

FORMAL SPECIFICATIONS FOR A DEPENDENCY GRAMMAR OF THE LITHUANIAN LANGUAGE

Gintarė Grigonytė, Erika Rimkutė
Vytautas Magnus University, Lithuania

Abstract

The first attempt to create some rules of Dependency Grammar (DG) for the Lithuanian language is introduced in this article. The need of a Lithuanian language parser was the background of this research. Concerning Lithuanian language processing, there are some key works on morphology level, but syntactical analysis is still lagging behind. That is the main reason why we consider formal specifications of DG being important for Lithuanian language processing. Up to now only verb phrases were syntactically analyzed while the grammar we propose involves a fuller coverage of the Lithuanian syntax. Our approach is based on corpus-based methods that let us extract, classify and evaluate DG rules. The main results of this research are discussed here as well.

Keywords: grammatical dependency, syntactic analysis, corpus, dependency rules, word order, insertion, word group

1. Introduction

Lithuanian language is a highly fleective language. There are some important works in automatic morphology area. The first attempts of automatic syntactic analysis are introduced in this article. The proposed specification for formal syntactic analysis is based on corpus-based rules. The main reason why we have chosen this method is a lack of available works and results concerning not only the formal grammar of the Lithuanian language but also computational linguistics in general.

Our linguistic resources were semi-automatically annotated Lithuanian corpus that consists of 1 million running words; the Corpus of the Contemporary Lithuanian Language (both corpuses were created at the Centre of Computational Linguistics, Vytautas Magnus University); the morphological analyzer *Lemuoklis*. The rules, presented bellow, include the most frequent met syntactic structures.

2. The specification of the formal dependency rules

Automatic analysis of the Lithuanian syntax is based on general principles of DG. According to I. Mel'čuk (Mel'čuk 1998: 13–15), DG is more relevant to the syntactic structures of a natural language than Phrase structure grammar. DG-based syntactic analyzer should recognize which word is dominant and which one is dependent.

We set four additional parameters for the descriptions of syntactic structures. They are as follows: dependency, word order, insertion, and priority. Dependency is the necessary attribute of links between words. Word order and insertion is possible but not mandatory. Priority is essential for the next stage of syntactic analysis (e.g. parser) and we will not pay much attention to this parameter here.

We mark **dependency** as an arrow which goes from the governing to the dependent word, e.g. $word1 \rightarrow word2$, which means $word1$ governs $word2$. Possible types of dependencies are represented in figure 1: a) one way forward dependency, $word1$ governs $word2$, e.g. $einu \rightarrow namo$ (*I am going* \rightarrow *home*); b) one way backward dependency, $word2$ governs $word1$, e.g. $mažas \leftarrow vaikas$ (*a little* \leftarrow *kid*); c) dual dependency, both words are on the same level, e.g. $ponas \leftrightarrow Jonas$ (*Mister* \leftrightarrow *John*).

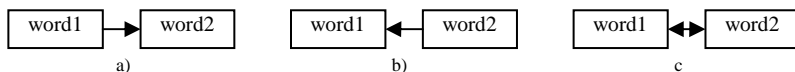


Figure 1. Possible types of dependencies between two words

Word order is the other important parameter concerned DG rules. Lithuanian language has free word order that is inconvenient for word dependency determinations. The most frequent sentence structure is SVO¹, e.g. $Jonas$ (S) $skaito$ (V) $knyga$ (O) (*John reads a book*).

The possible word order is presented in Figure 2. It is clear that dependency here is the same but the word order is different, e.g. $Jonas$ $skaito$ (*John reads*) vs. $skaito$ $Jonas$ (*literally reads John*).



Figure 2. Possible types of word order in a two-word combination

The other important criterion for syntactic description is **insertion**. An additional word can be inserted in already established structures. Usually new inserted words do not change syntactic dependency but modify its structure. It is necessary to evaluate the parameter of the possible insertion. The main tendency is for nominal words to be inserted with other nominal and adverbial word combinations while finite verb forms are never inserted in them. Verb phrases are characterized by a great variety of forms, various possible word order and insertions.

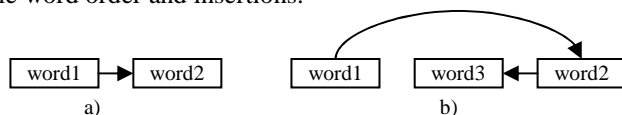


Figure 3. Example of insertion between two words

The insertion of an additional word between two words is represented in Figure 3. Insertion does not affect dependency, e.g. a) $skaito \rightarrow vaikas$ (*a kid* \rightarrow *reads*); b) $skaito \rightarrow [mažas \leftarrow vaikas]$ (*literally reads* \rightarrow [*little* \leftarrow *kid*]).

Up to now we have discussed three main relational parameters: dependency, word order and insertion. We have excluded **priority** as it is considered being the next step of

¹ S – subject, V – verb, O – object

analysis. Priority reveals the importance of different rules of the same group, e.g. adjective ← noun (*mažas ← vaikas (little ← kid)*), noun → adjective (*vaikas → mažas (literally a kid → little)*). The above mentioned rules indicate the same dependency, but the first one is more common therefore of a greater priority.

3. The dependency rules of the Lithuanian language

Grammar rules proposed here consist of two levels: the level of word groups (lower level) and the level of the combinations of word groups (upper level). We will discuss only word group level.

3.1. The dependency rules for word groups

3.1.1. Subject groups

Subject of the Lithuanian language can be simple or complex. A simple subject usually consists of one noun, e.g. *vaikas skaito (a kid reads; vaikas (a kid) is a subject)*. An extended simple subject consists of adjectives, participles, pronouns, numerals that are compatible with the noun in gender, number and case, e.g. *mažas vaikas (a little kid), pirmą knygą (the first book)* (see table 9). A complex subject consists of a few nouns in one group, e.g. *berniukai ir mergaitės (boys and girls)* (see table 1).

The same rules might be joined into a new complex group of rules, e.g. noun ↔ noun, [adjective ← noun] ↔ [adjective ← noun]. Various words or word groups can be inserted in a two word combinations, e.g. [adverb ← adjective] ← noun. Word order is important in that case also: adjectives, participles, numerals and pronouns usually precede nouns. A different word order is supposed to be inverted.

Table 1. The most frequent rules for a subject

Example	Word1	Word2	Word3	Dependency
<i>Jonas Jonaitis</i>	proper noun (G, N, C) ²	proper noun (G, N, C)	-	proper noun ↔ proper noun* ³
<i>ponas Jonas</i>	noun (G, N, C)	proper noun (G, N, C)	-	noun ↔ proper noun*
<i>gydytojas chirurgas</i>	noun (G, N, C)	noun (G, N, C)	-	noun ↔ noun*
<i>P. Jonaitis</i>	abbreviation	noun/proper noun	-	abbreviation ↔ proper noun*
<i>Berniukai ir mergaitės</i>	noun/proper noun	conjunction	noun/proper noun	noun/proper noun → conjunction → noun/proper noun

3.1.2. Predicate groups

Predicate can be simple and complex as well as subject. Finite verbs and participles compose simple predicates, e.g. *lyja (it's raining)*. The most common structures of complex predicates are represented in table 2.

Table 2. The most frequent predicates rules

Example	Word1	Word2	Word3	Word4	Dependency
<i>turėtų būti geras</i>	finite verb	infinitive	participle/ adjective	-	finite verb → infinitive → participle/adjective*

² G – gender, N – number, C – case, P – person

³ „*“ means additional parameters (word order, insertion, etc.) for dependency rules exist.

<i>turiu eiti</i>	finite verb	infinitive	-	-	finite verb → infinitive*
<i>buvo priverstas dirbti</i>	finite verb (<i>būti</i>)	participle	infinitive	-	finite verb → participle → infinitive*
<i>turi būti verčiamas dirbti</i>	finite verb	infinitive (<i>būti</i>)	participle	infinitive	finite verb → infinitive → participle → infinitive*
<i>galima dirbti</i>	participle/a djective	infinitive	-	-	participle/adjective → infinitive*
<i>buvo einantis</i>	finite verb (<i>būti</i>)	participle	-	-	finite verb → participle*
<i>galima būtų pasakyti</i>	participle	subjunctive verb (<i>būti</i>)	infinitive	-	subjunctive verb → participle → infinitive*
<i>būtų buvę galima padaryti</i>	subjunctive verb (<i>būti</i>)	participle (<i>būti</i>)	participle	infinitive	subjunctive → participle (<i>buvę</i>) → participle → infinitive*
<i>norėtųsi pailsėti</i>	subjunctive verb	infinitive	-	-	subjunctive verb → infinitive*
<i>būtų buvęs daromas</i>	subjunctive verb	participle	participle	-	subjunctive verb → participle (<i>buvęs</i>) → participle*
<i>noriu eiti miegoti</i>	finite verb	infinitive	infinitive	-	finite verb → infinitive → infinitive*
<i>norint padaryti</i>	gerund	infinitive	-	-	gerund → infinitive*
<i>bėgte nubėgo</i>	second infinitive	finite verb	-	-	second infinitive ← finite verb*
<i>bėgte nubėgti</i>	second infinitive	infinitive	-	-	second infinitive ← infinitive*
<i>gyventi yra gera</i>	infinitive	finite verb (<i>būti</i>)	adjective/ participle	-	finite verb ← [adjective/participle → infinitive]*
<i>skaito rašo</i>	finite verb	finite verb	-	-	finite verb ↔ finite verb*
<i>skaito ir rašo</i>	finite verb	conjunction	finite verb	-	finite verb → conjunction → finite verb*
<i>atrodo sveikas</i>	finite verb/infinitive (N)	adjective/parti ciple/noun (N, G)	-	-	finite verb/ infinitive → adjective/participle*

3.1.3. Attribute groups

Adjectives, participles, some pronouns and numerals compose attribute groups. Usually they are coordinated with noun in gender, number and case. Attributes that are not in concord with nouns are common in Lithuanian language. They are represented by Genitive of the noun, e.g. *tėvo knyga* (*father's book*). The main rules of attribute groups are given in Table 3.

Table 3. The most frequent attributive rules

Example	Word1	Word2	Dependency
<i>gerai žinomas</i>	adverb	adjective/participle	adverb ← adjective/ participle*
<i>turtingas pinigų</i>	adjective	noun Gen ⁴	adjective → noun Gen*
<i>gabus muzikai</i>	adjective	noun Dat	adjective → noun Dat*

⁴ Gen – Genitive, Acc – Accusative, Instr – Instrumental, Loc – Locative

<i>garsus pasiekimais</i>	adjective	noun Instr	adjective → noun Instr*
---------------------------	-----------	------------	-------------------------

3.1.4. Object groups

Object groups are mainly expressed by nouns in Genitive, Dative, Accusative and Instrumental cases, e.g. *skaito knygą* (read/reads a book), *didžiujasi sūnumi* (is/are proud of a son). There are some variants when an object group consists of a preposition and a noun in Genitive, Dative, Accusative and Instrumental cases. These variants have a fixed word order and the governing word is usually a preposition. Attributes of nouns can be inserted in object groups (see Table 4).

Table 4. The most frequent object groups

Example	Word1	Word2	Dependency
<i>ant dėžės</i>	preposition	noun/adjective/pronoun/ numeral/participle Gen	preposition → noun/adjective/ pronoun/ numeral/participle Gen*
<i>apie vaikų</i>	preposition	noun/adjective/pronoun/ numeral/participle Acc	preposition → noun/adjective/ pronoun/ numeral/participle Acc*
<i>su draugais</i>	preposition	noun/adjective/pronoun/ numeral/participle Instr	preposition → noun/adjective/pronoun/ numeral/participle Instr*

3.1.5. Modifier groups

Adverbs, half-participles, nouns in Instrumental, Locative cases compose modifier groups, e.g. *dirbti miške* (to work in a forest), *greitai bėgti* (to run fast). Prepositional constructions with nouns can also form a modifier group. The most frequent rules for modifier rules and their internal relations are shown in Table 5.

Table 5. The most frequent modifier groups

Example	Word1	Word2	Dependency
<i>be galo daug</i>	adverb	adverb	adverb ← adverb
<i>per greitai</i>	particle	adverb	particle → adverb
<i>prie miško</i>	preposition	noun Gen, Acc, Instr	preposition → noun Gen, Acc, Instr*

Rules that describe main parts of sentence belong to the upper (sentence) level. Joining rules of both levels enables us fully to describe syntactic relations of a simple sentence. Upper level rules can be classified into predicate and subject relation, predicate and object relation, predicate and modifier relation, subject and attribute relation according to their sentential functions. Due to the lack of space a more detail description of these rules are omitted.

Some of the lower level rules are applied in analyzing complex verb groups (Grigonytė 2004; Grigonytė et al. 2005). The future research will go in the direction of syntactic analyzer. Another application for the automatic syntactic analysis is morphological disambiguation of Lithuanian language (Rimkutė 2003).

4. The methodology of extraction of corpus-based rules

The methodology of extraction of corpus-based rules is described here. Automated analysis and classification of word groups was performed with the help of the Corpus of the Contemporary Lithuanian Language and morphological analyzer *Lemuoklis* (Zinkevičius 2000). Automatic analysis of word groups consists of the following stages: detection of text units, lemmatization of isolated words, classification of text units into relevant word groups.

The output of the lemmatization is shown below:

<w l="eiti(eina,ėjo)" m="finite verb, infinitive" l="eiti(eina,ėjo)" m="finite verb, participle">**eiti**</w>

After the lemmatization we have a morphological output that is used as an attribute for further classification. The main criteria of the classification are the length of a word group and possible word relations, i.e. subject, predicate, object, attribute and modifier relations in a group. After the revision of the classified text units by an expert the final structures were defined.

5. Conclusions

Our attempts to create the rules for formal syntactic analysis concentrates on the proposed methodology for the extraction of syntactic rules for formal grammar of the Lithuanian language. Subject, predicative, object, attribute and modifier groups were described and analyzed within two levels of the analysis. We also defined the necessary parameters and additional features that would increase the quality of automatic syntactic analysis. Our future plans include DG application for the parser of Lithuanian language.

References

- Grigonytė, Gintarė 2004. Dalinis sintaksinis lietuvių kalbos veiksmažodžių analizatorius (Partial Syntactic Analyzer of the Lithuanian Language). *Bachelor thesis*. Vytautas Magnus University, Kaunas.
- Grigonytė, Gintarė, Rimkutė, Erika 2005. Automatinis lietuvių kalbos veiksmažodžių grupių atpažinimas (Automatic Verb Phrases Recognition in the Lithuanian Language). In: *Informacinės technologijos 2005*, Kaunas: Kauno Technologijos universitetas. 315–320.
- Mel'čuk, Igor 1988, *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.
- Rimkutė, Erika 2003. Morfologinio daugiareikšmiškumo tipologija (the Typology of Morphological Ambiguity). In: Merkys, V.; Ambrasas, V.; Sauka, L. (eds.) *Lituanistica* 4 (56). 60–78.
- Zinkevičius, Vytautas 2000. *Lemuoklis* – morfologinei analizei (Morphological analysis with *Lemuoklis*). In: Gudaitis, L. (eds.) *Darbai ir Dienos* 24. 246–273.

GINTARĖ GRIGONYTĖ is an engineer-programmer of the Centre of Computational Linguistics at Vytautas Magnus University. She is a student of Software engineering programme at Kaunas University of Technology. Her Master thesis deals with a dependency analysis for Lithuanian language. Her research interests include computational linguistics, software engineering, automatic syntactic analysis. E-mail: g.grigonyte@hmf.vdu.lt

ERIKA RIMKUTĖ is a junior researcher of the Centre of Computational Linguistics at Vytautas Magnus University. She received her M. A. (Lithuanian language) at Vytautas Magnus University. Her research interests include corpus linguistics, computational linguistics, automatic morphological analysis and synthesis, morphological ambiguity and disambiguation and automatic syntactic analysis. Her doctoral study focuses on morphological ambiguity and disambiguation in the Lithuanian language. E-mail: e.rimkute@hmf.vdu.lt.