# A Combined Approach for Automatic Identification of Multi-Word Expressions for Latvian and Lithuanian

Justina Mandravickaitė, Tomas Krilavičius, and Ka Lok Man

***Abstract*—We discuss an experiment on automatic identification of bi-gram multiword expressions (MWE) in parallel Latvian and Lithuanian corpora. Raw corpora, lexical association measures (LAMs) and supervised machine learning (ML) are used due to the scarceness and quality of lexical resources (e.g., POS-tagger, parser) and tools. Combining LAMs with ML works well for other languages, our experiments show that it perform well for Lithuanian and Latvian as well. We analyse and discuss frequency thresholds in terms of potential MWE and LAMs values. Finally, combining LAMs with ML we have achieved 98,8% precision and 57,5% recall for Latvian and 96,9% precision and 61,8% recall for Lithuanian.**

***Index Terms*—hybrid-approach, lexical-association-measures, machine-learning, multi-word-expression.**

## I. Introduction

A MULTI-WORD EXPRESSION (MWE) is a sequence of $\geq 2$ words, which functions as a single unit at linguistic analysis, e.g. syntactic analysis. Identification of MWEs is one of the most challenging problems in NLP [1]. A number of methods are used to identify MWEs, e.g. lexical association measures [2], [3], machine learning [4]–[6] , deep learning [7], hybrid [8]–[11], etc., however, experiments are required to determine whether they can be transferred to Lithuanian and Latvian.

Latvian and Lithuanian languages belong to the Baltic language group and are synthetic languages (favor morphologically complex words), thus simple statistical approaches for identification of MWEs cannot provide satisfactory results, as the morphological richness results in lexical sparseness.

Statistical approaches which treat multiword expressions as a bag of words pay no attention to the variation of MWE components [12]. The relatively free word order in both languages does not improve the situation as well. Moreover, Lithuanian and Latvian lexical resources for complementing or replacing statistical approaches are limited.

Exploration of MWEs flexibility and handling exceptions could make the detection of MWE in Lithuanian easier. But even most of the hybrid methods cannot be implemented in a straightforward manner. Thus possibility of detecting Latvian and Lithuanian MWEs by combining lexical association measures and machine learning could be a right approach in this situation. Machine learning allows various properties of text to be encoded in feature vectors (lexical, morphological, syntactic, semantic, contextual, etc.) associated with output classes, as well as identifying complex non-linear relations. It permits capturing elaborate features in languages with complex morphology.

LAMs compute an association score for each collocation candidate assessing the degree of association between its components. These scores can be used for the extraction of collocation candidates, ranking them, or for classification (setting a threshold and dismissing all collocations below it). However, some association measures are very similar (e.g., Pointwise Mutual Information and Dice identify lexical collocations; T-score and Loglikelihood show grammatical collocations [13]).

Due to diversity of collocations different LAMs are good indicators to their detection. For example, for collocations where components statistically occur more often than incidentally,*Log-likelihood ratio*, $x^2$ *test*, *Odds ratio*, *Jaccard*, *Pointwise mutual information* perform better, while for collocations which occur in the different contexts than their components (non-compositionality principle) *J-S divergence*, *K-L divergence*, *Skew divergence, Cosine similarity* were suggested [14]. For discontinuous MWE (where other words occur among the components of MWE), *Left context entropy* and *Right context entropy* show better results [14].

Combining association measures helps in the collocation extraction task [15], [9], [16]. Improvement of the extraction procedure can be achieved by combining a relatively small number of measures. And so far there is no universal combination of association measures that works best in every situation, since the task of collocation extraction depends on the data, language and type/notion of MWEs.

Combination of lexical association measures (LAMs) and supervised machine learning algorithms was investigated by several authors, e.g. [8] used such approach for the extraction and evaluation of MWEs from the English part of Europarl Parallel Corpus, extracted from the proceedings of the European Parliament; extraction of nominal MWEs by application of the same method and from the French part of the same Europarl corpus is reported by [17]. Best combinations of LAMs are

Manuscript received October 10, 2017. This research was partly funded by a grant (No. LIP-027/2016) from the Research Council of Lithuania.

J. Mandravickaitė is with Baltic Institute of Advanced Technology and Vilnius University, Lithuania (corresponding author, e-mail: justina@bpti.lt).

T. Krilavičius is with Baltic Institute of Advanced Technology and Vytautas Magnus University, Lithuania (e-mail: t.krilavicius@bpti.lt).

K. L. Man is with Xi'an Jiaotong-Liverpool University, China (e-mail: kalok2006@gmail.com).

extensively reported in [15], [9], [16], [14].

## II. Methodology

We use lexical association measures (LAMs) combined with supervised machine learning algorithms in this investigation. The first part of the experiment (getting values of LAMs) was executed with MWEToolkit[1] [18] and for the second one (application of machine learning algorithms for MWEs candidates with LAMs values) we use WEKA [2] [19].

Firstly, using MWEToolkit, the candidate MWE bi-grams were extracted from the raw text. Then, values of 5 association measures (*Maximum Likelihood Estimation*, i.e. *relative frequency (mle)*, *Dice's coefficient (dice)*, *Pointwise Mutual Information (pmi)*, *Student's t score (t)* and *Log-likelihood score (ll)*) [18] were calculated. Afterwards, preliminary results were evaluated against the reference lists of bi-gram MWE for each language. The aforementioned reference lists were based on EuroVoc - Multilingual Thesaurus of the European Union [3].

In the following step, preliminary results were evaluated against the reference list of bi-gram MWE (converted to ARFF file with the values of **True** (MWE) and **False** (not MWE)) using WEKA. Selected algorithms (*Naïve Bayes* [20], *OneR* (rule-based classifier; [21]), and *Random Forest* [22]) were applied for automatic identification of MWEs. As the data was rather sparse we separately used two filters: *SMOTE* (it re-samples a dataset by applying the *Synthetic Minority Oversampling TEchnique*) [23] and *Resample* (it produces a random subsample of a dataset using either sampling with replacement or without replacement) [19].

The evaluation of classification results were based on standard measures - *Precision*, *Recall* and *F-measure*. As in [24], [25], Precision is the proportion of items, predicted by supervised machine learning algorithm, which are relevant to the query; Recall is proportion of items, predicted by supervised machine learning algorithm, which are relevant to the query and are predicted successfully. F-measure can be defined as the average of Precision and Recall when they are close, and in general it is the square of the geometric mean divided by the arithmetic mean in terms of the aforementioned Precision and Recall [24].

We have chosen Latvian and Lithuanian parts of JRC-Acquis Multilingual Parallel Corpus 4 [26]. It contains the total body of European Union law applicable to its member states. Currently it includes selected texts written since 1950s. Statistics for Latvian (LV) and Lithuanian (LT) parts of JRC-Acquis Multilingual Parallel Corpus are presented in Table I

We use 1/3 of each, Latvian and Lithuanian, parts of JRC-Acquis Multilingual Parallel Corpus, i.e. 9 million words for LV and LT each.

Our purpose was to get the best possible results without relying on the special linguistic tools, e.g. POS tagger, parser, i.e. to remain at least partially language agnostic. Thus preprocessing consisted of tokenising (one sentence per line) and lowercasing only.

TABLE I
Latvian (LV) and Lithuanian (LT) part of JRC-ACQUIS multilingual parallel corpus

| Characters | Words | Language |
|---|---|---|
| 196 452 051 | 27 594 514 | LT |
| 199 438 258 | 26 967 773 | LV |

As there are no known gold standards for MWE evaluation resources for Latvian and Lithuanian at the moment, we use EuroVoc, a Multilingual Thesaurus of the European Union, for evaluation of MWE candidates with calculated LAMs, extracted with MWEToolkit. We selected bi-grams only, as statistical methods were generally reported to be more successful with shorter n-grams [27]. We use separate lists (one for Latvian, one for Lithuanian) of these bi-gram MWEs for evaluation of MWE candidates with calculated LAMs values, converted to ARFF format (WEKA), where, beside numerical values of LAMs, logical values, showing, whether record is True (MWE) and False (not MWE), are included. Latvian reference list consists of 3608 bi-gram terms, while Lithuanian reference list has 3783 bi-gram items. Number of bigrams is different, because MWEs in Lithuanian/Latvian not always had their equivalents as bi-grams in other language and vice versa, e.g. coal - akmens anglys (Lithuanian), akmeņogles (Latvian); pasture fattening - ganomasis gyvulių penėjimas (Lithuanian), nobarošana ganībās (Latvian)

## III. Experimental Setup

We use LAMs combined with supervised machine learning. LAMs are calculated using MWEToolkit [18], and WEKA [19] is used to train selected classifiers on LAMs.

In this paper we disccuss experiments with bi-gram MWEs only, but we plan to extended definitions of LAMs to 3- and 4-grams, which is not always straighforward, and explore LAMs+ML approach for longer MWEs in future research.

Candidate MWE bi-grams were extracted from the raw text with MWEToolkit: frequencies of separate words and bi-grams are counted, *hapaxes*[4] are removed (or more thorough filtering by frequencies is performed), and values of 5 association measures (Maximum Likelihood Estimation, Dice's coefficient, Pointwise Mutual Information, Student's t score and Log-likelihood score) [18] are calculated. For each language, the results were evaluated against the reference lists, based on EuroVoc - Multilingual Thesaurus of the European Union.

The results were evaluated against the reference list of bi-gram MWE (converted to ARFF file with the values of **True** (MWE) and **False** (not MWE)) using WEKA. Selected algorithms (Naïve Bayes [20], OneR (rule-based classifier; [21]), Bayesian Network [28] and Random Forest [22]) were applied for automatic identification of MWEs. Feature vectors were constructed

[1] MWEToolkit http://mwetoolkit.sourceforge.net

[2] WEKA, http://www.cs.waikato.ac.nz/ml/weka/

[3] EuroVoc, the EU's multi-lingual thesaurus, http://eurovoc.europa.eu/

[4] Bi-grams that occurred in the corpus only once.

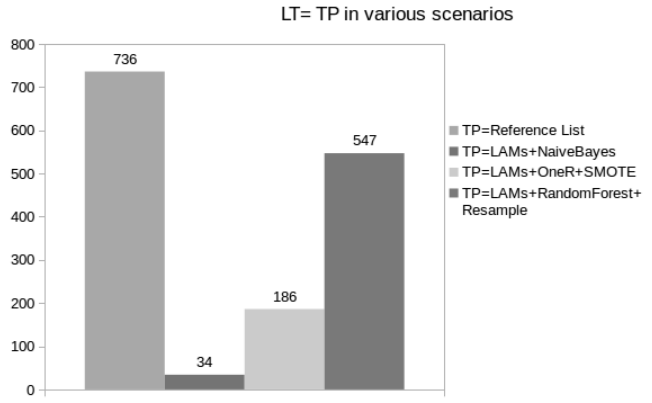


Fig. 1. Lithuanian TP in various scenarios

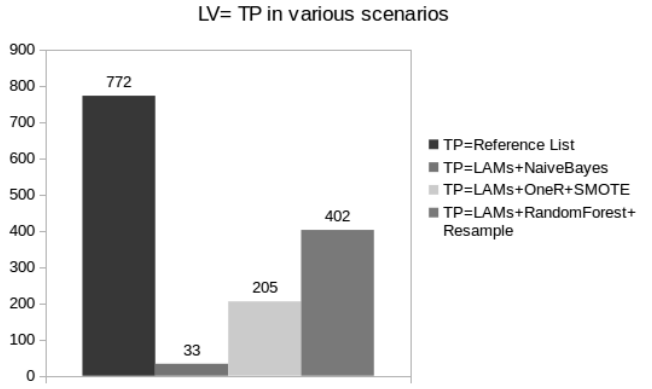


Fig. 2. Latvian TP in various scenarios

from LAMs values for each MWE candidate and its appearance in reference list (True/False).

SMOTE and Resample filters were used to deal with data sparseness. SMOTE re-samples a dataset by applying the Synthetic Minority Oversampling TEchnique) [23]. Resample produces a random subsample of a dataset using either sampling with or without replacement [19].

To evaluate performance we employ

1) precision $P = \frac{tp}{tp+fp}$,
2) recall $R = \frac{tp}{tp+fn}$ and
3) F-score $F_1 = 2 \cdot \frac{P \cdot R}{P+R}$,

, where $tp, fp$ and $fn$ are true positives (correctly identified MWEs), false positives (expressions incorrectly identified as MWEs) and false negatives (incorrectly identified as non-MWEs), correspondingly [24], [25].

Association measures and supervised machine learning algorithms were combined in 3 ways:

1) without any filter,
2) with the SMOTE filter and
3) with the Resample filter.

All the models were tested using standard 10-fold cross-validation.

## IV. Results

We performed experiments with 736 MWE present in the corpus from the reference list for Lithuanian, that is, 736 true positives (TP). For Latvian there were 772 compounds present in the corpus from reference list, i.e. we had 772 MWEs. For TP in different scenarios, see Figure 1 and Figure 2 Summary of experimental results performed in different scenarios (LAMs only, LAMs combined with a supervised machine learning algorithm, LAMs combined with a supervised machine learning algorithm and one of the filters – SMOTE or Resample) are presented in Table XIX

### A. Results of lexical association measures

LAMs used in this paper can be calculated via contingency table. Each observed frequency in a contingency table is marked as a numeric value, $o_{ij}$ , where $i$ and $j$ represent the presence or absence of each component

TABLE II
Contingency table of observed frequencies

| | **b** | **not b** | |
|---|---|---|---|
| **a** | $o_{11}$ | $o_{12}$ | $o_{1p}$ **(R1)** |
| **not a** | $o_{21}$ | $o_{22}$ | $o_{2p}$ **(R1)** |
| | $o_{p1}$ **(C1)** | $o_{2p}$ **(C2)** | $o_{pp}$ **(N)** |

of the n-gram. The marginal frequencies are the sums of each line and are sometimes marked as R1, R2, C1 and C2 respectively. The sum of the marginal frequencies presents the total number of n-grams and is denoted as N. For more details, see Table II. Each contingency table of observed frequencies has a table of expected frequencies, assuming that there is no association between the components of the given n-gram (Table III).

For our experiments we used 5 LAMs implemented in *mwetoolkit*. As mle is basically relative frequency, we provide equations for the remaining 4 LAMs [13]:

$$pmi = log\left(\frac{o_{11}}{e_{11}}\right) \quad (1)$$

$$dice = \frac{2o_{11}}{R1 + C1} \quad (2)$$

$$t - score = \frac{o_{11} - e_{11}}{\sqrt{o_{11}}} \quad (3)$$

$$ll = 2 * \sum_{ij} o_{ij} log\left(\frac{o_{ij}}{e_{ij}}\right) \quad (4)$$

1) **Pointwise Mutual Information (pmi)** ranks n-grams by comparison of the frequency of the MWE candidate to the frequency of the components of the MWE [3]. According to [2], this measure is biased towards low-frequency n-grams.
2) **Dice coefficient (dice)** takes into consideration the frequency of n-gram components occurring together and their individual frequencies [3].
3) **T-score (t)** determines whether the association between two words is non-random [3]. According to [2], this measure produces conservative values.
4) **Log-likelihood ratio (ll)** measures the difference between the observed values and the expected values [3]. It has been reported by [2] for it to perform better for lexical word (or content word, as opposing to function word) collocations.

TABLE III
CONTINGENCY TABLE OF EXPECTED FREQUENCIES

| | **b** | **not b** |
|---|---|---|
| **a** | $e_{11} = \frac{R1C1}{N}$ | $e_{12} = \frac{R1C2}{N}$ |
| **not a** | $e_{21} = \frac{R2C1}{N}$ | $e_{22} \frac{R2C2}{N}$ |

5) **Maximum Likelihood Estimation (mle)** is basically relative frequency. According to [2], frequencies can be used as a "baseline" measure.

For **Lithuanian**, bi-grams with the highest mle values provided mainly co-occurrences (see Table IX; the exceptions are the first and last bigrams). T-score appeared to favour lower frequency bi-grams. 4 of the 10 bi-grams with the highest t-score values were part of longer n-grams, 4 were potential MWE and 2 were co-occurrences. The highest values of pmi belonged mostly to potential MWE with exception of 3 co-occurrences. Dice mostly identified named entities, in our case - terms in Latin and Lithuanian. Finally, ll gave the highest values mainly to potential MWE, however, 1 co-occurence and 1 bi-gram that was a part of longer MWE were present as well. For more details, see Tables IX-XIII.

For **Latvian**, the highest 2 values of mle were assigned to co-occurences, however, the rest of bi-grams were potential MWE. T-score favoured low frequency bi-grams, although there were more potential MWE among the items with the highest values than in Lithuanian case. The highest values of pmi were given to grammatical MWE (3 bi-grams), co-occurrences (5) and 2 MWE. Dice provided the highest values mostly for proper names and terms in Latin, 1 MWE in Slovak (possibly due to error in corpus development) and 2 MWE. Finally, ll appeared to favour higher frequency bi-grams and the majority of the highest values were given to potential MWE, although 4 co-occurrences were present as well. For more details, see Tables XIV-XVIII.

### *B. Frequency Analysis*

Frequencies of MWE candidates have a significant impact on LAMs values (as they are measures of "attraction" between MWE components based on their statistical properties), hence we perform a brief analysis based on MWE candidate frequencies.

LITHUANIAN MWE candidates frequencies range from 2 (it is generally a good idea to filter out candidates with frequency of 1 in order to obtain more reliable association measures) to 31311. There are 159085 candidates with frequency of 2 (or DIS-LEGOMENA). The highest frequency (31311) has 1 MWE candidate. Thus average frequency of Lithuanian MWE candidates is 8.59. For more information on frequencies of Lithuanian MWE candidates, see Table IV.

Frequencies of True positives (TP) or correctly identified MWE range from 2 to 6667. Thus average frequency of Lithuanian TP is 33. Also, there are 136 DIS-LEGOMENA and 1 TP with the highest frequency (6667). For TP with the highest frequencies, see Table VI.

As for classification errors, the experiment with the best results (see Section IV-C) have 58 false negatives (FN; expressions incorrectly identified as non-MWEs). Their frequencies range from 20 (the lower frequency bound of MWE candidates used in the latter experiment) to 1307. Also, there are 3 false positives (FP; expressions incorrectly identified as MWEs) and 1 of them have the lower frequency bound of 20 while other 2 are more frequent.

TABLE IV
FREQUENCY RANGES OF LITHUANIAN MWE CANDIDATES

| **Frequency range** | **Number of entries** |
|---|---|
| 501-31311 | 388 |
| 6-500 | 82266 |
| 2-5 | 274076 |

TABLE V
FREQUENCY RANGES OF LATVIAN MWE CANDIDATES

| **Frequency range** | **Number of entries** |
|---|---|
| 48142-497 | 708 |
| 496-6 | 115332 |
| 5-2 | 322614 |

LATVIAN MWE candidates frequencies range from 2 to 48142. There are 180272 candidates with frequency of 2 (or DIS-LEGOMENA). The highest frequency (48142) has 1 MWE candidate. Thus average frequency of Latvian MWE candidates is 10,69 For more elaborated information on frequencies of Latvian MWE candidates, see Table V.

Frequencies of True positives (TP) or correctly identified MWE range from 2 to 5921. Thus average frequency of Latvian TP is 27. Also, there are 152 DIS-LEGOMENA and 1 TP with the highest frequency (5921). For TP with the highest frequencies, see Table VII.

The experiments with the lowest classification errors (see IV-C) have 189 false negatives (FN; expressions incorrectly identified as non-MWEs). Their frequencies range from 5 (the lower frequency bound of MWE candidates used in the latter experiment) to 528. Also, there are 3 false positives (FP; expressions incorrectly identified as MWEs) and 1 of them has the frequency of 7 (slighly higher then the lower bound of 5) while other 2 are more frequent.

### *C. Classification Results*

We experimented with 736 (LT) and 772 (LV) MWEs present in the corresponding corpus from the reference. See Table XIX for summary of experimental results (LAMs only, LAMs combined with a supervised machine learning and a filter).

Reference list is based on EuroVoc which mostly contained the EU institutions related terms, hence MWEs mostly fit into 3 categories: Noun + Noun, Adjective + Noun and Abbreviation or Acronym + Noun. However, as we did not use either POS tagger or parser (see the beginning of the paper), detailed morpho-syntactic analysis is in our future plans.

Using only the lexical association measures implemented in the MWEToolkit combined with the reference list for evaluation gave low results. Best Recall was 21.4%

TABLE VI
LITHUANIAN TP WITH THE HIGHEST FREQUENCIES

| TP | Frequency |
|---|---|
| europos bendrijos (european communities) | 6667 |
| balsų dauguma (majority of votes) | 1520 |
| transporto priemonės (vehicles) | 1307 |
| jungtinis komitetas (joint commitee) | 685 |
| transporto priemonė (vehicle") | 543 |
| europos parlamentas (european parliament) | 526 |
| europos bendrija (european community") | 408 |
| jungtinė karalystė (united kingdom) | 399 |
| juridinis asmuo (legal person) | 394 |
| europos sąjunga (european union) | 392 |

TABLE VII
LATVIAN TP WITH THE HIGHEST FREQUENCIES

| TP | Frequency |
|---|---|
| eiropas kopienas (European communities) | 5921 |
| apvienotā komiteja (joint committee) | 717 |
| eiropas parlaments (European parliament) | 528 |
| apvienotā karaliste (United Kingdom) | 402 |
| eiropas kopiena (European community) | 369 |
| darbības joma (scope) | 352 |
| valsts iestādes (state institution) | 349 |
| eiropas savienība (European Union) | 347 |
| juridiska persona (legal person) | 346 |
| intervences aģentūra (intervention agency) | 340 |

TABLE VIII
THRESHOLDS OF LAMS VALUES FOR LATVIAN AND LITHUANIAN

| Lithuanian | Latvian | LAMs |
|---|---|---|
| 0.123 | 0.018 | dice |
| 336,774 | 126,549 | ll |
| 0.000 | 0.000 | mle |
| 7.742 | 7.632 | pmi |
| 11.179 | 7,452 | t |

for Latvian and 19.5% - for Lithuanian. Best Precision was 0.5% for Latvian and 0.8% for Lithuanian. Finally, best F-measure was 0.8% and 1.3% for Latvian and Lithuanian respectively. These results were observed after several gradual frequency filtering, setting collocation boundaries via LAMs value curves (see Table VIII) and adjustments in terms of range of candidate MWEs. Out of all 5 LAMs, relative frequency or mle measure proved to be nearly useless in our case. Thus in LAMs scenario it seems that almost any candidate MWE out of the 436 498 (Latvian) and 356 730 (Lithuanian) was identified as an MWE. Thus, association measures did not suffice for the successful extraction of MWEs for Latvian and Lithuanian in our case.

Association measures and supervised machine learning algorithms were combined in 3 ways: (i) without any filter, (ii) with the SMOTE filter and (iii) with the Resample filter. All the models were tested using standard 10-fold cross-validation.

To explore how the results change by using different lower bounds of MWE candidates frequencies, for both languages we performed experiments in 3 flavours of candidate frequencies: 1) MWE candidates with frequencies $\geq 2$, 2) MWE candidates with frequencies $\geq 5$ and 3) MWE candidates with frequencies $\geq 20$.

We report only the best obtained results. For more details, see Table XIX. The best configuration for both languages in our case is Random Forest classifier combined with the Resample filter. As for different lower bound of MWE candidate frequencies, more detailed filtering show better results for Lithuanian than Latvian. In the latter case MWE candidates with frequencies $\geq 5$ seem to be optimal.

The best results for LATVIAN are P=98.8%, R=57,4% and $F_1 = 72.6\%$. They were achieved by classifying MWE candidates with $\geq 5$ frequencies and configuration Random Forest + Resample. However, Recall was slightly better when MWE candidates with $\geq 5$ frequencies were classified (R=57.5). More thorough filtering by frequencies gave lower results in this case.

For LITHUANIAN the best results are P=96.9%, R=61.8% and $F_1 = 75.5\%$. They were achieved by classifying MWE candidates with $\geq 20$ frequencies and configuration Random Forest + Resample. Contrary to Latvian part, results for Lithuanian continued to improve with more filtering by frequencies.

Thus results show that combining LAMs with supervised ML improves extraction of MWEs for both languages.

## V. CONCLUSION

We report our experiment for extraction of MWEs, that is, bi-gram terms for Latvian and Lithuanian. Because of the lack of lexical resources and availability or accuracy of special lexical tools (e.g. POS-tagger, parser), we used raw corpora and combination of lexical association measures and supervised machine learning. This experimental setup improved our results in comparison with using association measures only.

Combining lexical association measures and supervised machine learning, the best experimental setup for languages (Latvian and Lithuanian) consisted of all 5 lexical associatoin measures (Maximum Likelihood Estimation, T-score, Pointwise Mutual Information, Dice Coefficient and Log-Likelihood Ratio), Random Forest algorithm and Resample filter. Also, filtering out low frequency ($< 5$ for Latvian and $< 20$ for Lithuanian) MWE candidates led to improvement of the results.

Our future plans include experiments for automatic extraction of different types of MWEs for Latvian Lithuanian and a greater diversity of MWEs.

## APPENDIX

TABLE IX
LITHUANIAN: 10 BI-GRAMS WITH THE HIGHEST MLE VALUES

| MWE candidate | Frequency | mle |
|---|---|---|
| ūkio produktais (agricultural products) | 95 | 9.9870E-06 |
| tai ji (that is it/her) | 95 | 9.9870E-06 |
| tų produktų ([of] these products) | 95 | 9.9870E-06 |
| praneša viena (announces one/single [smth.]) | 95 | 9.9870E-06 |
| šios medžiagos (these materials) | 95 | 9.9870E-06 |
| kurią turi (which [she/he] has) | 95 | 9.9870E-06 |
| reikia atlikti (need to be done) | 95 | 9.9870E-06 |
| kaip antai (such as) | 95 | 9.9870E-06 |
| kelių valstybių (several states) | 95 | 9.9870E-06 |
| žmonių sveikatos ([of] people health) | 95 | 9.9870E-06 |

TABLE X
LITHUANIAN: 10 BI-GRAMS WITH THE HIGHEST T-SCORE VALUES

| MWE candidate | Frequency | t-score |
|---|---|---|
| nustatančiame išsamias (setting detailed [smth.]) | 62 | 9.9999 |
| viena dvyliktoji (one twelfth) | 3 | 9.99974 |
| pateiktos bendruose (presented in common [smth.]) | 3 | 9.9997 |
| regioninę plėtrą (regional development) | 3 | 9.9997 |
| sudaromos sutartys (agreements are concluded) | 14 | 9.9996 |
| atitinkantys regionai (coresponding regions) | 3 | 9.9996 |
| tarifų lengvatos (rate exemptions) | 29 | 9.9995 |
| chemijos produktą (chemical product) | 6 | 9.9995 |
| vartojami gyvuliams (used for livestock) | 2 | 9.99959 |
| šiuose susitarimuose ([in] these agreements) | 11 | 9.9995 |

TABLE XI
LITHUANIAN: 10 BI-GRAMS WITH THE HIGHEST PMI VALUES

| MWE candidate | Frequency | pmi |
|---|---|---|
| turėtų būti (should be) | 10099 | 99.8464 |
| taip pat (also) | 8575 | 92.3750 |
| laukinių medžiojamųjų ([of] wild game) | 100 | 9.9995 |
| laisvai cirkuliuoti (freely circulate) | 100 | 9.9994 |
| misijos vadovas (head of mission) | 100 | 9.9988 |
| įgaliotasis atstovas (authorized representative) | 100 | 9.9975 |
| trečioje įtraukoje ([in] the third indent) | 100 | 9.9968 |
| šiam sprendimui ([for] this decision) | 100 | 9.9965 |
| taikoma direktyva (applicable directive) | 115 | 9.9964 |
| eeb nuostatas (provisions of the eec) | 177 | 9.9953 |

TABLE XII
LITHUANIAN: 10 BI-GRAMS WITH THE HIGHEST DICE VALUES

| MWE candidate | Frequency | dice |
|---|---|---|
| vertybiniams popieriams ([for] securities/stock) | 73 | 1 |
| būčiau dėkingas ([I] would be grateful) | 52 | 1 |
| glycine max (glycine max (soybean)) | 19 | 1 |
| xanthomonas campestris (xanthomonas campestris) | 19 | 1 |
| erwinia amylovora (erwinia amylovora (fireblight)) | 19 | 1 |
| objektinio stiklelio (object lense (optics)) | 18 | 1 |
| vezikulinio stomatito (vesicular stomatitis) | 18 | 1 |
| guignardia citricarpa (guignardia citricarpa) | 17 | 1 |
| citricarpa kiely ([guignardia] citricarpa kiely) | 17 | 1 |
| oficialiajam leidiniui ([for] official publication) | 16 | 1 |

TABLE XIII
LITHUANIAN: 10 BI-GRAMS WITH THE HIGHEST LL VALUES

| MWE candidate | Frequency | ll |
|---|---|---|
| portugalijos stojimo (Portuguese accession) | 166 | 997.2996 |
| privalo turėti (must have) | 221 | 994.9513 |
| du kartus (twice) | 147 | 992.4616 |
| perduoda komisijai (hand over to the commission) | 259 | 991.7834 |
| tokie standartai (such standards) | 27 | 99.9950 |
| atitinkamas sankcijas (appropriate sanctions) | 20 | 99.9871 |
| turėjimas ketinant (having intent on) | 11 | 99.9770 |
| išvardytos prekės (listed goods) | 29 | 99.9736 |
| leidimo turėtojo ([of] permit holder) | 20 | 99.9313 |
| kinijoje išaugintų (produced in China) | 13 | 99.9267 |

TABLE XIV
LATVIAN: 10 BI-GRAMS WITH THE HIGHEST MLE VALUES

| MWE candidate | Frequency | mle |
|---|---|---|
| šo nolīgumu (this agreement) | 982 | 9.9911E-05 |
| var izmantot (can be used) | 982 | 9.9911E-05 |
| saimnieciskās darbības (economic activities) | 98 | 9.9708E-06 |
| vielas iekļaušanu (inclusion of the substance) | 98 | 9.9708E-06 |
| noteiktā procedūra (the prescribed procedure) | 98 | 9.9708E-06 |
| tirgu kopīgo (common market) | 98 | 9.9708E-06 |
| ražošanas jaudu (production capacity) | 98 | 9.9708E-06 |
| civilās aviācijas (civil aviation) | 98 | 9.9708E-06 |
| mutandis piemēro (applied as necessary) | 98 | 9.9708E-06 |
| lai konstatētu (to find out) | 98 | 9.9708E-06 |

TABLE XV
LATVIAN: 10 BI-GRAMS WITH THE HIGHEST T-SCORE VALUES

| MWE candidate | Frequency | t |
|---|---|---|
| pastāvīgs uzņēmums (permanent residence) | 4 | 9.9999 |
| metāla priekšmetu (metal item) | 2 | 9.9999 |
| uzlikt peļņai (make a profit) | 2 | 9.9999 |
| sākotnējā emisija (initial emission) | 5 | 9.9998 |
| atzinuma iegūšanas (obtaining an opinion) | 3 | 9.9998 |
| arodbiedrību pārstāvji (trade union representatives) | 3 | 9.9998 |
| atliekvielu maksimālos (maximum residue levels) | 6 | 9.9997 |
| piegādātāja rakstiska (supplier [is] written) | 4 | 9.9997 |
| tajā griezusies (approached him [for smth.]) | 2 | 9.9997 |
| posmā nepieņēma (the stage was not accepted) | 2 | 9.9995 |

TABLE XVI
LATVIAN: 10 BI-GRAMS WITH THE HIGHEST PMI VALUES

| MWE candidate | Frequency | pmi |
|---|---|---|
| kā arī (as well) | 9064 | 93.4401 |
| ar ko (with what) | 9793 | 93.1338 |
| vērā eiropas (into europe) | 8896 | 92.9436 |
| attiecas uz ([it] refers to) | 8691 | 91.7347 |
| un jo (and so on) | 9030 | 90.5741 |
| amerikas savienotajās (united america) | 100 | 9.9990 |
| produktus no (products from) | 134 | 9.9987 |
| pa jūru (by the sea) | 100 | 9.9969 |
| pasākumiem pret (measures against [smth.]) | 104 | 9.9957 |
| pārbaužu organizēšanu (organization of inspections) | 100 | 9.9957 |

TABLE XVII
LATVIAN: 10 BI-GRAMS WITH THE HIGHEST DICE VALUES

| MWE candidate | Frequency | dice |
|---|---|---|
| just them (just them) | 85 | 1 |
| kaustiskā kalcinētā (caustic soda) | 48 | 1 |
| christiane scrivener (christiane scrivener (proper name)) | 29 | 1 |
| glycine max (glycine max (soybean)) | 19 | 1 |
| frans andriessen (frans andriessen (proper name)) | 19 | 1 |
| erwinia amylovora (erwinia amylovora (fireblight)) | 19 | 1 |
| guignardia citricarpa (guignardia citricarpa) | 17 | 1 |
| citricarpa kiely ([guignardia] citricarpa kiely) | 17 | 1 |
| európske spoločenstvo (european community (foreign - Slovak)) | 16 | 1 |
| hematopoētisko nekrozi (hematopoietic necrosis) | 16 | 1 |

TABLE XVIII
LATVIAN: 10 BI-GRAMS WITH THE HIGHEST LL VALUES

| MWE candidate | Frequency | ll |
|---|---|---|
| lēmumu nr. (decision no.) | 683 | 999.9924 |
| ar protokolu (by protocol) | 484 | 999.3219 |
| pirmo ievilkumu (first indent) | 193 | 998.9650 |
| derīguma termiņu (expiration date) | 174 | 998.5713 |
| starpību starp (the difference between [smth.]) | 212 | 997.1847 |
| šajā gadījumā (in this case) | 469 | 996.6280 |
| izņēmuma kārtā (exceptionally) | 148 | 996.3629 |
| padomei ziņojumu (council report) | 225 | 996.2061 |
| ko ievieš (implemented by [somebody/smth.]) | 326 | 995.7639 |
| valsts tipa (country type) | 343 | 995.5439 |

TABLE XIX
SUMMARY OF THE RESULTS FOR LATVIAN AND LITHUANIAN WITH DIFFERENT LOWER BOUNDS OF MWE CANDIDATE FREQUENCIES

| Scenario | Precision | Recall | F-meas. |
|---|---|---|---|
| **Latvian** | | | |
| LAMs (freq. $\geq$ 2) | 0.2% | **21.4%** | 0.4% |
| LAMs (freq. $\geq$ 5) | 0.3% | 12.3% | 0.6% |
| LAMs (freq. $\geq$ 20) | **0.5%** | 3.9% | **0.8%** |
| LAMs + RandomForest + Resample (freq. $\geq$ 2) | 93.3% | **57.5%** | 71.2% |
| LAMs + RandomForest + Resample (freq. $\geq$ 5) | **98.8%** | 57.4% | **72.6%** |
| LAMs + RandomForest + Resample (freq. $\geq$ 20) | 98.7% | 53.5% | 69.4% |
| **Lithuanian** | | | |
| LAMs (freq. $\geq$ 2) | 0.2% | **19.5%** | 0.4% |
| LAMs (freq. $\geq$ 5) | 0.4% | 11.8% | 0.8% |
| LAMs (cand. freq. $\geq$ 20) | **0.8%** | 4.2% | **1.3%** |
| LAMs + RandomForest + Resample (freq. $\geq$ 2) | 90.0% | 56.5% | 69.4% |
| LAMs + RandomForest + Resample (freq. $\geq$ 5) | 95.8% | 60.8% | 74.3% |
| LAMs + RandomForest + Resample (freq. $\geq$ 20) | **96.9%** | **61.8%** | **75.5%** |

## References


[1] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword expressions: A pain in the neck for nlp," *Computational Linguistics and Intelligent Text Processing*, pp. 189–206, 2002.

[2] S. Evert, "The statistics of word cooccurrences: word pairs and collocations," Ph.D. dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Germany, 2005.

[3] G. I. Lyse and G. Andersen, "Collocations and statistical analysis of n-grams," *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian„ Studies in Corpus Linguistics, John Benjamins Publishing, Amsterdam*, pp. 79–109, 2012.

[4] M. Yazdani, M. Farahmand, and J. Henderson, "Learning semantic composition to detect non-compositionality of multiword expressions." in *EMNLP*, 2015, pp. 1733–1742.

[5] A. Rondon, H. de Medeiros Caseli, and C. Ramisch, "Never-ending multiword expressions learning." in *MWE@ NAACL-HLT*, 2015, pp. 45–53.

[6] M. Lapata and A. Lascarides, "Detecting novel compounds: The role of distributional evidence," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 235–242.

[7] N. Klyueva, A. Doucet, and M. Straka, "Neural networks for multi-word expression detection," *MWE 2017*, p. 60, 2017.

[8] L. Zilio, L. Svoboda, L. H. L. Rossi, and R. M. Feitosa, "Automatic extraction and evaluation of mwe," in *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 2011.

[9] P. Pecina and P. Schlesinger, "Combining association measures for collocation extraction," in *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006, pp. 651–658.

[10] M. Garcia, M. Garcıa-Salido, and M. Alonso-Ramos, "Using bilingual word-embeddings for multilingual collocation extraction," *MWE 2017*, p. 21, 2017.

[11] J. Mandravickaitė, T. Krilavičius, and K. L. Man, "Automatic identification of multi-word expressions for latvian and lithuanian," in *Lecture Notes in Engineering and Computer Science:Proceedings of the International MultiConference of Engineers and Computer Scientists, 15-17 March, 2017, Hong Kong*, vol. 2, 2017, pp. 706–709.

[12] S. Sharoff, "What is at stake: a case study of russian expressions starting with a preposition," in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. Association for Computational Linguistics, 2004, pp. 17–23.

[13] S. Evert and B. Krenn, "Using small random samples for the manual evaluation of statistical association measures," *Computer Speech & Language*, vol. 19, no. 4, pp. 450–466, 2005.

[14] P. Pecina, "A machine learning approach to multiword expression extraction," in *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, vol. 2008, 2008, pp. 54–61.

[15] ——, "Lexical association measures: Collocation extraction," Ph.D. dissertation, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, 2008.

[16] ——, "Lexical association measures and collocation extraction," *Language resources and evaluation*, vol. 44, no. 1-2, pp. 137–158, 2010.

[17] M. Dubremetz and J. Nivre, "Extraction of nominal multiword expressions in french." in *MWE@ EACL*, 2014, pp. 72–76.

[18] C. Ramisch, *Multiword expressions acquisition: A generic and open framework*. Springer, 2014.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[20] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[21] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine learning*, vol. 11, no. 1, pp. 63–90, 1993.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[24] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *International Journal of Machine Learning Technology*, pp. 37–63, 2011.

[25] J. W. Perry, A. Kent, and M. M. Berry, "Machine literature searching x. machine language; factors underlying its design and development," *Journal of the Association for Information Science and Technology*, vol. 6, no. 4, pp. 242–254, 1955.

[26] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, "The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages," *arXiv preprint cs/0609058*, 2006.

[27] S. Bartsch and S. Evert, "Towards a firthian notion of collocation," *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network Internet Lexicography, OPAL–Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache, Mannheim, to appear*, 2014.

[28] J. Su, H. Zhang, C. X. Ling, and S. Matwin, "Discriminative parameter learning for bayesian networks," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1016–1023.