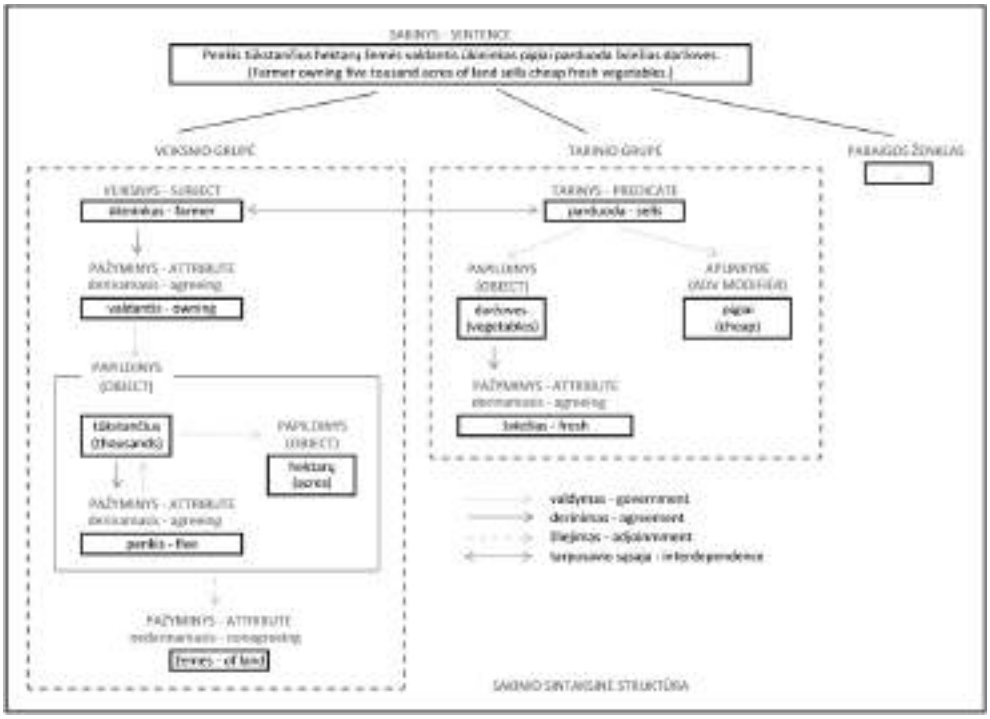

FIGURE 8. Syntactic structure of the sentence with a semantically irresolvable word group *Penkis tūkstančius hektarų žemės valdantis ūkininkas pigiai parduoda šviežias daržoves (Farmer owning five thousand acres of land sells cheap fresh vegetables)*

5. THE AIMS OF ANNOTATION

The statistical method of natural language processing, especially when used in machine translation, does not produce satisfactory results for the Lithuanian language. Statistical parsing does not provide exhaustive information either. Moreover, the software created for other languages does not give sufficient results for the Lithuanian language. V. Daudaravičius worked for a long time on the automatic syntactic analysis of the Lithuanian language and wrote the following: "Naivu manyti, kad metodai, kurie sėkmingai taikomi anglų kalbai, tinka ir kitoms kalboms. – It is naive to think that the methods, which are successfully used for the English language, are suitable for other languages too." [Daudaravičius 2012: 3].

The goal of VILSINTEKS is to collect reliable comprehensive information about the grammar of the Lithuanian language in praxis. With high quality automatic syntactic analysis created via a non-statistical approach, non-statistically

based machine translation systems are enabled, especially in the generation of Lithuanian sentences.

VILSINTEKS seeks to reflect specific features of the Lithuanian language. Exhaustive information about Lithuanian sentences will be provided in an obvious representation for the general usage. The annotated corpus will be freely accessible on the Internet. Currently, one can get examples of the sentences with the given word by using the text corpus created in Vytautas Magnus University. VILSINTEKS will provide examples of the syntactic structure with the given features, for example, a simple sentence with the pronoun as a subject, and others. The application of the Russian corpus is similarly described: it can be used in the mode of the search for examples, which illustrate the given linguistic phenomenon [НКРЯ –National Corpus of the Russian Language]

The first 500-1000 sentences of 2 million words in the corpus of contemporary Lithuanian language (journalism) will be annotated manually. Later, the annotation will be semi-automatic. The automatic syntactic analysis of the Lithuanian language thus created will be used for the annotation with the follow-up human review.

During the annotation process the software will be improved with due regard to the mistakes identified. The most important task is to prepare a well-functioning automatic syntactic analysis of the Lithuanian language, which can be used by the creation of a transfer machine translation system. Using the TRANSFER method, the second step is the syntactic analysis. It can be expected that the syntactic annotation of the corpus will serve to improve the automatic syntactic analysis.

Excel tables are used for annotation because they give the obvious representation and it is easy to get the XML format from Excel tables. The VILSINTEX information will be freely accessible in XML format too. Thus, it can be used for statistical research on the Lithuanian language and for information retrieval about the grammar of Lithuanian, among other uses.

6. CONCLUSIONS

The graph is used for the syntactic structure of Lithuanian sentences because a tree cannot reflect all the syntactic information which a Lithuanian sentence contains. Two examples of statistical annotation show the insufficiency of the tree for representing the syntactic structure of Lithuanian sentences.

A twofold annotation is used for the Lithuanian corpus: the table describes the exhaustive information about the sentence and its words, and syntactic structure of the sentence is represented in a picture.

In the table the information represented is of three types: information about the whole sentence (interrogative, elliptical, impersonal); the information about each word of the sentence (morphologic, syntactic and semantic data); and non-grammatical information (missing words of elliptical sentences, antecedent of pronoun inside and outside the sentence, number expression of numeral and vice versa, full word correspondents of acronyms).

The syntactic structure of the Lithuanian sentence is divided at the first stage into three parts: the subject group, the predicate group and the sentence-end character. The last one is very important because in some cases it is the only criterion which permits a decision as to the type of the sentence (declarative or interrogative).

Syntactic relations are depicted in the syntactic structure of the sentence with different colors and style of line according the type of relation.

Semantically irresolvable word groups are located in one box by representing the syntactic relationships within.

The goal of VILSINTEKS is to provide exhaustive and reliable information about Lithuanian sentences. It can be used for the statistical research of the Lithuanian language, and for the information retrieval regarding the Lithuanian grammar. It can also be of service by creating a non-statistical machine translation system.

VILSINTEKS aims to collect the grammatical information about the Lithuanian grammar in praxis.

The first sentences are annotated manually. Later, the annotation will be semi-automatic. Excel tables are used for annotation because they give the obvious representation for human usage, and the annotated corpus will be freely accessible on the Internet. It is easy to get the XML format from Excel tables. The VILSINTEKS information will be freely accessible in XML format too. Thus, the information of VILSINTEKS can be used for statistical research of the Lithuanian language and for the information retrieval about Lithuanian grammar.

REFERENCES

Allen James 1987: *Natural Language Understanding*. Amsterdam: The Benjamin/Cummings Publishing Company.

Ambrazas Vytautas 1997: *Dabartinės lietuvių kalbos gramatika*. Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Atwell Eric, Demetriou George, Hughes John, Schiffrin Amanda, Souter Clive and Wolcock Sean 2000: Comparing linguistic interpretation schemes for English corpora. – *ICAME* 24, 7–24.

Church Kenneth Ward 1988: A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. – ANLC '88: Proceedings of the second conference on Applied natural language processing. Association for Computational Linguistics Stroudsburg, PA, 136–143.

Daudaravičius Vidas 2012: *Teksto skaidymas pastoviųjų junginių segmentais*. Daktaro disertacijos santrauka. Kaunas: Vytauto Didžiojo universitetas.

Hajič Jan, Böhmová Alena, Hajičová Eva and Hladká Barbora Vidová 2000: The Prague Dependency Treebank: A Three-Level Annotation Scenario. – Treebanks: Building and Using Parsed Corpora, ed. by A. Abeillé. Amsterdam: Kluwer, 103–127.

Kay Martin, Gawron Mark, J. and Norvig Peter 1994: *Verbmobil: A Translation System for Face-to-Face Dialog*. Stanford: CSLI.

Kapočiūtė-Dzikienė Jurgita, Krupavičius Algis, Nivre Joakim 2013: Lithuanian Dependency Parsing with Rich Morphological Features. – *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*. Seattle, Washington, USA, 18 October, 12–21.

Köhler Reinhard 2012: Quantitative Syntax Analysis. – *Quantitative linguistics* 65, ed. by Reinhard Köhler, Gabriel Altmann, Peter Grzybek. De Gruyter Mouton.

Mille Simon, Wanner Leo and Burga Alicia 2012: Treebank Annotation in the Light of the Meaning-Text Theory. – *Linguistic Issues in Language Theory* 7(16), 1–12.

Swamy M.N.S, Thulasiraman K. 1981: *Graphs, Networks, and Algorithms*. New York, Chichester Toronto: Willey Interscience publication.

Šveikauskienė Daiva 2005: Graph Representation of the Syntactic Structure of the Lithuanian Sentence. – *Informatica* 16(3), 407–418.

Vaišnienė Daiva, Zabarskaitė Jolanta 2012: The Lithuanian Language in the digital age. – *META-NET White paper series*, ed. by Georg Rehm, Hans Uszkoreit. Berlin, Heidelberg: Springer-Verlag.

НКРЯ – *Национальный корпус русского языка 2003–2013*. Prieiga internete: http://www.ruscorpora.ru/corpora-usage.html

High Quality Syntactic Annotated
Corpus of Lithuanian – VILSINTEKS

Aukštos kokybės sintaksiškai anotuotas lietuvių kalbos tekstynas VILSINTEKS

SANTRAUKA

Pastaruoju metu labai daug kalbama ir rašoma apie tai, kad lietuvių kalba mažai kompiuterizuota. Perkama kitoms kalboms sukurta programinė įranga, kuri lietuvių kalbos atveju dažniausiai neduoda patenkinamų rezultatų. Jau atėjo metas, kai reikia pradėti aptarti lietuvių kalbos kompiuterizavimo kokybę, užuot kalbėjus apie menką jos kompiuterizavimą.

Straipsnyje aprašomas aukštos kokybės sintaksiškai anotuotas lietuvių kalbos tekstynas, kuriame pateikta patikima informacija. Anotavimas atliekamas dviem lygmenimis: išsami informacija apie sakinį nurodoma lentelėje ir sintaksinė struktūra nubraižoma grafiškai. Anotuojant didelis dėmesys skiriamas sintaksinių ryšių vaizdavimui. Jie parodomi skirtingų spalvų ir tipų linijomis. Straipsnyje aprašomi neskaidomi žodžių junginiai – tai žodžių grupės, kurios tik kartu gali išplėsti kitą žodį, ir tik visą grupę gali pažymėti ją išplečiantis žodis. Kitų sakinio žodžių ryšys su vienu iš neskaidomo junginio dėmenų neturi prasmės. Struktūroje neskaidomi junginiai sudedami į vieną bloką parodant vidinius sintaksinius ryšius. Straipsnyje pateikiami sintaksiškai anotuotų sakinių pavyzdžiai. Sakinio struktūrai vaizduoti naudojamas grafas, nes medis, kuris sėkmingai taikomas anglų kalbos sakinių struktūrai, negali atspindėti visos sintaksinės informacijos, esančios lietuviškame sakinyje. Tai labai gerai matyti statistiniu metodu anotuotų sakinių pavyzdžiuose, kurie pateikiami šiame straipsnyje. Parodomos dviejų sakinių struktūros, kuriose trūksta labai svarbios informacijos: vienoje neparodytas sintaksinis ryšys tarp veiksnio ir tarinio, kitoje nėra ryšio tarp žodžių, kurie derinami skaičiumi ir gimine. Palyginimui straipsnyje pateikiami šie sakiniai, anotuoti ir VILSINTEKS naudojamu metodu.

Pradėto kurti tekstyno tikslas – sukaupti išsamią ir patikimą informaciją apie lietuvių kalbos gramatiką, atliekant sakinio analizę be klaidų, t. y. kai kompiuterio darbo rezultatus dar peržiūri žmogus. Anotuotas tekstynas bus viešai prieinamas internete. Kaip dabar iš VDU tekstyno galima gauti pateikto žodžio pavartojimo pavyzdžius, taip VILSINTEKS tinklapyje bus galima gauti sintaksinių struktūrų, kurios turi tam tikrų požymių – pavyzdžiui, vientisinių sakinių, kuriuose veiksniu eina įvardis ir pan., – pavyzdžių.

Anotuojant naudojamos *Excel* lentelės, nes jos leidžia vaizdžiai pateikti sakinio struktūrą ir lengvai pertvarkyti informaciją į XML formatą, kuris plačiai taikomas paieškai. Taigi anotuotą tekstyną bus galima panaudoti statistiniams lietuvių kalbos tyrimams, taip pat informacijai apie lietuvių kalbos gramatiką išgauti ir kt.

Įteikta 2013 m. spalio 28 d.

DAIVA ŠVEIKAUSKIENĖ
Lietuvių kalbos institutas
Petro Vileišio g. 5-216, LT-10308 Vilnius, Lietuva
daiva.fmf@gmail.com