Automatic Lemmatisation of Lithuanian MWEs

Loïc Boizou **Jolanta Kovalevskaitė** **Erika Rimkutė**
Centre of Computational Linguistics
Vytautas Magnus University
Kaunas, Lithuania
{l.boizou, j.kovalevskaite, e.rimkute}@hmf.vdu.lt

Abstract

This article presents a study of lemmatisation of flexible multiword expressions in Lithuanian. An approach based on syntactic analysis designed for multiword term lemmatisation was adapted for a broader range of MWEs taken from the Dictionary of Lithuanian Nominal Phrases. In the present analysis, the main lemmatisation errors are identified and some improvements are proposed. It shows that automatic lemmatisation can be improved by taking into account the whole set of grammatical forms for each MWE. It would allow selecting the optimal grammatical form for lemmatisation and identifying some grammatical restrictions.

1 Introduction

In Lithuanian, in addition to fully fixed multiword expressions (MWEs), there are many MWEs with one or more constituents (possibly all) which can be inflected. Therefore, these “flexible” MWEs appear in texts in several forms[1]: as shown by the corpus data, the Lithuanian verbal phraseme *pakišti koją*, meaning ‘to put a spoke in somebody’s wheel’, has a form with the verb in definite past *pakišo koją*, a form with the verb in past frequentative (*pakišdavo koją*), and future (*pakiš koją*) (for more examples, see Kovalevskaitė (2014)). If we extract MWEs of a strongly-inflected language like Lithuanian from a raw text corpus using statistical association measures, we often get MWEs with their different grammatical forms (GF). However, in lexical databases and terminology banks, a single lemma is usually used in order to represent the MWE independently from its concrete forms which appear in the corpus.

In traditional Lithuanian dictionaries of idioms, there is no problem of MWE lemmatisation, because data are collected manually and represented following the rule that a verb of an idiom is provided in infinitive form (Paulauskas (ed), 2001), e.g.: *Savo vietą žinoti* ‘to know one‘s place’, *Vietos neturėti* ‘to have nowhere to go’. The recent Dictionary of Lithuanian Nominal Phrases[2], which was compiled semi-automatically from a corpus, contains the unlemmatised list of MWEs. Therefore, automatic phrase lemmatisation could help in organizing the dictionary data and in improving the user interface.

This article describes the main problems occurring during the lemmatisation of Lithuanian MWEs. The concept of lemmatisation is quite clear for single words, but for MWEs, it can be understood differently as it will be discussed in part 3. Although the accuracy of lemmatisation of individual words is high (99% for lemmatisation and 94% for grammatical form identification (Rimkutė and Daudaravičius, 2007)), the lemmatisation of single words included in MWEs cannot produce well-formed MWE lemmas. Indeed, base forms of Lithuanian multiword terms which should occur in dictionaries and terminology databases are not the same as the sequences of lemmas of their constitutive words (Boizou et al., 2012, p. 28). For example, if we lemmatise each constitutive word in MWEs *taupomųjų bankų*, *taupomuoju banku* (‘savings bank’ in genitive plural and instrumental singular), the result is *taupyti bankas* (infinitive ‘to save’ and nominative singular ‘bank’), because the morphological annotator of the Lithuanian language (Zinkevičius, 2000; Zinkevičius et al., 2005) assignes infinitive as the proper lemma for participles and other verb forms. The structure of such improperly lemmatised MWE fails to reflect the syntactic relations (agreement and gov-

[1] Grammatical forms of the same MWE (or phraseme) can be labelled as phraseme-types (Kovalevskaitė, 2014).

[2] http://donelaitis.vdu.lt/lkk/pdf/dikt_fr.pdf

ernment) which ensure the grammatical cohesion between the constitutive words of an MWE.

This work is focused on syntagmatic lemma, which is the form of the MWE where the MWE syntactic head is lemmatised and the necessary adaptations are made in order to ensure the morphosyntactic unity of the MWE. With a similar approach, a tool for automatic Lithuanian sintagmatic lemmatisation called JungLe was first developed and trained with multiword terms during the project ŠIMTAI 2 (semi-automatic extraction of education and science terms). The first experiments with the Lithuanian multiword terms showed accuracy close to 95% (Boizou et al., 2012), but it can be related to a relatively low variety of term structures. For this study, JungLe was adapted for a broader set of types of Lithuanian compositional and non-compositional MWEs, e.g. idioms, collocations, nominal compounds, MW terms, MW named entities, MW function words, proverbs, etc.

2 Data

This study is based on the data from the Dictionary of Lithuanian Nominal Phrases (further on, dictionary). The database of the dictionary[3] consists of 68,602 nominal phrases. It has to be mentioned, that the term nominal phrase refers to all MWEs which contain at least one noun: it can be phrases with a noun as a syntactic head, as well as phrases with a verb or an adjective as a syntactic head. In this article, the terms MWE and phrase are used interchangeably.

In the dictionary, the phrases are of different length: from two-word phrases (31,853 phrases) to phrases comprising 46 words (1 phrase). The major part of the dictionary phrases is made up of two-word phrases (46.4%), whereas three-word phrases and four-word phrases form accordingly 28.7% and 10.1% of the dictionary (Rimkutė et al., 2012, p. 19). The phrases are not lemmatised, but given in the form as they appear in the corpus, e.g.:

- *mobilaus ryšio telefonas*, *mobilaus ryšio telefono*, *mobilaus ryšio telefoną*, *mobilaus ryšio telefonus* ('mobile phone' in various grammatical forms: nominative singular, genitive singular, accusative singular and accusative plural);
- *nenuleisti rankų*, *nenuleido rankų*, *nenuleidžia rankų* ('not to give up', lit. 'not to lower hands', in various grammatical forms: infinitive, definite past, present tense).

As Lithuanian is a strongly inflected language, it is an advantage that users can see phrases in the form they are used in the corpus. Phrases were extracted by the method of Gravity Counts (Daudaravičius and Marcinkevičienė, 2004, p. 330) from the Corpus of Contemporary Lithuanian Language (100 m running words; made up of periodicals, fiction, non-fiction, and legal texts published in 1991-2002). Gravity Counts helps to evaluate the combinability of two words according to individual word frequencies, pair frequencies or the number of different words in a selected 3 word-span. As a result, it detects collocational chains as text fragments, not as a list of collocates for the previously selected node-words.

After automatic extraction of collocational chains from the corpus, manual procedures were performed: transformation of collocational chains into phrases (the procedure is described in detail in Marcinkevičienė (2010) and Rimkutė et al. (2012)). According to the lexicographical approach, linguistically well-formed collocational chains have to be grammatical, meaningful, and arbitrary. Therefore, some chains were shortened, complemented, joined or deleted manually. At present, the dictionary database contains phrases without additions (1) and with additions: additions can be explicitly specified (2) or not (3), e.g.:

1. *ne tuo adresu* 'under a wrong address';
2. (*gauti*; *suteikti*) *daugiau informacijos* '(get; give) more information';
3. *atkreipiant dėmesį į* (...) 'paying attention to (...)'.

Phrase type	Number of lemmas	Number of lemmas with 2 or more GF
2-word	18,581	6,585 (35,4%)
3-word	10,970	2,245 (20,5%)
4-word	3,333	477 (14,3%)
In total	32,884	9,307 (28,3%)

Table 1: Statistical information about MWEs from the type (1).

[3] http://tekstynas.vdu.lt/page.xhtml?id=dictionary-db

Only two-, three-, and four-word phrases from the type (1) were filtered out for this study (see Table 1). As already mentioned, these phrases make up 85% of the whole dictionary database, and thus can be considered as the most typical multiword units in Lithuanian (Marcinkevičienė, 2010; Rimkutė et al., 2012). Most of them are idioms, phraseologisms and collocations, although there are many multiword terms as well.

It was calculated that two-word expressions have on average 1.71 different grammatical forms, three-word expressions – around 1.35 different grammatical forms, and four-word expressions – 1.22 different grammatical forms. The maximum number of grammatical forms for one lemma are respectively 21 (*aukšta mokykla* 'high school'), 15 (*mobilaus ryšio telefonas* 'mobile phone') and 8 (*kandidatas į seimo narius* 'candidate as MP'). For this study, only the 9,307 MWEs with two or more grammatical forms were selected. The remaining MWEs, those for which only one form was identified automatically, are excluded, as they do not always need to be lemmatised and require a further study.

The next section describes the main approaches applied to the process of automatic lemmatisation of MWEs.

3 Approaches to Lemmatisation of MWEs

In Lithuanian, a great number of MWEs can appear in different grammatical forms. As such, they do not differ from variable simple words. Accordingly, a lot of Lithuanian MWEs consist of nouns, verbs and/or adjectives that are used in a particular grammatical form. Some of these word classes can have from a few to dozens of different grammatical forms. Traditionally, for the set of grammatical forms of each variable word, one basic form is assigned. The latter, a lemma, is a convenient representation of the whole set of grammatical word forms. Although in principle a lemma could be an artificial form (a stem, for example), the tradition is to select as a lemma one form from the whole set of grammatical forms, e.g. in Lithuanian:

- nominative singular form for nouns (except for plural nouns);
- nominative singular masculine positive indefinite form for adjectives;
- positive form for adverbs (if they vary in degree);
- infinitive for verbs (including participles).

In the field of computational linguistics, it is common to use artificial lemmas for MWEs, because they can be easily generated by automatic means. There are two main kinds of artificial lemmas:

a) It is possible to use a lemmatic sequence which is the sequence of lemmas of each constitutive word of the MWE[4]. Using the morphological annotation tool for Lithuanian (the tool is described in Zinkevičius (2000) and Zinkevičius et al. (2005)), each grammatical form of the multiword term, *bendrosioms mokslo programoms* 'framework programme', is annotated morphologically as follows:

- <word="bendrosioms" lemma="bendras" type="bdv., teig, nelygin. l., įvardž., mot. g., dgs., N."/>[5]
- <word="mokslo" lemma="mokslas" type= "dkt., vyr. g., vns., K."/>[6]
- <word="programoms" lemma="programa" type="dkt., mot. g., dgs., N."/>[7]

The lemmatic sequence, e.g. for the previous example *bendras mokslas programa*, is often used in the field of automatic term recognition (e.g., Loginova et al. (2012, p. 9)) to represent a term or another type of MWE. Nonetheless, such a substitute, which lacks grammatical cohesion between the parts of the MWE, appears as a heap of words, which is unnatural for human users[8].

[4] The difference between syntagmatic lemma (with morphosyntactic relations between constitutive words) and lemmatic sequence (the sequence of lemmas of constitutive words) is relevant only for MWEs, not for single words.

[5] The field *type* contains the following grammatical features: adjective, positive, undefined, positive degree, feminine, plural, dative.

[6] Grammatical features: noun, masculine, singular, genitive.

[7] Grammatical features: noun, feminine, plural, dative.

[8] In about 5% of the studied phrases, the sequences of isolated lemmas incidentally correspond to their natural lemma, e.g. *vyras ir moteris* 'man and woman', *valstybinis simfoninis orkestras* 'national symphony orchestra'. Such cases require the following conditions: the nominal syntactic head is masculine singular, the only syntactic relation inside the term is agreement or implies invariable words, degree and definiteness must not be retained in the lemma, no participle is implied.

b) The second frequent method is stemming, that is, dropping of endings. For example, the forms of the previously mentioned MWE *bendroji mokslo programa*, *bendrosios mokslo programos* and *bendrosioms mokslo programoms* can be represented as: *bendr moksl program*. This option is even more artificial for Lithuanian, since, in addition to the loss of syntactic cohesion, this approach generates shortened words without endings, which do not exist in Lithuanian.

Other approaches attempt to provide a natural lemma, i.e., by either choosing the most frequent form as a lemma, or generating a correct syntagmatic lemma from grammatical forms. Taking the most frequent form of the lemma avoids mistakes in generation, but the result is that the set of basic forms is heterogeneous: some MWEs will be in nominative, some will be in accusative, genitive or in some other case, some will be in the plural form, others - in the singular.

Automatic lemmatisation according to the syntactic structure of each MWE ensures the constitency of basic forms, but it is the most complicated process. The tool JungLe, which is described in Boizou et al. (2012), was specifically designed for this task. This software analyses an MWE and attempts to distinguish three types of syntactic components (as a concrete example the multiword term *individualus studijų grafikas* 'individual study plan/schedule' is provided):

- syntactic head (e.g., the noun *grafikas* 'plan/schedule');
- words congruent with the head (e.g., the adjective *individualus* 'individual');
- other words, that is, words governed by the head and their own dependents (e.g., the noun in genitive case *studijų* 'study').

The generation of the syntagmatic lemma requires the syntactic head to be lemmatised (for terms, the syntactic head is a noun, but there is more diversity with other types of MWEs). Words (usually in the genitive case) governed by the head and their own dependents remain in their grammatical form, e.g., *švietimo {lygmuo}* 'education level', *socialinių mokslų {sritis}* 'field of social sciences'.[9]

The most difficult case concerns words congruent with the head, since they often have to be corrected to remain congruent with the head once it is lemmatised. If the head is masculine singular, the adaptation usually requires only taking lemmas for each congruent word, e.g., in the multiword term *individualus studijų grafikas* 'individual study plan/schedule' (the adjective *individualus* 'individual' agrees with the noun *grafikas* 'schedule', not with the noun *studijų* 'study'). When the syntactic head is feminine, congruent words must also be put in their feminine form, e.g., *nuotolinės studijos* 'distance studies' (instead of **nuotolinis studijos*, where the masculine singular indefinite positive form of the adjective *nuotolinis* is incongruent with the feminine plural head *studijos*).

Besides, some lexico-grammatical features, e.g., definiteness, comparative/superlative degrees, are usually semantically relevant, so that they have to be kept in the syntagmatic lemma, which requires to generate the proper form, even when the head is masculine and singular, e.g. *Senasis ir Naujasis testamentas* 'The Old and New Testament' (where the adjectives *Senasis* and *Naujasis* are in the definite form, instead of **Senas ir Naujas testamentas*), *aukštesnioji žemės ūkio mokykla* 'high school of agriculture' (where the adjective *aukštesnioji* is in the definite comparative form, instead of **aukštas žemės ūkio mokykla*).

Syntagmatic lemmatisation also requires to lemmatise participles in a different manner than single words. Indeed, participles are traditionally lemmatised as verbs in infinitive form. For example, the single word lemmatisation of the term *perkeliamieji gebėjimai* 'transferable skills' gives a result *perkelti gebėjimai*, that is a sequence of an infinitive (*perkelti* 'to transfer') and a noun in nominative (*gebėjimai* 'skills'). The correct syntagmatic lemmatisation requires participles to be corrected in gender, number and case only, in order to remain congruent with their lemmatised head, e.g. *perkeliamasis gebėjimas*.

All required generations are made by a light-weighted generative module. This module uses to the largest possible extent the information provided by the morphological analyser, which works on a single word basis. Its generative capacities are restricted to the nominative forms, since noun, adjective and participle lemmas are in the nomina-

[9] Here and below the syntactic head is indicated in curly brackets.

tive form. Aiming at facilitation of the process, the generation proceeds either from a single lemma or a grammatical form. For example, lemmas for participles are generated from the grammatical form, because it helps to avoid the problem of numerous verbal paradigms in Lithuanian, while adjectives are generated from the lemma. Indeed, some endings of nouns and adjectives (e.g., *-(i)ų* genitive plural) hide the declension paradigm (which is necessary for the selection of the correct feminine ending), so that it is better to decide from the lemma, which expresses the adjectival declension paradigm by its ending[10], e.g., *nuotolinis* ‘distant‘ (third adjectival declension paradigm) → *nuotolinė*. The whole process is very similar to Thurmair (2012, p. 257).

4 Syntagmatic Lemmatisation of Lithuanian MWEs: Evaluation and Results

In this part of the article, we present our results: what problems are solved by syntactic analysis, and what problems still remain and pose challenges for automatic MWE lemmatisation.

Two-, three- and four-word phrases were automatically lemmatised with the help of JungLe tool and the results were evaluated manually (see Table 2). JungLe generates a lemma for each MWE grammatical form separately, so that more than one lemma can be provided for the same MWE, especially when it is difficult to identify automatically to which word an attribute in genitive belongs, e.g., two lemmas, both inaccurate, were provided for the MWE *bendroji dalinės nuosavybės teisė* ‘general partial ownership’, where the first adjective *bendroji* ‘general’ is congruent with the MWE head *teisė* ‘law’ and the second adjective *dalinės* ‘partial’ with the noun *nuosavybės* ‘property’ (which depends on the MWE head). In the first provided lemma **bendroji dalinė nuosavybės teisė*, *dalinė* incorrectly agrees with *teisė*, and in the second one, **bendrosios dalinės nuosavybės teisė*, *bendrosios* incorrectly agrees with *nuosavybės*.

As each grammatical form is lemmatised separately, in some cases there is more than one lemma for the same MWE. Thus, lemmatisation accuracy was assessed for individual grammatical forms of MWEs. Table 2 shows that the highest accuracy is with two-word phrases; however, the number of incorrectly lemmatised MWEs increases for three- and four-word phrases. It shows that the syntactic complexity increases with the length of the MWEs.

Phrase type	Number of lemmas	Number of GF	Correctly lemmatised GF
2-word	6,585	19,822	91.56%
3-word	2,245	6,110	80.57%
4-word	477	1,206	76.43%
In total	9,307	27,138	

Table 2: Statistical information about the analysed MWEs (2 or more forms only).

The analysis of the automatic lemmatisation revealed three groups of errors[11]: a) agreement errors (number, gender); b) government errors; c) lexico-grammatical errors (degree, definiteness, lexical plural).

4.1 Agreement Errors

Many errors occur with numerals, e.g., **beveik du trečdalis* ‘*nearly two third’ (it should be *beveik du trečdaliai*, ‘nearly two thirds’, with *trečdalis* ‘third’ in the plural form), also **aštuoni mėnuo* ‘*eight month’ (while it should be *aštuoni mėnesiai*, ‘eight months’, with *mėnuo* ‘month’ in the plural form). Many of these errors can be eliminated by applying proper rules in the syntactic analysis.

During the syntactic analysis gender errors occur when the composition of an MWE is more complex, e.g., **vienas ar kita grupė* ‘one or the other group’(instead of *viena ar kita grupė*, with *viena* ‘one’ and *kita* ‘other’ in the feminine form). We can see that the coordinating link could be the factor determining the agreement errors[12].

[10] Ending *-as* for the first adjectival declension, *-ias* for the second, *-is* and *-ys* for the third and fourth, and *-us* for the fifth.

[11] Some errors of lemmatisation occur due to errors of the previous morphological analysis, e.g., the lemma for the MWE *arbatinio šaukštelio* (‘tea spoon’, sing. Gen.) is provided incorrectly as **arbatinio šaukštelis* (genitive singular + nominative singular, instead of *arbatinis šaukštelis*, nominative singular for both the adjective and the noun), because of an improper morphological analysis: *arbatinis* was annotated as a noun, not an adjective.

[12] Some similar errors, which must be corrected, appear with the genitive case, e.g. *Afrikos ir Azija* (genitive and nominative, it should be *Afrika ir Azija* ‘Africa and Asia’, nominative and nominative), **daina ir šokių ansamblis* (nominative noun, conjunction, genitive noun, nominative noun), instead of *dainų ir šokių {ansamblis}* (genitive noun, conjunction, genitive noun, nominative noun) ‘song and dance ensemble’.

It was observed that a large number of agreement errors take place when one of the attributes is an apposition, i.e., a noun which has to agree in a case (sometimes gender and number) with the adjacent noun, e.g., **šalies gavėja* (it should be *šalis gavėja*, 'the recipient party'), **mergelės Marija* (it should be *mergelė Marija*, 'the Virgin Mary', with both parts in nominative). One the other hand, such attributes are not numerous, and such errors could be solved by looking at other cases than genitive.

4.2 Government Errors

Problems mainly arise when the tool fails to correctly identify the syntactic head of a phrase. Such a problem usually occurs when the head is not at the end of an MWE, e.g. **paskolos {studijos}* (instead of *{paskola} studijoms* 'study loan'), **atliko savo {darbas}* 'carried out their work' (instead of *{atlikti} savo darbą*, where the verb is the head). Problems also occur in phrases, where the head is a half-participle or a gerund: **įsigaliojus naujasis civilinis {kodeksas}* 'when the new civil code came into effect' (it should be *{įsigaliojus} naujajam civiliniam kodeksui*, i.e. where dative is required).

It must be noticed that in some cases the representation of the lemma does not correspond to a natural linguistic form. It occurs in collocations which contain a conjugated verb (*pakilo*) with a (nominative) subject, e.g. *pakilo temperatūra* 'temperature rised'. In the assigned lemmas, conjugated verbs are substituted for infinitive forms (*pakilti*). Infinitives cannot have a subject in Lithuanian, and therefore the MWE subject could be presented in brackets in the nominative form (e.g. *pakilti (temperatūra)* 'to rise (temperature)'). A further exception comes from the MWEs with a gerund, since the logical subject of a gerund is not expressed as a nominative, but as a dative complement, e.g., *atsitikus nelaimei* 'a disaster occurs'. In such cases, we propose to assign two different lemmas: one lemma, which retains the grammatical form without change, *atsitikus nelaimei* (gerund + dative complement), as used in gerund grammatical form; and the second lemma, where the gerund is substituted for the infinitive and the dative complement is substituted for a nominative form in bracket, e.g. *atsitikti (nelaimė)* (infinitive + nominative), as in the previous example.

We should also mention, among other complicated lemmatisation instances, the loss of grammatical forms which carry the meaning of an MWE, e.g., *atstovų teigimas* ('representatives' assertion') could be considered as a correctly generated lemma; however, after a closer investigation of the grammatical forms, we can see that in this MWE the syntactic head is always used in the instrumental case (*teigimu*), i.e., *atstovų teigimu* ('according to the representatives'), thus the lemma should keep this form. Similarly, the lemma of an MWE *balsavimo paštas* ('voting post') is not accurate, as the syntactic head (*paštas*) should be in the instrumental case, i.e., *balsavimas paštu* ('voting by mail'), while the lemma of a phrase *visa išgalė* ('all possible measures') should be *visomis išgalėmis* ('by all possible measures'), because this phrase as an MWE is used only in the form of instrumental plural.

4.3 Lexico-grammatical Errors

There are many errors made by JungLe where a lemma has to be assigned to nouns which are used in plural in the phrase, e.g., **žmogaus teisė ir laisvė* 'human right and freedom', instead of *žmogaus teisės ir laisvės*, 'human rights and freedoms'; **jungtinė tauta* 'united nation', instead of *Jungtinės Tautos*, 'United Nations'; **visa Baltijos šalis* 'the whole Baltic country', instead of *visos Baltijos šalys* 'all Baltic countries', **Vilniaus ir Šalčininkų rajonas* 'Vilnius and Šalčininkai district', instead of *Vilniaus ir Šalčininkų rajonai* 'Vilnius and Šalčininkai districts'. As number errors were considered the examples when a lemma looked correct at first sight, i.e., a lemma is provided in the same number as in the dictionary. However, from the usage data (all forms of a phrase) one can see that certain MWEs are used only in plural, e.g., *meteorologinės sąlygos* 'meteorological conditions', *mineralinės trąšos* 'mineral fertilizers', *mirties aplinkybės* 'death circumstances'. All these phrases, which are made of an adjective or a genitive noun followed by a noun, are incorrectly lemmatised in the singular form, e.g. **meteorologinė sąlyga*, **mineralinė trąša*, **mirties aplinkybė*. Many of the above-mentioned nouns can be used in plural and singular, when they are used independently, but they can be restricted to one of these numbers inside MWEs. Traditional grammars and dictionaries do not provide necessary information to solve this problem, which could often be resolved if we take into account actual usage from the corpus.

There are two types of degree errors: a) in some phrases a particular degree form is used, thus, the same form should be in the lemma (*Aukščiausiasis Teismas* ‘supreme court’; *daugiau kaip dveji metai* ‘more than two years’); b) there are phrases, where an adverb or an adjective is used in several degrees: then different phrases can contain adjectives or adverbs of different degrees (cf. *įvairūs / įvairiausi būdai* (‘various/ the most various ways’) and *skirti daug/daugiau/daugiausia dėmesio* (to pay a lot of/more/ most attention)). Analysis of all forms of an MWE can help to distinguish a) and b) phrases.

Errors of definiteness often occur in phrases joined by coordination, when one adjective is provided in the definite form while the other one is indefinite, e.g., **Senas ir Naujasis testamentas* (‘Old and New Testament’).

After the examination of errors and problematic cases created by JungLe, we can draw a conclusion that automatic lemmatisation is aggravated by:

1. syntactic heads in the genitive form: when there are several nouns in the genitive in the MWE, it leads to attachment ambiguities;

2. the length of an MWE: the longer the phrase, the more complicated syntactic structure; the accuracy of lemmatisation decreases (see Table 2);

3. problems of lexico-grammatical nature, when a grammatical form depends on a lexical meaning (here, errors of number must be emphasized).

It must be emphasized that the numbers in Table 2 show the situation after the first extension of JungLe. The results can still be improved significantly. Some errors can be corrected by improving the grammar used by JungLe for syntactic analysis, some of them require adding new capacities to JungLe, other errors will be difficult to correct without human intervention.

5 Discussion and Recommendations

The traditional morphological analyser, which analyses every word individually, cannot produce natural lemmas for MWEs. It is necessary to carry out the syntactic analysis for automatic assignation of natural lemmas for different grammatical forms of MWEs. But beside syntactic analysis of MWEs, we often need to take into account the usage data of a particular MWE and to apply additional criteria. The automatic syntagmatic lemmatisation tool was tested on the data from the Dictionary of Lithuanian Nominal Phrases, which are characterized by a high variety. For this reason, it can be stated that the essential features, as well as problems, of automatic lemmatisation of Lithuanian MWEs were identified.

One of the most important lemmatisation issues that is difficult to solve is the problem of an attribute which is incongruent with a noun and usually expressed in genitive. Most commonly, such problems (in automatic lemmatisation) are inevitable, because ambivalent syntactic relations can exist in MWEs composed of the same words, e.g., the lemma for MWE grammatical forms *administracinės teisės pažeidimų*, *administracinės teisės pažeidimą*, *administracinės teisės pažeidimus* ‘breach of administrative law’ (where *administracinis* ‘administrative’ is congruent with *teisė* ‘law’) should be *administracinės teisės pažeidimas*, while the lemma for MWE grammatical forms *administracinį teisės pažeidimą*, *administracinių teisės pažeidimų*, *administracinius teisės pažeidimus* ‘administrative breach of law’ (where *administracinis* ‘administrative’ is congruent with *pažeidimas* ‘breach’) should be *administracinis teisės pažeidimas*. In order to set the right lemma, the noun with which the adjective agrees must be correctly assigned.

The head of a phrase in genitive can influence adjective agreement errors, too. For instance, the genitive grammatical form *periodinio mokslo leidinio*, where it is unclear if *periodinio* ‘periodic’ is congruent with *mokslo* ‘science’ or *leidinio* ‘publication’, could formally be lemmatised as **periodinio mokslo leidinys* ‘a publication of periodic science’ or *periodinis mokslo leidinys* 'a periodic scientific publication’ by looking at the internal syntactic structure of the term. In order to disambiguate syntax correctly, we need to compare other (unambiguous, i.e., cases other than the genitive) forms of the term, e.g., *periodiniams mokslo leidiniams* (in dative plural), which shows that the adjective *periodinis* ‘periodic’ is congruent with the noun *leidinys* ‘publication’, therefore, this MWE should properly be lemmatised as *periodinis mokslo leidinys*.

The problems concerning the genitive case would decrease, if the usage criterion was applied, i.e., if lemma was identified considering all forms of the MWE. For example, it is especially complicated to lemmatise an MWE with all genitive cases, e.g., it is impossible to identify an accurate lemma for MWEs *valstybinio socialinio draudimo biudžeto* ('the budget of state social insurance'), *fizinių asmenų pajamų mokesčių* ('income taxes of natural persons'). In such cases, a rule should be applied: if the same phrase is used in genitive and in other cases, the lemma should be identified on the basis of phrases with other cases than genitive.

The usage criterion would help to avoid the number errors. Quite often this criterion proved the rule that if different grammatical forms of an MWE are in plural, then the lemma should keep the plural form too. For example, a dictionary of nominal phrases provides two grammatical forms: *laužas ir atliekos*, *laužo ir atliekų* 'debris and waste', in both phrases the noun *laužas* is in the singular form, while *atliekos* is used in plural. Thus, when merging the two MWEs to one lemma, *atliekos* has to remain in the plural form. During the lemmatisation of the forms *žvėris ir paukščius*, *žvėrių ir paukščių*, *žvėrys ir paukščiai* ('beasts and birds', repectively accusative plural, genitive plural, nominative plural), we have to assign plural lemmas for both nouns – *žvėrys ir paukščiai*, because all forms of these nominal phrases are in the plural form. This is especially important for names, cf. **Lietuvos geležinkelis* (it should be *Lietuvos geležinkeliai*, 'Lithuanian Railways'), **Vilniaus šilumos tinklas* (it should be *Vilniaus šilumos tinklai*, 'Vilnius Heating Network').

Based on the usage data, it would be possible to distinguish between the MWEs where a certain word is used only in one form of the degree (*Aukščiausiasis Teismas*, superlative, 'Supreme Court'), and those where several forms of a degree are used (*įvairūs būdai* and *įvairiausi būdai*, positive and superlative, 'various ways').

When applying the usage criterion, it is important to remember that in this case the accuracy of the tool will be linked to the corpus data: the rarer the phrase, the higher the risk for the tool to make a mistake. For example, if we recognize only two forms of a particular phrase, and they are both in the plural form, the tool can come to a false conclusion that the lemma of that phrase is also in plural, although that phrase could also be used in singular. But such a risk is significant for rare MWEs only.

It is possible that next to the usage criterion, other criteria will have to be introduced. For example, in order to avoid lemmatisation errors related to definiteness, it would be worthwhile to invoke not only the usage, but, also, frequency criterion. Indeed, according to the data, *nekilnojamas turtas* (with the indefinite form of the adjective *nekilnojamas*) and *nekilnojamasis turtas* (with the definite form of the adjective *nekilnojamasis*), which both mean 'real property', are concurrently used. However, one can expect the standard form, the definite one, to be more frequent, as it is a term.

The evaluation of the research results has revealed that the accuracy of the MWE lemmatisation is not only influenced by the accuracy of the syntactic analyser, but, also, by the variability of MWEs. If we come across a phrase which has two variants, then a separate lemma will be assigned to each variant during the automatic lemmatisation, e.g., *užrašų knygutė* and *užrašų knygelė* ('a notebook', the difference lies in the diminutive suffix of the nouns). However, several forms of degree, different forms of definiteness could be used in the same MWE; for this reason, we have to discuss how to reflect all this in a lemma. The substituting component could be presented in angle brackets: *skirti* [*daug/daugiau*] *dėmesio* 'to pay [much/more] attention'; [*nekilnojamas/nekilnojamasis*] *turtas* 'real property' (with a definite or indefinite adjective). Thus, this would indicate that some syntagmatic lemmas contain substituting components.

References

Loïc Boizou, Gintarė Grigonytė, Erika Rimkutė, and Andrius Utka. 2012. Automatic Inference of Base Forms for Multiword Terms in Lithuanian. In *Proceedings of the Fifth International Conference Human Language Technologies – The Baltic Perspective*, pages 27–35.

Vidas Daudaravičius and Rūta Marcinkevičienė. 2004. Gravity Counts for the Boundaries of Collocations. *International Journal of Corpus Linguistics*, 9(2):321–348.

Jolanta Kovalevskaitė. 2014. Phraseme-type and Phraseme-token: a Corpus-driven Evidence for Morphological Flexibility of Phrasemes. *Res Humanitariae*, XVI, pages 126–143.

Elizaveta Loginova, Anita Gojun, Helena Blancafort, Marie Guégan, Tatiana Gornostay, and Ulrich Heid. 2012. Reference Lists for the Evaluation of Term

Extraction Tools. In *Proceedings of the 10th International Congress on Terminology and Knowledge Engineering (TKE)*, pages 177–192, Madrid, Spain.

Rūta Marcinkevičienė. 2010. *Lietuvių kalbos kolokacijos*. Vytauto Didžiojo universitetas, Kaunas, Lithuania.

Jonas Paulauskas (ed). 2001. *Frazeologijos žodynas*. Lietuvių kalbos institutas, Vilnius, Lithuania.

Erika Rimkutė and Vidas Daudaravičius. 2007. Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. *Kalbų studijos*, 11:30–35.

Erika Rimkutė, Agnė Bielinskienė, and Jolanta Kovalevskaitė (eds). 2012. *Lietuvių kalbos daiktavardinių frazių žodynas*. Vytauto Didžiojo universitetas, Kaunas, Lithuania.

Gregor Thurmair and Vera Aleksić. 2012. Creating Term and Lexicon Entries from Phrase Tables. In *Proceedings of the 16th EAMT Conference*, pages 253–260.

Vytautas Zinkevičius, Vidas Daudaravičius, and Erika Rimkutė. 2005. The Morphologically Annotated Lithuanian Corpus. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pages 365–370.

Vytautas Zinkevičius. 2000. Lemuoklis – morfologinei analizei. *Darbai ir Dienos*, 24:245–273.