Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-27

Automatic Inference of Base Forms for Multiword Terms in Lithuanian

Loïc BOIZOU [a], Gintarė GRIGONYTĖ [b], Erika RIMKUTĖ [a] and Andrius UTKA [a,1]
[a] *Center of Computational Linguistics, Vytautas Magnus University, Kaunas*
[b] *Institute of Computational Linguistics, University of Zurich*

Abstract. This paper reports on a specific problem of automatic terminology extraction in Lithuanian – base form inference. While the process of lemmatisation is properly carried out by existing tools, problems arise with normalizing multiword terms. It can be described as the discrepancy between the base form (i. e. lemma) of a term and the sequence of the base forms of constituent lexical items within a term. Lithuanian is a strongly inflected language and the lemmatisation of each word separately within a multiword term breaks the syntactic relations expressed by inflection (case, gender, number) which need to be kept in order to ensure the cohesion of the term.

Keywords. term extraction, syntagmatic lemmatisation, Lithuanian

Introduction

Domain terminology is a valuable resource which can be widely applied in text processing (e.g. document indexing, retrieval) and information inferring (e.g. relation extraction, ontology building) systems. The reliability and the applicability of terminologies largely depend on the method they are built: human made or automatically extracted. The source of human created domain terminologies for Lithuanian are various paper dictionaries and the online Lithuanian *Term Bank* [1]. As domain specific terminologies are of very dynamic and changing nature, the paper dictionaries are often outdated and could only be used as a source for basic domain terminology, naturally their applicability in text-processing is very limited. The online *TermBank* potentially remains as a valuable domain specific terminology dictionary, however it seems that its functions are prescriptiveness and regulation rather than the comprehensive presentation of terminology. For instance, the terminology of *science and education* in *Term Bank'2012* contains 1355 terms in total, where 90 are tagged as approbated, 5 as recommended for approbation, and even 1260 as unacceptable; the terminolgy of *politics* – 581 terms (169 – approbated, 412 – recommended for approbation, and 0 – unacceptable); the terminology of *public safety* – 483 terms (480 – approbated, 3 – recommended for approbation, and 0 – unacceptable). Therefore, we can claim that the need of domain terminologies in Lithuanian is prevalent.

[1] Corresponding Author: Andrius Utka, Centre of Corpus linguistics, Vytautas Magnus University, K. Donelaiio str. 52-206, LT-44244 Kaunas, Lithuania; E-mail: a.utka@hmf.vdu.lt.

Efforts on the Lithuanian terminology extraction from domain corpora have been presented in [2]. This paper focuses on a specific problem of automatic terminology extraction in Lithuanian, i.e. the inference of base forms of terms in Lithuanian. The method described in this paper has been implemented in the *JungLe* tool. This work is a part of the project ŠIMTAI 2[2].

In sections 1.1, 1.2 and 1.3 we describe in detail the problem of syntagmatic lemmatisation and its underlying causes in morphology and syntactic structure of a multi-word in Lithuanian language. Section 2 presents the approach to automatic inference of base forms of multiword terms in Lithuanian and describes the *JungLe* tool.

1. The Problem

Base forms of terms which should occur in dictionaries and terminology databases are not the same as lemmas. While there is no problem in mapping a base form for a single word terms in Lithuanian, attaching a correct base form for a multi word terminological unit is not a trivial task. Consider for instance two examples of terms – single-word and two-word presented with the following information: a) use case, b) lemma, c) normalized term form, and d) term in English:

a) **mokykl-os** (n. plu. fem. gen.)	**universitet-ų rektor-iams** (n. plu. masc. gen. + n. plu. masc. dat.)
b) mokykla	universitetas rektorius
c) mokykl-a (n. sing. fem. nom.)	universitet-o rektor-ius (n. sing. masc. gen. + n. sing. masc. nom.)
d) school	university rector

The normalization of a term can be described as the discrepancy between the base form of a term and the base forms of constituent elements within a term. The base form of the term is its lexical-conceptual representation and is the form that is preferred in terminology banks. A final domain terminology is going to contain normalized term forms, but not lemmas or a list of concrete use cases.

Therefore the task of term recognition in Lithuanian involves an additional step of base form derivation next to detection of grammatical variants in the text. In some cases of multiword terms direct derivation from the lemma is hard, since several morphological elements need to be coordinated. Consider for instance an example illustrating a derivation of different number, case, gender and a degree of comparison:

a) **aukšt-os-ios-e mokykl-os-e** (adj. plu. fem. loc. comp. d. + n. plu. fem. loc.)
b) aukštas mokykla
c) aukšt-oj-i mokykl-a (adj. sing. fem. nom. comp. d. + n. sing. fem. nom.)
d) higher school

There are several solutions for deriving a normalized form of a term: a) collecting all possible use cases of a term from the reference corpora and selecting the base form, b) inferring the base form of a term from a use case of a term. The first solution is not always reliable as it requires term to be used in its base form, which is not necessarily

[2]"Automatic Identification of Educational and Scientific Terminology (ŠIMTAI 2)" supported by The National development programme of Lituanistics (2009-2015) (grant No. LIT-2-44)

present in the reference corpus. So these unsolved cases would require inference of the base form as well.

In the following sections we explain the morphological complexity of possible term forms and describe observed syntagmatic patterns in multi-word terms which we later apply in deriving the base form of a term.

1.1. The Role of Grammatical Features for Base Form Inference

When inferring base forms of Lithuanian terms, one needs to take into account certain grammatical categories. The most important of them is a part-of-speech. The other important grammatical categories are number and case, while gender category for this task is not critical. However, the gender becomes important, during the case syncretism, e.g. the word *darbuotojų* in the combination *mokslo darbuotojų* (en. *academic workers*) by a morphological analysis tool is analyzed as plural genitive with either feminine (base form – *mokslo darbuotoja*) or masculine (*mokslo darbuotojas*) gender. In both cases the base form of the term should be given in masculine gender – *mokslo darbuotojas*.

There are several variations in the category of number, e.g. in some cases only the first constituent of the term *studentų atstovybė* (en. *students' agency*) has to be in plural; in others all the constituents in terms *akademiniai įgūdžiai* (en. *academic abilities*), *auditorinės darbo valandos* [en. *class hours*] have to be in plural; while all the constituents of the terms *fakulteto taryba* (en. *Council of the faculty*), *universiteto autonomija* (en. *university's autonomy*), *bendrasis priėmimas* (en. *general admission*) have to be in singular.

The corpus analysis shows that some of the term constituent words are used only in plural form (e.g. *asignavimai* [en. *assignations*], *duomenys* [en. *data*], *studijos* [en. *studies*], *rūmai* [en. *House*], *žinios* [en. *knowledge*], *pinigai* [en. *money*], *pareigos* [en. *duties*], even though dictionaries present them in singular as a default form.

Due to this mismatch between real usage and dictionary information, the morphologic analyzer [3] which is based on dictionary information would wrongly assign singular base forms for terms like *mokinio pasiekimas* [en. *a schoolchild's achievement*], *švietimo resursas* [en. *resource of education*], *praktinis gebėjimas* [en. *practical ability*], *fizinis mokslas* [en. *physical science*], *akademinis užsiėmimas* [en. *academic activity*], as they should be in plural – *mokinio pasiekimai, švietimo resursai, praktiniai gebėjimai, fiziniai mokslai*).

The assignment of appropriate number is sometimes complicated due to homonymy, e.g. *studija* (en. *scientific written work*) and *studijos* (en. *studies*). This phenomenon has caused wrong assignment of base forms for the terms: *bakalauro studija* (should be – *bakalauros studijos*, en. *Bachelor studies*), *nuotolinės studija* (*nuotolinė studija*, en. *distance studies*).

However, sometimes information about a part-of-speech, a number, a gender and a case is not enough. In order to solve the assignment of a base form, some word combinations need to have syntactic information present. For instance, the base form of the combination *aukštoji mokykla* (en. *high school*) may only be correctly assigned, if we have the rule, that adjective and participle combine with a noun, but if pronominal forms of adjective or participle are encountered, then pronominal forms should be preserved in their base forms.

In some cases the word order of the base form of a term differs from the word order that can be observed in the corpus, e.g. *tipų studija* (should be – *studijų tipas*, en. *type of studies*), *lygmens kvalifikacija* (*kvalifikacijų lygmuo*, en. *qualification level*).

1.2. Grammatical Term Structure

In order to abstract major syntagmatic templates for multi-word term expressions and use them as rules for the syntagmatic lemmatizer, we have analyzed grammatical structures of around 800 terms in the domain of the Education and Science. Table 1 presents length distributions of the terms.

Table 1. Length distribution of terms.

Length of terms in words	Number of occurrences	Proportion
1	155	19.87 %
2	474	60.77 %
3	125	16.03 %
4	22	2.82 %
5	4	0.51 %
In total:	780	100 %

The most frequent templates for inferring base forms of terms are summarized as following. The structure of two-word terms can be generalized by 3 main grammatical patterns:

1. NOUN GEN. + NOUN SG. NOM. (52,7% of two-word terms) (e.g. *studijų sritis* (en. *field of study*), *profesijos mokytojas* (en. *profession teacher*), *mokslo institutas* (en. *institute of science*))
2. ADJ. SG. NOM. + NOUN SG. NOM. (42,2%) (e.g. *aukštoji mokykla* (en. *high school*), *moksliniai tyrimai* (en. *scientific research*), *aukštasis mokslas* (en. *higher education*)),
3. PARTIC. SG. NOM. + NOUN SG. NOM. (5%) (e.g. *baigiamasis darbas* (en. *final paper*), *pasirenkamasis dalykas* (en. *arbitrary subject*), *suaugusiųjų švietimas* (en. *education of adults*)).

Three-word terms can mainly occur in 7 grammatical patterns. The most frequent ones are:

1. ADJ. GEN. + NOUN GEN. + NOUN NOM. (48% of three-word terms) (e.g. *neformalus suaugusiųjų mokymas* (en. *informal education of adults*), *aukštojo mokslo institucija* (en. *high education institution*)).
2. NOUN GEN. + NOUN GEN. + NOUN NOM. (24,8%) (e.g. *studijų krypties reglamentas* (en. *study field regulation*), *studijų krypčių aprašas* (en. *study field inventory*).
3. ADJ. NOM. + ADJ. NOM. + NOUN NOM. (12%) (e.g. *netiksliniai moksliniai tyrimai* (en. *inexpedient scientific research*), *nebiudžetiniai finansiniai ištekliai* (en.*non-budget financial resources*), *netiesioginis centralizuotas valdymas* (en. *indirect centralised management*)).

Four-word terms can mainly occur in 9 grammatical patterns:

1. ADJ. NOM. + NOUN GEN. + NOUN GEN. + NOUN (22,7%): *tarptautinė mokslo duomenų bazė* (en. *international database of scientific data*);

2. ADJ. NOM. + ADJ. GEN. + NOUN GEN. + NOUN (18,2%): *bendrasis universitetinio lavinimo dalykas* (en. *general subject of university education*);
3. ADJ. GEN. + NOUN GEN. + PARTIC. NOM. + NOUN (13,6%): *specialiųjų poreikių turintis mokinys* (en. *pupil with special needs*);
4. NOUN GEN. + CONJ. + NOUN GEN. + NOUN (13,6%): *mokslo ir studijų institucija* (en. *institution of science and studies*);
5. ADJ. GEN. + NOUN GEN. + ADJ. NOM. + NOUN (9,1%): *aukšto lygio moksliniai tyrimai* (en. *high level scientific research*);
6. PARTIC. NOM. + ADJ. NOM. + NOUN GEN. + NOUN (9,1%): *pripažinta tarptautinė duomenų bazė* (en. *approbated international database*);
7. ADJ. NOM. + ADJ. NOM. + ADJ. NOM. + NOUN (4,5%): *bendroji nacionalinė kompleksinė programa* (en. *common national complex programme*);
8. NOUN GEN. + ADJ. GEN. + NOUN GEN. + NOUN (4,5%): *valstybės mokslinių tyrimų Įstaiga* (en. *state institution of scientific research*);
9. NOUN GEN. + NOUN GEN. + NOUN GEN. + NOUN (4,5%): *technologijos mokslų studijų sritis* (en. *study field of technology*);

Since we have only a few five-word terms in the acquired list of terminology structural generalizations cannot be made about this group of terms.

1.3. Term Cohesion in Lithuanian

Morphology plays an important role as a factor of cohesion (as well as word order) in nominal syntagms including multiword terms in Lithuanian. Morphology ensures the differentiation of the three main syntactic relations of multiwords:

1. agreement (or congruence) of grammatical features;
2. government (or reaction), that is the selection of a given case as a dependency mark;
3. adjoinment (or coordination), that is a combination of words when one member is related in terms of meaning, but is independent in terms of grammatical form.

Prototypically the relation of agreement ties adjectives, particles, some pronouns, numerals to a noun head, while the relation of government ties nouns to a noun head and nouns to verbs [4] and [5]. Other possible combinations include: agreement between nouns (combined nouns, e.g. *mokslininkas stažuotojas* [en. *postdoctoral fellow*], lit. **researcher trainee*), and governed adjectives (with nominalisation of adjectives/participles, e.g. *suaugusiųjų mokymas* [en. *adult education*], lit. **grown-up education*). The relation of adjoinment is less important for terminology, as typically this relation is characteristic to combinations of adverbs and verbs, participles and conjugated verbs, infinitives and conjugated verbs. It should be noted that other researchers classify syntactic relations differently, i.e., according to [6] there are two relations: government and modification. The latter includes the adjoinment relation.

Consequently, lemmatization, which returns all the nominals at the nominative case, breaks the cohesion of a term. That is why the list of lemmatised word forms of a multiword term is no more a coherent syntactic unit, but rather an unconnected sequence of words.

2. Method

In this section we describe the approach for inferring the base form of multiword terms based on grammatical term structures (described in section 1.2) which preserves grammatical cohesion.

2.1. Design Principles

The software *JungLe* (from *junginiu lemuoklis*, that is, syntagm lemmatiser) is not a full lemmatiser. It is designed as a layer assisting the general lemmatiser of the Lithuanian language Lemuoklis [3]. *JungLe* to a large extent uses information for each word form (contextual form, lemma and grammatical information) provided by *Lemuoklis* in order to derive the base form of a term.

JungLe is implemented in Haskell. It resorts to a set of Haskell modules internally designed for tasks related to NLP like, for instance, the datatype of annotated word, the dependency grammar module, the interface module for the annotation format used by Lemuoklis.

2.2. The JungLe Algorithm

In *JungLe* the syntagmatic lemmatization is a two step process: identification of syntactic relation in a term followed by re-lemmatization – assigning matching paradigms for each words in a term.

2.2.1. Identification of Syntactic Relations within a Term

The syntagm lemmatization has to process three different types of word types inside a term:

- the head (the syntactic top node)
- the congruent words (with the head)
- the non-congruent words (with the head)

The first step is the head identification. It must be noticed that for an overwhelming majority of terms the head is the last word.

JungLe differentiates between nominative case terms and non-nominative case terms. In the case of nominative case terms, the head is the last nominative noun, or the last nominative adjective, if there is no nominative noun. In non-nominative terms, the syntactic analysis needs to be carried out. By syntactic analysis we mean a simplified dependency grammar which describes the structure of nominal syntagms without embedded preposition.

The foremost step is to recognize the head of a term. Then its congruent words are detected by looking for other words which forms have the features of number, gender and case.

Given the relatively low number of grammatical term structure models, there are only few causes of mistake, i.e., the main one is the ambiguity of the genitive case which arises when a genitive adjective appears along with several genitive nouns and that makes the governing node unclear. Consider for instance:

- *mokslinių tyrimų institutų* 'institutes of scientific research' ⟶ *mokslinių tyrimų institutas* or **mokslinis tyrimų institutas*
- *paskolų gyvenimo išlaidoms* 'credit for life spending' ⟶ **paskolų gyvenimo išlaidos* or *paskolos gyvenimo išlaidų*

The first example above illustrates a typical case where the ambiguity is whether the adjective is govern by the first or the second noun. In the second example (which is quite rare), the ambiguity is harder to solve because it concerns the determination of the top node *paskolos* 'credit' (the right one) or *išlaidos* 'spending'.

The step of the identification of the different syntactic components of terms is followed the generation of the suitable term lemma.

2.2.2. Re-lemmatisation

The basic rules followed while generating the base form of a term are: a) the base form of the head is set as the lemma, b) the non-congruent words keeps their contextual form and c) the congruent words need to be re-lemmatised.

The first goal of the re-lemmatisation is to restore the congruence with the head. *JungLe* ensures that the gender and number of the congruent words have to match the gender and number of the head. All word forms have to be in the nominative, which is the lemmatic case. Besides, this re-lemmatisation has to preserve the lexicogrammatical features of definiteness and degree, which take part in the term structure.

For participles, it requires an additional step to rebuild a new lemmatic stem. According to the Lithuanian grammatical tradition, participles are lemmatised as infinitive which does not fit the term structure, therefore *JungLe* lemmatizes participles in a similar way like adjectives, i.e. in case, number and gender.

The re-lemmatisation contains two phases: stemming and generation. The stemming involves the removing of the ending and the depalatisation, if needed:

- . . . či- ⟶ . . . t-
- . . . dži- ⟶ . . . d-

The generation of the lemma is based on a cascading grammar structure and a simple string concatenation, if necessary with an adaptation of the stem when required by the consonant alteration rules before palatalising endings:

- . . . t- ⟶ . . . č-
- . . . d- ⟶ . . . dž-

The cascade of test is based on: the head's grammatical features of number and gender; on the paradigm of the removed ending (during the stemming) and the grammatical annotations of the word form (for the lexicogrammatical features). It must be emphasized that the generative component, which is hard-coded, is restricted as it concerns only the nominative case of the main adjective/participle paradigms.

3. Evaluation

The preliminary evaluation of *JungLe* is carried out on the basis of the list of 827 extracted multiword terms which were annotated by experts. The mistakes which arise

from JungLe come either from incorrect morphological analysis carried out by Lemuoklis or from incorrect re-lemmatisation. Table 2. summarizes different types of inference mistakes for base forms of terms.

Table 2. Types of inference mistakes for base forms of terms.

incorrect input	2	0.2 %
incorrect analysis	6	0.7 %
incorrect re-lemmatisation	10	1.2 %
incomplete re-lemmatisation	27	3.3 %
number of terms	827	100 %

JungLe provides an accuracy close to 95 %. The qualitative analysis shows some obvious tendencies:

- Incorrect analysis, which is rather rare, arises from the genitive ambiguity. Such an ambiguity cannot be resolved at the syntagm level.
- Incorrect re-lemmatisation is due to mistakes in the generation module, which generates an ungrammatical lemma. The cause of the problem is the stemming of definite participles.
- Incomplete re-lemmatisation is also caused by some mistakes in the generation module: most of the cases (17) are related with the pluralia tantum noun *studijos* 'studies' (e.g. **universitetinės studija* instead of *universitetinės studijos*), which appears frequently in multiword terms; other cases are related to the suffix *-in* (e.g. **socialinę stipendija* instead of *socialinė stipendija*) and to definite forms (e.g. **aukštosiose mokykla* instead of *aukštoji mokykla*).

4. Conclusion

In this paper we have presented an approach for syntagm lemmatisation of multiword terms in Lithuanian. The approach is based on detecting syntactic governance and adjusting the grammtical form of the congruent word. The evaluation of the *JungLe* showed an accuracy close to 95 %.

Term normalization allows to generate a canonical term representation form, independent of term's contextual variation. The *JungLe* tool, which reaches a high accuracy with minimal programming redundancy provides the missing link between the corpus item obtained by automatic terminological extraction methods and the dictionary data. Thus, automatically extracted terminological data can reach dictionary databases in a shorter time.

Acknowledgments

This study is a part of the project "Automatic Identification of Educational and Scientific Terminology (ŠIMTAI 2)" supported by The National development programme of Lituanistics (2009-2015) (grant No. LIT-2-44).

References

[1] http://terminai.vlkk.lt/pls/tb/tb.search
[2] G. Grigonytė, E. Rimkutė, A. Utka, L. Boizou, *Experiments on Lithuanian Term Extraction*, NEALT Proceedings Series **11** (2011), 82–8.
[3] Zinkevičius, Vytautas, *Lemuoklis – morfologinei analizei [Morphological analysis with Lemuoklis]*. In: Gudaitis, L. (ed.) Darbai ir Dienos 24 (2000). 246–273.
[4] V. Labutis, *Lietuvių kalbos sintaksė*, Vilniaus universiteto leidykla, Vilnius, 1998.
[5] V. Ambrazas (ed.), *Dabartinės lietuvių kalbos gramatika*, Mokslo ir enciklopedijų leidykla, Vilnius, 1996.
[6] A. Holvoet, A. Judžentis (ed.), *Sintaksinių ryšių tyrimai*, Lietuvių kalbos institutas, Vilnius, 2003.