Experiments on Lithuanian Term Extraction

Gintarė Grigonytė, Erika Rimkutė, Andrius Utka and Loic Boizou
Centre of Computational Linguistics (http://donelaitis.vdu.lt)
Vytautas Magnus University
Kaunas, Lithuania
{g.grigonyte, e.rimkute, a.utka, l.boizou}@hmf.vdu.lt

Abstract

This paper explores the problem of extracting domain specific terminology in the field of science and education from Lithuanian texts. Four different term extraction approaches have been applied and evaluated.

1 Introduction

Term extraction nowadays is becoming automated and a well defined process that contains phases of NLP in the levels of morphology, syntax, and sometimes semantics (Sager, 1990; Cabre, 1992). Even though NLP tools have never reached very high reliability, there is a wide choice of term extraction applications for widely spread Indo-European languages like English (Pantel and Lin, 2001), French (Daille, 1994), Polish (Piskorski et al., 2004), and Russian (Mitrofanova and Zakharov, 2009).

The research strategies of term extraction can be divided into statistically-based and linguistically-based[1] (Cabre et al., 2001). Rarer languages often do not have the luxury of linguistic tools for automatic text processing. One can argue that a possible solution could be statistically-based tools, which are claimed to be language independent. However, there is a lack of evaluation of such tools for rare languages.

Even though there are some rapid advances in Lithuania's HLT[2], automatic term extraction is still quite a new and unexplored field. First attempts of using commercial term extraction tools for Lithuanian were described by Zeller (2005).

The present paper deals with the automatic extraction of Lithuanian domain specific terminology in the field of education and science. In the following subsections we will describe the terminology situation in Lithuania and several Lithuanian language specific pitfalls that are relevant for linguistically or statistically based term extraction systems.

1.1 Terminology Situation in Lithuania

The main volume of Lithuanian terminology is available at the Lithuanian Terminology bank[3].

The Terminology bank is being run and constantly updated by the Commission of the Lithuanian Language[4] together with the Office of Lithuanian Seimas[5]. Presently, the bank keeps records of 150 thousand terms and their definitions of various domains, e.g. machinery, computer science, medicine, etc. Naturally, there is a large number of domain specific databases and dictionaries in various institutions that do not always include officially accepted terms.

In Lithuania until now terms have been composed, chosen and approbated on the basis of inconsiderable amount of texts, intuition, and the norms of the Lithuanian language. This is a traditional prescriptive way of term definition that does not satisfy contemporary needs of the language.

However, there is a great urge for changes in the Lithuanian terminology, as now there is a constant lack of terms and a large number of incorrectly translated terms. Furthermore, new variants of terms occur much faster than the definition of a term. Quite often the standardized terms are not willingly accepted by the society.

Obviously this paper takes the descriptive path, as it is an attempt to find an efficient and robust way to extract a domain specific terminology without any prescriptive judgment.

[1] Hybrid approaches combine both strategies: usually linguistic analysis followed by statistical filtering.

[2] More about Lithuanian HLT in Marcinkevičienė and Vitkutė-Adžgauskienė (2010).

[3] http://terminai.vlkk.lt:10001/pls/tb/tb.search

[4] http://vlkk.lt/

[5] http://www.lrs.lt/

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 82–89

1.2 The Language Related Problem

Lithuanian is a highly inflective language. For example, Lithuanian nouns, adjectives and particles typically have 7 cases in singular and 7 in plural, which makes 14 different wordforms of a single-word. Additionally some Lithuanian nouns, adjectives, participles, pronouns, and numerals can be used in three different genders (feminine, masculine, and neuter), which again adds to a variety of forms. This proliferation of inflections makes the statistical automatic identification of terms more complicated, as distinct wordforms of terms appear very infrequently.

The solution for this is the morphological tagging, which again is complicated due to many morphological categories and morphological ambiguity, which exists in Lithuanian in spite of rich variety of wordforms. Unlike in other languages like for instance Malay morphological categories do not necessarily resolve ambiguity as ambiguity is present within lemmas, e.g., "laiko" (noun "time" and verb "hold"), and within wordforms, e.g. "prekės" (sing. noun gen. and pl. noun nom.).

Besides, the linguistic approach needs an answer to the question, which of grammatical categories are necessary for the successful extraction of term candidates and which can be ignored. One thing is obvious that the part-of-speech category is not enough for Lithuanian.

The categories of gender and number are not very helpful in distinguishing between terms and non-terms. For example, if we consider the most productive two-word term combination N + N in Lithuanian, then in the terms like *dėstytojų kompetencija* (competence of teachers), *studentų atstovybė* (students' organization) the first noun should have plural Genitive form, while in the terms like *fakulteto taryba* (faculty board), *universiteto autonomija* (autonomy of the university) the first noun should have singular Genitive form.

In other cases the second noun needs to be either in plural (*akademiniai įgūdžiai* (academic skills), *auditorinės darbo valandos* (class hours)) or in singular (*bendrasis priėmimas* (common enrollment), *mokslinis leidinys* (scientific publication)). The category of gender may not be a distinguishing feature either, as each constituent of a term can potentially be in feminine or masculine gender.

It seems that the only useful additional feature in the N + N combination is the Genitive case of the first constituent, as it remains stable across many variants.

Additional complications arise with three-word or longer terms. For example, if we take the term *mokslinių tyrimų įstaigos atestacija* (certification of institution of scientific researches), then its structure is

A pl. Gen + N pl. Gen + N sg. Gen + N sg. Nom

Such long terms often consist of several combinations of words, where their syntactic relations might differ. For example, in the term *neformaliojo suaugusiųjų švietimo įstatymas* (law of informal education of adults) with a structure of A sg. G. + N pl. G. + N sg. G. + N sg. N syntactic relations are spread as follows: (figures indicate the elements of word combinations):

1 ⟵ 3

2 ⟵ 3

[123] ⟵ 4

Beside language related problems, there are some universal terminology identification problems typical to all languages. One of such problems is determining term boundaries. For example, the word combination *dėstytojų ir mokslininkų kvalifikacijos bei kompetencijos atitiktis* (correspondence of qualification and competence of teachers and scientists) may give birth to one term or several terms: 1) *dėstytojų kvalifikacijos bei kompetencijos atitiktis* (correspondence of qualification and competence of teachers), 2) *mokslininkų kvalifikacijos bei kompetencijos atitiktis* (correspondence of qualification and competence of scientists), 3) *dėstytojų ir mokslininkų kvalifikacijos atitiktis* (correspondence of qualification of teachers and scientists), 4) *dėstytojų ir mokslininkų kompetencijos atitiktis* (correspondence of competence of teachers and scientists) (more combinations of possible terms (concerning their boundaries) are possible).

Finally, the question, whether a particular stable word combination is a term or not, is faced by both human experts and computer programs. However, even if a word combination is a term, yet another judgment on its specificity, i.e. domain term vs. general term, is required. The solution of such problem is possible only with the help of an expert.

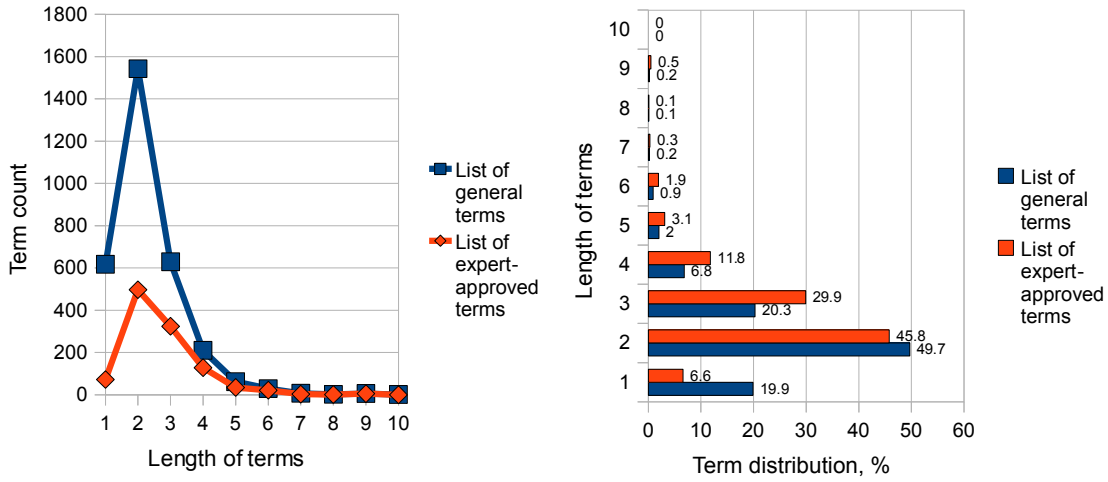

Figure 1: Absolute frequency distribution (1st part) and relative frequency distribution (2nd part) of term lengths.

2 Experimental Settings

2.1 The Corpus

An experimental 103,893 token corpus of education and science has been compiled specifically for term extraction experiments. This corpus consists of laws, orders, regulations, resolutions, memoranda, descriptions, overviews, notes, reports, newsletters, programmes, summaries, and standards. The texts mainly encompass the following topics: high education policy, research policy, continuous education, and professional education policy. Two reference lists have been manually created for the analysis of terms and evaluation purposes: the first one is a larger list of all general terms, and the second one is the special list of terms in the field of education and science.

Firstly, 5 linguists have identified a list of general terms (in total 3,106 lemmas). Problem cases and disagreements have been solved by consultation between the linguists. Then this provisional list has been further reviewed by the education expert, who has identified the terms of science and education. As a result of this review, the list has shrank to 1,085 terms. The expert mostly removed general administrative terms and terms that are not related to the previously mentioned topics.

The analysis of distribution of term lengths in the reference lists has been performed, as term length is very important characteristic to many term extraction methods. The analysis of term types has shown that the term length can vary from one to ten words (see Figure 1). Although two-word terms are by far most frequent (49.7 % of the total number of terms), they cover only a half of all terminology.

In the expert list, the part of single and two-word terms has significantly decreased, while the part of three and four-word terms has significantly increased. The distribution of term lengths shows that terms with lengths from 1 to 4 words make up 96.7% of all terms in the general list and 94.1% in the expert list. Many extraction methods typically leave out single-word and four-word terms, although it is evident from this analysis that they are quite important in the overall coverage of terminology.

2.2 Statistical Approaches

Statistically-based term extraction aims at detecting syntagmatic collocations or keywords, which are relevant for the domain[6]. Many statistic measures can be applied for term extraction task starting with frequency rank, Mutual Information (Church and Hanks, 1989), Dice coefficient (Smadja et al., 1996), T-score (Church et al., 1991), Log-likelihood (Dunning, 1993), C-value (Frantzi and Ananiadou, 1996), and others.

A major advantage of statistical TE systems is that they do not require huge databases with term patterns constructed by humans and also do not require running through language analysis pipelines. The assumption in using statistical models in TE is that words which tend to co-occur together are related and therefore they are likely term candidates.

[6]These features are referred as unithood and termhood.

In this section, we will deal with three statistical approaches, namely keyword cluster identification, keyword extraction with machine learning, and collocation extraction.

2.2.1 Keyword Clusters

When analysing abilities of extracting terms by the linguistic program *WordSmith Tools* (Scott, 2008), the idea of extracting *keyword clusters* as terminology candidates has emerged. The idea is based on the assumption that the most frequent clusters of keywords in a given text may also point to terminology.

Keywords are identified by comparing a given text's frequency list to the frequency list of a large reference corpus by using Ted Dunning's (Dunning, 1993) log likelihood test. The final step is calculating keyword clusters that are two or more words, which are found repeatedly near each other (1-3 intervening words may be present). Below is the list of 10 most frequent keyword clusters and their frequencies:

mokslo [.] studijų (studies of science)	*524*
švietimo [.] mokslo (education science)	*522*
aukštojo mokslo (higher education)	*474*
mokslinių tyrimų (scientific research)	*390*
suaugusiųjų švietimo (adult education)	*292*
lietuvos respublikos (lithuanian republic)	*284*
protų nutekėjimo (brain drain)	*228*
švietimo ministerija (ministery of education)	*220*
profesinio mokymo (professional teaching)	*184*
neformaliojo [.] švietimo (non-formal education)	*164*

Even though the resulting list seems very promising, the comparison in terms of precision and recall with the reference lists has not shown good results. The best result with the whole keyword list has produced 9.7% of recall and 1.8% of precision when compared to the expert's list and 7.5% of recall and 3.7% of precision when compared to the list of general terms (see Table 2 for all results).

2.2.2 Machine Learning: KEA

Many NLP tools require a process of machine learning, which allows the system to improve the quality of results by its own experience or the supervision of a human.

The keyword extractor KEA[7] implements such methods. The core of the KEA system is a statistically-based algorithm with a machine learning system generating an extraction model. The multi-featured algorithm rates keywords taking in account four components:

- degree of specificity;
- position in the text;
- length of the phrase;
- node degree.

The learning process, which is based on Naive Baysian Method using the software WEKA[8], requires a manual extraction of keywords from a training corpus. A clear overview of the whole system may be found in Medelyan (2005). Besides, the core may be extended by language specific Java components as a stop-words list and a stemmer.

From a terminological perspective, it must be noticed that a keyword is not the same as a term. However, in a corpus made of specialized texts, we may expect a significant matching between the set of keywords and the set of terms.

The task of machine learning has been based on a subset of the test corpus consisting of four texts, representing approximately 22,000 words, about 1/5 of the whole experimental corpus. The manually selected keywords have been the terms appearing in the sub-corpus. Given the highly inflected nature of the Lithuanian language, two attempts of machine learning have been carried out. In the first case, the list of manually extracted keywords has included only the main nominal forms (Nominative, Accusative and Genitive, if relevant singular and plural). In the second case, the list has been restricted to the base forms (lemma). The extraction results have been identical for both methods.

Once the extraction model has been built, three different approaches have been tested : 1) with no extension, 2) with a stop-words list and 3) with both a stop-words list and a stemmer. The first approach has produced results of insufficient quality with such terms as *mokslo ir* (science and). This has led to the conclusion that a stop-word list would enhance the quality. On the basis of the results given by the rough extraction, a list of more than one hundred stop-words has been compiled. In order to avoid an artificial *ad hoc* correction of the results, only words of some specific groups have been included in the stop-words list:

- grammatical words such as prepositions, con-

[7] http://www.nzdl.org/Kea/

[8] http://www.cs.waikato.ac.nz/ml/weka/

junctions, particles, some adverbs and frequent forms of pronouns;
- some expressions of quantity;
- main abbreviations;
- some general verb forms (to be, to show, modal verbs).

The use of this restricted stop-word list obviously has improved the results, avoiding some common mistakes of highly improbable terminological combinations.

Besides, the third variant adding a stemmer has also been tested, but the quality of the results has been the worst of all three. The Lithuanian results have appeared as a confirmation of a statement in the documentation of KEA's source code - "We have obtained better results for Spanish and French with NoStemmer".

Given the preliminary results of the different methods, the term candidates have been extracted on the basis of the second variant, that is, with a stop-words list only. Subsets of different sizes have been extracted (10,000, 5,000, 1000, 500 forms of term candidates).

2.2.3 Collocation Extraction Methods

Collocations and terminology are related concepts, but this relationship is not a synonymic or a simple one. It is a well known fact that the term of collocation is very broadly understood, and therefore it is important at the very start to define which notion of collocation is used in the current paper. In this paper we will deal only with statistical collocations, that is with sequences of words that co-occur more often than would be expected by chance.

Automatically extracted collocations have been used for terminology extraction process many times (e.g., Daille, 1996, Azé et al., 2005 etc.). As a rule, collocation extraction methods are used as the first step in creating a terminology candidate list, which is then further processed by ranking and extracting the relevant items. However, there are at least two problems associated with these methods. The first problem is that there is quite a large number of statistical collocation extraction methods (e.g. Azé et al., 2005 deal with 13 methods), and it is not a trivial task to choose the best one for the terminology extraction task. The second problem is that the majority of collocation extraction methods are limited to extracting two to three word collocations, while the range of term lengths, as our analysis has shown, is rather more broad.

The latter statement can be supported by the analysis of term lengths in 103,893 word experimental corpus (see Figure 1).

Due to these reasons, it has been decided to try only the tools and methods that extract collocations of variant length, i.e. LICE (Gravity Counts)[9], and leave out other tools that extract fixed length collocations.

Gravity counts (Daudaravičius and Marcinkevičienė, 2004) is a method for determining borders or collocations. It is based on the idea that all words in a text are more or less tied, that is, the degree of attraction between them may be stronger or weaker. Gravity count G for two words x and y in this order is calculated according to the formula:

$$G(x,y) = \log\left(\frac{f(x,y)\cdot r(x)}{f(x)}\right) + \log\left(\frac{f(x,y)\cdot l(y)}{f(y)}\right) \quad (1)$$

where f(x) is the frequency of x, f(y) the frequency of y, f(x,y) the frequency of the two word co-occurring in this order, r(x) the number of different words to the right of x and l(y) the number of different words to the left of y.

The software LICE, which is designed to implement the gravity counts, has been used to collect collocations with the aim of comparing them with the set of terms manually extracted. It must be emphasized that LICE does not extract only multiword expressions, since the program is designed to indicate the limits of collocations. Thus, if a group of consecutive words shows a significant degree of attraction, they appear as a collocation, but words which are loosely tied to others appears separately in the results given by LICE.

The result given by LICE has been processed in order to extract the multiword expressions. Then, in order to improve the result in the same way it has been done for KEA, a stop-words list (the same as prepared for KEA) has been used as a second filter. It must be emphasized that even after the filtration the number of expressions remains very high (more than 16,000 terms).

2.3 Linguistically-based Approach

The process of linguistically-based term extraction is a pipeline that may include morphological analysis, syntactic parsing, and a module of linguistic rules (patterns) that describes terms.

[9] LICE is an experimental piece of software used for internal research at CCL VMU.

N Gen	N Nom			638
A Nom	N Nom			610
N Nom				484
A Gen	N Gen	N Nom		168
N Gen	N Gen	N Nom		145
A Nom	N Gen	N Nom		73
PART Nom	N Nom			66
A Nom	A Nom	N Nom		42
A Gen	N Gen	N Gen	N Nom	37
V inf				31

Table 1: Morphological patterns in the general list of terms (where N - noun, A - adj, PART - participle, Nom - Nominative, Gen - Genitive).

The approach is language dependent as terms in different languages have different morphological patterns. Morphological patterns may include part-of-speech categories for analytic languages (e.g. English), or additional grammatical categories such as cases for synthetic languages (e.g. Lithuanian), or syntactic categories (e.g. noun phrases).

Typically this approach requires an annotated corpus, which needs to have an appropriate annotation scheme (e.g. POS, POS+case, or syntax). The term extraction tools simply extract all occurrences of required patterns from the annotated corpus and produce a list of term candidates that can be manually reviewed, statistically processed, or filtered with the help of stop-word lists.

Linguistic rules can be coded as regular expressions and directly used for identifying term candidates. An example of such a rule for a single-word term is [noun], for two-word terms - [noun]+[noun] and [adjective]+[noun]. Morphological patterns that have not been coded into a term extractor will produce low recall, while non-terms that coincide with the programmed patterns will reduce precision. Lopes et al. (2010) have shown that linguistic approaches produce better results than statistical ones, besides they also emphasize the fact that linguistic approaches are more complicated in comparison to easy adaptable statistical methods.

For the present study, the manually extracted terminological list has been morphologically tagged[10] and a frequency list of morphological patterns has been built. In order to avoid unnecessary diversity of the patterns, only categories of part-of-speech and case have been considered. The list of top ten grammatical patterns with frequencies of occurrence in the lemmatized list of general terms is given in Table 1.

A set of 27 morphological patterns has been selected for the extraction of term candidates from the annotated experimental corpus. The list of patterns includes mostly combinations of nouns and adjectives, sometimes with the intervening conjunction *ir* 'and'. In order to limit the noise, only multiword patterns have been considered. The maximal length of morphological patterns is five words. A Haskell function has been specially developed for this goal. Some deficiencies of this approach result from the tagging process which can give inaccurate analysis or fail to analyse an unknown word.

2.4 Evaluation

The evaluation of a term extraction system can be addressed with measuring against the *gold standard*. Two term reference lists have been set (see section 2.1) for the purpose of evaluating the four different term extraction approaches.

The quality performance of the term extraction system can be evaluated in terms of *precision* and *recall*. Which are the equivalent of inverted measures of *silence* and *noise* proposed by (Cabre et al., 2001).

2.5 Results

All the test results of the above tested methods and tools are summarized in the Table 2.

Concerning the recall, there are several objective reasons, which have negatively influenced the results. In case of KEA and the linguistic approach, candidate terms longer than 5 words have not been extracted, while these patterns represent between 1.4% and 2.8% of the manually extracted terms. Similarly, the linguistic, keyword clusters approach and LICE, have not taken into account single word terms, which represent between 6.6% and 19.9%.

Moreover, some discrepancy comes from terms in the expert's reference list, that actually have not been present the experimental corpus. The expert's reference list includes 158 items absent from the corpus, which represents 14.5% of the list and has a strong influence on the recall rates. Except for few direct additions, for example *akademikas*

[10]The tagging has been performed using the morphological analyzer developed at CCL VMU. Rimkutė and Daudaravičius (2007) have established that the precision of the tagger is 94% for establishing grammatical categories and 99% for lemmatisation.

Method	Tools	Reference	Candidates	Lemmas	Match*	Recall	Precision
Keyword clusters	WSmith	General list	10777	5959	219	7.5	3.68
			5000	3096	186	6.4	6
			1000	734	105	3.6	14.3
			500	397	69	2.4	17.4
		Expert list	10777	5959	105	9.7	1.8
			5000	3096	88	8.1	2.8
			1000	734	53	8.1	7.2
			500	397	35	3.2	8.8
Keywords	KEA	General list	10000	6398	865	29.7	13.5
			5000	3165	629	21.6	19.9
			1000	703	238	8.2	33.9
			500	381	157	5.4	41.2
		Expert list	10000	6398	269	24.8	4.2
			5000	3165	197	18.2	6.2
			1000	703	77	7.1	11
			500	381	51	4.7	13.4
Gravity Count	LICE	General list	16593	14627	1124	38.6	7.7
		Expert list	16593	14627	388	35.8	2.7
Linguistic approach	Tagger, scripts	General list	25058	18990	1801	61.8	9.5
		Expert list	25058	18990	713	65.7	3.8

*number of automatically extracted terms that match terms in reference lists.

Table 2: Evaluation of Terminology Extraction Tools.

'academician', it is mainly due to a significant process of normalization of terms occurring in the corpus by operations on the syntactic structure by the expert. For example, the terms *docento pedagoginis vardas* 'pedagogical title of docent', *vakariniai kursai* 'evening courses' and *priėmimas į universitetą* 'enrollment in university' appear respectively in the expressions *docento ir profesoriaus pedagoginiai vardai* 'pedagogical titles of docent and Professor', *kursai (dieniniai, vakariniai, tęstiniai, trumpalaikiai ir kt.)* 'courses (full-time, evening, continuing, short)' and *priėmimo į VU* 'enrollment in UV' (or *priėmimo į valstybines aukštąsias mokyklas* 'enrollment in public high schools'). None of these examples could be found by any of the tested extraction methods.

The number of extracted candidate terms has a strong influence on the level of precision. For example, with LICE and the linguistic approach, the number of extracted patterns is very high, with more than 15,000 lemmatized expressions, which has generated mechanically much noise in comparison with the manually extracted reference lists consisting of 1,085 and 3,106 terms. Besides, term candidates have not been rated by both of the methods, which does not allow to extract a meaningful subset of comparable number.

We are aware that results could be further evaluated, taking into account measures for partial matches, however the lack of necessary tools has not allowed us to include in the present paper.

The overall analysis of the results shows that the linguistic approach that extracts term candidates on the basis of morphological patterns has appeared to be quite reliable and most promising according to the measure of recall (61.8% and 65.7%). While in terms of precision, the keyword approach with machine learning has produced better results. Both these methods may be considered as the most perspective, as the linguistic approach could identify a thousand more correct terms than the keyword approach, and the keyword approach has picked up the smallest number of non-terms.

3 Conclusions

We have looked at the term extraction task for Lithuanian from the perspective of existing term extraction tools. Four different methods, i.e., three statistical and one linguistic, have been applied and evaluated against manually constructed reference lists.

The evaluation of term extraction methodology has lead to the following conclusions:

- Most of Lithuanian domain specific terms are two-word or three-word noun phrases.

- The majority of the Lithuanian terms are very rare.

- The best performing methods in terms of recall pick up low precision.

- The statistical modeling for term detection on such a tiny corpus is not very reliable. Thus linguistically based term candidate detection appeared to perform better, i.e. the combined recall and precision levels have been the highest with this

method.

- The analysis has shown that all the methods have problems in extracting both multiword and single-word terms, as well as determining which terms are domain specific and which ones are general. The possible solution would be the combination of several methods for each of these tasks.

- Domain specificity of a term has not been analysed is this paper. A possible approach would be a statistical measure of specificity that expresses the difference of usage of a term between a general corpus and a domain specific corpus.

- An increased quality, i.e. increased precision, may be obtained by improving the linguistic filtering of noisy candidate term lists in order to extract only expressions matching the usual structure of Lithuanian terms.

- A significant improvement of the results may be expected for the machine learning method (KEA) with a more extensive learning process and a processing of each file of the corpus separately.

- Hybrid approaches have not been analysed in this paper, however they may turn out to be very useful in reducing the noise produced by the linguistic approach.

Acknowledgments

The presented research is funded by a grant (No. LIT-2-44) from the Research Council of Lithuania in the framework of the project "Švietimo ir mokslo terminų automatinis identifikavimas – ŠIMTAI 2" (Automatic Identification of Education and Science Terms). The authors would like to thank anonymous reviewers for their comments.

References

Azé J., Roche M, Kodratoff M., and Sebag M. 2005. *Preference Learning in Terminology Extraction: a ROC-based Approach*, Proceedings of Applied Stochastic Models and Data Analysis. p. 209-219.

Cabre T., Estopa R., and Vivaldi J. 2001. *Automatic term detection. Recent advances in computational terminology*. p. 53-88.

Cabre T. 1992. *Terminology: theory, methods and applications*. John Benjamins.

Church K.W., and Hanks P. 1989. *Word Association Norms, Mutual Information and Lexicography*. In Proc: ACL'89, p. 76-83.

Daille B. 1994. *Towards Automatic Extraction of Monolingual and Bilingual Terminology*. In Proc: COLING'94, p. 515-524.

Daille B., Habert B., Jacquemin C., and Royaut J. 1996. *Empirical observation of term variations and principles for their description*. Terminology, 3(2):197-258.

Daudaravičius V., and Marcinkevičienė R. 2005. *Gravity counts for the boundaries of collocations*. Corpus Linguistics, 9(2):321-348.

Dunning T. 1993. *Accurate Methods for the Statistics of Surprise and Coincidence*. Computational Linguistics, 19(1): 61-74.

Frantzi K., and Ananiadou S. 1996. *Extracting Nested Collocations*. In Proc: COLING'96, p. 41-46.

Lopes L., de Oliveira L. H. M., and Vieira R. 2010. *Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches*. In Proc: 9th Int. Conf. on Computational Processing of the Portuguese Language.

Marcinkevičienė R., and Vitkutė-Adžgauskienė D. 2010. *Developing the Human Language Technology Infrastructure in Lithuania*. In Proc: 4th Int. Conf. Human Language Technologies - The Baltic Perspective. IOS Press.

Medelyan O. 2005. *Automatic Keyphrase Indexing with a Domain-Specific Thesaurus*. Master Thesis. University of Freiburg, Germany.

Mitrofanova O., and Zakharov V. 2009. *Automatic Analysis of Terminology in the Russian Corpus on Corpus Linguistics*. In Proc: 5th Int. Conf. NLP, Corpus Linguistics, Corpus Based Grammar Research.

Pantel P., and Lin D. 2001. *A Statistical Corpus-Based Term Extractor*. In Proc: 14th conf. Advances in Artificial Intelligence, E. Stroulia and S. Matwin (Eds.). Springer-Verlag, London, p. 36-46.

Piskorski J., Homola P., Marciniak M., Mykowiecka A., Przepiorkowski A., and Wolinski M. 2004. *Information Extraction for Polish Using the SProUT Platform*. In Proc. of Intelligent Information Systems 2004. Springer Verlag.

Rimkutė E, and Daudaravičius V. 2007. *Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas*. In Kalbų studijos, 11:30-35.

Sager J. 1990. *A Practical Course in Terminology Processing*. John Benjamins.

Scott M. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Smadja F., McKeown K., and Hatzivassiloglou V. 1996. *Translating collocations for bilingual lexicons: a statistical approach*. Computational Linguistics, 22(1):1-38.

Zeller I. 2005. *Automatinis terminų atpažinimas ir apdorojimas*. PhD thesis, VMU, Lithuania.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955