Looking for a Needle in a Haystack: Semi-automatic Creation of a Latvian Multi-word Dictionary from Small Monolingual Corpora

Inguna Skadiņa
University of Latvia, Institute of Mathematics and Computer Science
E-mail: inguna.skadina@lumii.lv

Abstract

Multiword expressions (MWEs) are an indispensable part of almost any dictionary. However, the identification of missing MWEs that have recently appeared in a language is not a simple task. In this paper we describe automated methods for MWE identification in a rather small Latvian text corpora. We propose starting with the application of statistical measures to identify a wide range of MWEs and then applying linguistically motivated filters to clean the list of initially extracted MWE candidates. We show that for morphologically rich languages, such as Latvian, in cases with a small amount of language data better results can be achieved with lemmatized data. We also demonstrate that in the case of a small general domain (balanced) corpus, automatic methods can be used to find good MWE candidates – terminological units, named entities and some lexicalized phrases. However, finding idiomatic expressions in small, general domain corpora is looking for a needle in a haystack: only a larger or more expressive corpus can help in the identification process.

Keywords: multi-word expressions, low resourced languages, collocations, named entities, terminology

1 Introduction

Multi-word expressions (MWEs), often defined as “lexical items that (a) can be decomposed in multiple lexemes and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity” (Baldwin and Kim 2010: 269), are indispensable part of almost every dictionary – general or terminological, monolingual or multi-lingual. Most commonly used MWE categories include idioms, phrasal verbs, multi-word conjunctions and prepositions, multi-word terms and named entities. While idiomatic expressions and other MWE categories that are used in a language for many years are usually fixed in printed and electronic dictionaries, idiomatic expressions, verbal constructions and terms (as well as named entities) that have more recently appeared in a language are usually not included, because manual identification (recognition) of such missing lexical items is a difficult task.

MWEs are frequently seen as a “pain in neck” (Sag et al. 2002: 1), because identification and processing of MWEs is a complicated task for many natural language processing applications. Different methods of how MWEs could be identified and extracted have been researched for several decades. These include statistical, linguistic-based and hybrid approaches (e.g., Ramisch 2015, Constant et al. 2017). Some methods are designed for specific MWE categories, e.g., noun compounds or light-verb constructions, while others try to cover different MWE categories.

The role of MWEs in natural language processing, especially parsing, has been addressed in the recent COST action ParseMe - PARSing and Multi-word Expressions (Savary et al. 2015). One of the outcomes of the PARSEME project is a survey of the state of the art techniques for MWE processing (Constant et al. 2017). This survey aims “to shed light on how MWEs are handled in NLP applications” (Constant et al. 2017: 839), in particular, in parsing and MT tasks. The results show that most research on MWE identification, extraction, annotation and translation addresses widely used languages with large language corpora, while much less work has been done on languages that lack such broad resources.

However, the problem of MWE identification and extraction for the Latvian language is being addressed in a large-scale national research project, “Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian” (Gruzitis et al. 2018). This aims to create multi-layered semantically annotated language resources for Latvian, anchored in widely acknowledged multilingual representations (AMR, PropBank, FrameNet, Universal Dependencies, Grammatical Framework, BabelNet, DBpedia), that are required for the development of natural language understanding and generation applications.

An important role in this set of language resources is assigned to the tools for identification, extraction and annotation of multiword expressions. These tools aim to extract lists of good quality MWE candidates, which, (1) can be delivered as open experimental MWE lexicon for Latvian, and, (2) after manual inspection, will be added to the largest Latvian open lexical database, *tezaurs.lv* (Spektors et al. 2016).

Automated processing of multi-word expressions in Latvian is mostly studied in the context of machine translation. During decades when rule-based machine translation systems were dominant, special MWE dictionaries were created manually or semi-automatically. Such an approach was also chosen by Deksne et al. (2008) for an English-Latvian rule-based machine translation system. The authors proposed using a special, manually created, dictionary of MWEs together with a set of MWE processing rules, and to include additional MWE processing step during parsing. In the era of statistical machine translation (SMT), Pinnis and Skadiņš (2012) investigated a term translation problem for domain specific SMT. Using automated methods, Pinnis (2013) also created a multilingual term dictionary (which includes multi-word terms, too) and demonstrated its importance in statistical machine translation. Finally, Skadiņa (2016) reported improvements in machine translation output when an automatically extracted MWE dictionary is integrated into a domain specific machine translation system.

All these solutions use automatically extracted MWE dictionaries in another natural language processing task, namely machine translation. In the case of statistical machine translation, the dictionary of automatically extracted MWE candidates can contain noise, e.g. parts of MWEs (shorter phrases), widely used phrases that are not terms, or even some frequently used sequences of words. In the case of lexicon building and supplementing, where incorrect phrases create additional work for lexicographers, the quality of extracted MWEs and proportion of incorrect candidates is very important.

In this paper we describe the process and strategies for finding Latvian MWEs using a rather small amount of data. We propose using statistical measures at first and then apply linguistic filters to avoid ungrammatical, but frequent sequences of words. We demonstrate that in the case of a small general domain (balanced) corpus automatic methods can be used to find good MWE candidates – terms, named entities and some lexicalized phrases. However, a rather small balanced corpus is not suitable for the identification of idiomatic expressions.

2 Strategies for MWE Identification and Extraction

The Latvian language is often mentioned among morphologically rich under-resourced languages (e.g., Skadiņa et al. 2012). For morphologically rich languages, MWE identification and extraction usually consists of two steps – at first morpho-syntactic patterns are applied to extract the initial list of MWE candidates, then this list is filtered by means of statistical measures (e.g. Pinnis et al. 2012, Ramisch 2015). The main limitation of this approach is that it convers only MWEs that represent particular linguistic patterns (usually noun phrases, sometimes verb phrases), leaving other MWEs out.

The aim of our work is to support automatic identification of different categories of MWEs that, after manual inspection, could be then added to the Latvian explanatory dictionary *tezaurs.lv.* We thus propose starting with the application of statistical measures. This allows us to identify a wide range of MWE categories, although it could also result in a high amount of ungrammatical constructions. Thus, as the next step, we apply linguistically motivated filters (patterns) to clean the list of initially extracted MWE candidates.

2.1 Data

Three different datasets were used in our experiments: the Balanced Corpus of the Modern Latvian language (Levāne-Petrova, 2012), the Latvian-Lithuanian parallel corpus (Utka et al. 2012) and Open Subtitles corpus (Lison and Tiedemann 2016). Depending on the related experiment these corpora were used as the original raw text corpus, lemmatized corpus or morphologically annotated corpus. Table 1 provides general information about these data sets.

Table 1: Corpora used for experiments

Corpus	Size			
	Sentences (thousands)	Tokens (million)	Unique tokens (thousands)	Unique lemmas (thousands)
Balanced Corpus of the Modern Latvian language	148	5,54	408,01	111,59
Latvian-Lithuanian parallel corpus	223	3,24	307,53	87,88
Open Subtitles corpus	454	2,37	117,01	56,44

The Balanced Corpus of the Modern Latvian language only modern, standard Latvian language texts that were written no more than 20 years ago. The corpus was collected using the following balancing criteria: 55% periodicals (27% national newspapers, 22% regional newspapers, 14% internet news, 13% special periodicals and 24% popular periodicals), 20% fiction, 10% scientific publications, 8% legal texts, 5% different other texts and 2% parliamentary transcripts.

The starting point for the Latvian-Lithuanian parallel corpus were texts that are written either in Latvian or Lithuanian and then translated into the other language. However, during the collection process it was discovered that such texts are insufficient to reach the goal of eight million words. Thus, legal texts, usually written in English and then translated into Baltic languages, were also included in the corpus. The resulting parallel corpus contains 19.3% texts that were originally written in Latvian, 39.3% texts that were written in Lithuanian, and 41.4% EU legal texts translated into Baltic languages. Texts originally written in Baltic languages represent the following domains: modern fiction (86%), periodicals (5.9%), popular literature (5.6%). In our experiments only texts that were originally written in Baltic languages were used, as we found noise in the translated legal texts.

2.2 Application of Statistical Measures

Different statistical measures are well known means for extraction of MWEs, especially collocations. Among the widely used measures the most popular are the t-score, mi-score, log-likelihood and Dice score (e.g., Manning and Schütze 1999).

In our experiments we applied different combinations of frequency, mi-score and t-score. The mi-score (mutual information score) measures the strength of association (frequency of co-occurrence vs. separate occurrence). A mi-score of three or higher is usually considered to be significant. However, for low frequency words the mi-score could be misleading, as demonstrated in Table 2, where none of the top 10 MWE candidates is an MWE. The t-score measures the confidence of association and can also be applied for low frequency words, with a t-score of two or higher considered to be statistically significant (Hunston 2002), although it also recognizes frequent word combinations (e.g., in Table 2, *kas ir* (which is)*, tas ir* (it is)).

Our first experiments were performed on the Balanced Corpus of Modern Latvian. The Collocate tool (Barlow 2004) was used for calculation of statistical measures. At first the mi-score and t-score were used individually and the most frequent MWE candidates were investigated. We then applied the t-score as an initial filter, and afterwards sorted the results by mi-score and frequency. The threshold for the t-score was set at 2.5, while for the mi-score it was set at 3. The higher t-score, as recommended by Hunston (2002), was set to avoid unnecessary noise. Word sequences consisting of two to five words were investigated.

Table 2 summarizes the top 10 word sequences extracted from the Balanced Corpus of Modern Latvian with the mi-score, t-score and combination of both. The word sequences that are bold in the table could be considered as good MWE candidates[1] – some of them (mostly short ones) are already included in existing dictionaries, while others could be added after investigation by a lexicographer.

Table 2: Top 10 word sequences extracted with different statistical measures
(bolded MWE candidates could be accepted as MWEs).

identified by mi-score		**identified by t-score**		**identified by t-score, filtered by mi-score**	
ordered by mi-score	ordered by frequency	ordered by t-score	ordered by frequency	ordered by mi-score	ordered by frequency
“(Caune, Rata, Grigule, Sviklis, Ugaine”	**kā arī** (also)	kā arī (also)	kā arī (also)	nolikums” (Latvijas Vēstnesis, 168 (3116), 22.10.2004.	**stājas spēkā** (enter into force)
“Pirts, baseini, vanna, solārijs, sports”	tas ir (it is)	(Ar grozījumiem, kas izdarīti ar (with amendments made by)	tas ir (it is)	nolikums” (Latvijas Vēstnesis, 129 (3705), 10.08.2007.)	**kas stājas spēkā** (that enters into force)
kurējās uguns vilinot knišļus gaiņājot	**stājas spēkā** (enter into force)	**likuma redakcijā, kas stājas spēkā** (the law version that comes into force)	kas ir (it is)	nolikums” (Latvijas Vēstnesis, 76 (3652), 11.05.2007.)	**likumu, kas stājas spēkā** (the low that enters into force)
aizā kurējās uguns vilinot knišļus	to, ka (the fact that)	ne tikai (not only)	**stājas spēkā** (enter into force)	nolikums” (Latvijas Vēstnesis, 124 (3072), 06.08.2004.)	likumu, kas stājas (the low that enters)
“(Pranka, Lāce, Trupovniece, et al.”	ar to (with this)	kas ir (it is)	to, ka (the fact that)	nolikums” (Latvijas Vēstnesis, 70 (2835), 13.05.2003.)	**Ministru kabineta** (Cabinet of Ministers)

[1] In some cases post-processing (removal of delimiters) is necessary

identified by mi-score		**identified by t-score**		**identified by t-score, filtered by mi-score**	
“Lāce, Trupovniece, et al. 2003/”	ne tikai (not only)	tas ir (it is)	ar to (with this)	izdarīti ar 10.06.1998., 25.11.1999., 20.06.2001.,	grozījumiem, kas izdarīti ar (amendments made by)
uguns vilinot knišļus gaiņājot zvērus	par to, (about it)	bet arī (also)	ne tikai (not only)	pensiju shēmas līdzekļu pārvaldītāju reģistrā (register of the pension scheme asset managers)	grozījumiem, kas izdarīti (amendments made)
rullī (2.lasījums. Steidzams) Datums: 09.11.2006.	kas stājas (that enters)	ar to (with this)	par to, (about it)	fondēto pensiju shēmas līdzekļu pārvaldītāja (funded pension scheme asset manager)	kas izdarīti ar (made by)
klusā aizā kurējās uguns vilinot	**kas stājas spēkā** (that enters into force)	to, ka (the fact that)	**kas stājas spēkā** (that enters into force)	fondēto pensiju shēmas līdzekļu pārvaldītājs (funded pension scheme asset manager)	(Ar grozījumiem, kas izdarīti ar (with amendments made by)
ugunsgrēks, zibens spēriens, zādzība, vētra	ir ļoti (is very)	tā ir (it is)	ir ļoti (is very)	fondēto pensiju shēmas līdzekļu pārvaldītāju (funded pension scheme asset manager)	(Ar grozījumiem, kas izdarīti (with amendments made)

The table clearly demonstrates the strengths and weaknesses of each approach: when MWE candidates are ordered by statistical significance then longer word sequences are identified (*likuma redakcijā, kas stājas spēkā – in the form of law which has effect*), while ordering by frequency identifies short, but stable phrases, e.g., multi-word conjunctions (*kā arī - as well as, ne tikai – not only)*.

When MWE candidates were selected and ordered by mi-score the top 10 word sequences were noun phrases or word sequences with a very high (more than 70) mi-score, but none of them was an MWE. In the case of the t-score, both short (*bet arī – but also)* and longer (*ar grozījumiem, kas izdarīti ar - with amendments that has been made with*) MWE candidates are identified. Although the top 10 MWE candidates identified by t-score include several MWEs, more than half of the identified MWE candidates are frequent word sequences or parts of phrases.

Finally, the t-score was applied as the first filter and then the candidate list was filtered by the mi-score. The Top 10 MWE candidates (ordered by mi-score) include four acceptable MWE candidates (others are typical initial phrases of legal documents). All four MWE candidates are complex noun phrases (terms): three are morphological variants (inflected forms) of the phrase ‘*fondēto pensiju shēmas līdzekļu pārvaldītājs*’ *(manager for funded pension scheme assets)* and the fourth is another term – ‘*pensiju shēmas līdzekļu pārvaldītāju reģistrā*’ *(in a register of funded pension scheme managers).* When this MWE candidate list is sorted by frequency, seven of the top 10 MWE candidates can be accepted as MWEs. These MWEs are either verbal constructions (e.g., *stājas spēkā - enter into force*) or nouns followed by a relative clause (e.g., *grozījumiem, kas izdarīti - amendments made*).

These three initial experiments demonstrated that in the case of a small corpus the most promising approach uses a combination of t-score and mi-score.

Latvian is a morphologically rich language, and thus application of statistical measures on a small corpus allows us to find only frequent phrases that in many cases are already in dictionaries (e.g. multi-word conjunctions, *kā arī – as well as*). To delve further and obtain not so trivial (although useful) data, we applied a Latvian lemmatizer (Paikens et al. 2013) and repeated the same set of experiments with lemmatized data. This allowed us to find more MWEs – we found many named entities (people's names and their occupations, as well as the related organization names) and terminological units from different domains. Table 3 shows the top 10 MWE candidates that were identified with a t-score and then filtered with the mi-score. Four named entities and five terms are among top 10 MWE candidates in the MWE candidate list that is ordered by mi-score. When the list is ordered by frequency, three complex function words (*kā arī - also, kaut kas - something, pēc tas - after*) and two frequent MWEs (*pants punkts - article* and already mentioned *stāties spēkā – enter into force*) are included.

Table 3: MWE candidates extracted from the lemmatized corpus (MWE candidates in bold could be accepted as MWEs).

Word sequences with highest mi-score	Most frequent word sequences
Arco Real Estate ‘ ‘ (company name)	kā arī (also)
pārvalde priekšnieks palīdze Linda Zubāne (assistant chief of adminastration Linda Zubāne)	pants punkts (article)
Černobiļa AES avārija sekas likvidēšana (Chernobyl nuclear plant disaster recovery)	, kas būt (which is)
šķirne ‘ Koričnoje Novoje ‘ (named entity)	tas , ka (the fact that)
jaukt dispersija kovariāt analīze iegūt (mixed variance covariance analysis provides)	**kaut kas** (something)
ar akūts katarāli strutot endometrīts (with acute catarrhal stomach endometritis)	tas , kas (that/what ...)
ar hronisks katarāli strutot endometrīts (with chronic catarrhal stomach endometritis)	būt ļoti (to be very)
pārvalde priekšnieks palīdze Ieva Sietniece (assistant chief of adminastration Ieva Sietniece)	viens no (one of)
Valmiera / Rūjiena / Strenči-1 (list of names)	**stāties spēks** (enter into force)
līcis piekraste krasts kāpa aizsargjosla (costal protection zone)	pēc tas (after)

2.3 Filtering MWE Candidates

The identified word sequences that are extracted using statistical measures are not always grammatical, as demonstrated in Tables 2 and 3. Moreover, the list of MWE candidates contains word sequences that are not MWEs (e.g., phrases or word sequences that are frequent in a particular corpus), and thus need to be removed from the list. Therefore, after selection of initial MWE candidates, statistical and morpho-syntactic filters are used for the final selection of MWE candidates.

Statistical filters are used to avoid unnecessary noise that is typical for MWEs with a low confidence score. In the case of the mi-score, we found that a high frequency (and mi-score in a range of four to 11) is a better signal that the string could be an MWE than a high mi-score and low frequency (e.g. below 10). In the case of the t-score – high frequency together with a high t-score is a signal of a good MWE candidate. Finally, if the t-score is used as the initial filter and mi-score is used as the second filter, then: (1) most of the MWE candidates will be frequent and a have mi-score value be between 10 and 35, or, (2) will have a high mi-score and low frequency.

The simple regular expressions and morpho-syntactic filters allow to filter out word sequences that are ungrammatical. Regular expressions are used to filter out sequences of tokens that start or end with a punctuation mark, include parentheses or numbers. For instance the word sequence '*par to ,*' (*about,*) from Table 2 ends with a comma and thus needs to be removed or replaced with '*par to*'. Language specific regular expressions include words that in a specific position makes an MWE candidate ungrammatical, e.g., *un* (*and*), *vai* (*or*) as the last word, or, *būt* (*to be*) at the beginning or end of a string consisting of two words (e.g. *būt ļoti* (*to be very*) in Table 2).

Morpho-syntactic filters are used to filter ungrammatical MWE candidates, as well as to extract specific categories of MWEs, e.g., verbal phrases. The most complicated case is an ungrammatical sequence that contains parts of two or more phrases (e.g., in Table 2: *kurējās uguns vilinot knišļus gaiņājot – fire burned luring flies fight*) or contains only part of the phrase (e.g., in Table 3: *ir ļoti – is very*). In such cases the process of filtering patterns needs to be defined carefully, to avoid situations when good MWEs are removed. For instance, the verbal phase *stājas spēkā* (*comes into force*) could be mistakenly removed, as it contains a verb followed by noun in the locative form.

Finally, in the case of overlapping MWE candidates (e.g. *stājas spēkā (comes into force), stājas spēkā ar (comes into force from)*, or *kas stājas spēkā (which comes into force)*) the choice of the most appropriate MWE needs to be made by a lexicographer.

3 Application and Results

We evaluated our method on three different Latvian language corpora: the Balanced Corpus of Modern Latvian, Latvian-Lithuanian corpus and Open Subtitles corpus. The choice of these corpora was justified by the aim of this research – to provide good MWE candidates for a Latvian explanatory dictionary. Therefore, we excluded well-known domain specific corpora, such as JRC Acquis or EMEA, because the term extraction problem (as a special category of MWEs) has been researched by Pinnis et al. (2012).

3.1 Balanced Corpus of the Modern Latvian Language

The Balanced Corpus of the Modern Latvian language was the starting point and the main resource of our research. This corpus (and its updated versions that are under construction) is the main resource on which other language resources (such as the universal dependency treebank, FrameNet and PropBank for Latvian) are currently created.

Our initial hypothesis was that this corpus is a good source to identify different types of MWEs that occur in Latvian rather frequently. However, as was demonstrated in the previous section, we found that applying simple statistical measures to this corpus allows us to identify good MWE candidates for the legal domain (e.g., *stājas spēkā - comes into force*). The main reason is the rather strict language of legal texts: although legal documents form only 5% of the corpus texts, typical legal domain phrases, that appear again and again, are identified as legal domain terminology entries in our MWE candidate list.

As was demonstrated in the previous section, in the case of a lemmatized corpus our method allows us to find many named entities and terminological units from different domains. Most of the identified MWEs consist of two or three words, and thus in the next experiment we identified strings of words up to three words long – these strings were identified by mi-score or t-score and then filtered by the former. The list of the top 10 MWE candidates is shown in Table 4. When MWEs are identified by mi-score, all the top 10 word sequences are MWEs. However, most of MWE candidates are named entities, the only exception is '*Pīrsons hī kvadrāts*' (*Pearson's chi-square*). In the case when MWEs are identified by t-score, fiver strings are named entities, four are terms and one (JP NVO RV) is a string of characters. When the top 20 MWE candidates were analyzed, eight of them were terms.

Table 4: Top 10 lemmatized MWE candidates selected with t-score and mi-score (MWEs in bold are terms, while others are named entities, except JP NVO RV).

mi-score	t-score
Legacy by Angosturs (named entity)	Arco Real Estate (company name)
Eastgate Properties Limited (company name)	Satja SAI Baba (company name)
Nike Riga Run (named entity – event)	**Pīrsons hī kvadrāts** (Pearson's chi-square)
ģenerāldirektors Jespers Koldings (general director Jesper Kolding)	katarāli strutot endometrīts (catarrulous endometritis)
Arco Real Estate (company name)	JP NVO RV
fon den Brinkena (name)	**amonijs nitrāts slāpeklis** (ammonium nitrate nitrogen)
Satja SAI Baba (company name)	Ge Money Bank (named entity – bank)
Satja Sai Baba(company name)	Parex Asset Management (named entity)
Latvian Art Theory (named entity)	New York Time (named entity)
Pīrsons hī kvadrāts (Pearson's chi-square)	jaukt dispersija kovariāt (mixed variance covariance)

This experiment shows that t-score allows better to identify terms that can be included into electronic dictionary. Therefore the threshold for t-score was raised up to 10: 8 terms, one named entity (LPP/ LC – name of party) and one sequence of words (° *C temperatūra*) was identified between top 10 candidates (Figure 1).

Frequency	**Mi-score**	**MWE candidate**
33	27.988560	ģenētiski modificēt kultūraugi (genetically modified crops)
34	27.120896	ģenētiski modificēt mikroorganisms (genetically modified microorganism)
42	26.717372	noziedzīgs nodarījums izdarīšana (committing criminal offences)
112	26.346762	LPP / LC (name of party)
137	26.146896	ģenētiski modificēt organisms (genetically modified organism)
168	25.801889	fondēta pensija shēma (funded pension schema)
9	25.669791	konkurētspējīga priekšrocība pārnešana (competitive advantage transfer)
8	25.222240	civila aizsardzība aizsargbūve (civil defence protection structure)
37	25.169589	° C temperatūra (° C temperature)
31	25.112838	infekcija slimība izraisītājs (infectious disease agent)

Figure 1: Frequency, mi-score for top 10 MWE candidates identified by t-score>=10.

As the project is organized around the top 2,000 Latvian verbs, in our next experiment, after the application of statistical measures to the lemmatized corpus, we filtered out only MWE candidates that contain a noun and verb in a person form (Figure 2). From the top 10 MWE candidates, seven could be accepted as MWEs: four of these are included in *tezaurs.lv*, while other three are included in the Latvian-English dictionary (Veisbergs 2005). It has to be mentioned that three of the four MWEs that are present at *tezaurs.lv* are formed by a verb in a person form followed by a noun in the locative form.

Frequency	**Mi-score**	**MWE candidate**
2006	44.740771	**stāties V spēks N** (come into force)
623	24.896466	**pieņemt V lēmums N** (to make decision / decide)
290	16.915062	**dot V iespēja N** (to enable)
247	15.360174	**tikt V gals N** (to manage)
218	14.595137	veikt V pētījums N (to do research)
141	11.759475	**sniegt V informācija N** (provide information)
130	11.393867	**pievērst V uzmanība** N (pay attention)
115	10.505124	tiesības N saņemt V (rights to receive)
104	10.187839	**ienākt V prāts N** (to come into one's head)
92	9.581201	aizvērt V acs N (to close eyes)

Figure 2: Frequency and mi-score for the top 10 verbal phrases consisting of verb (V) and noun (N), with the MWEs in bold.

We can conclude that the Balanced Corpus of Modern Latvian is a good source for automatic identification of named entities (people’s names and their occupations, as well as the related organization names) and terminological units from different domains. The increase in the threshold allows us to obtain good terminological entries with high precision. When a specific phrase pattern is considered, the result depends on that particular phrase’s construction and the quality of the pattern. We can also see that the size and balancing criteria of this corpus limits the ability of automatic methods with regard to finding idiomatic expressions.

3.2 Latvian-Lithuanian Corpus

To investigate the applicability of our methods for identification of other types of MWEs, not only named entities and terms, we applied the same strategy to the Latvian part of the Latvian-Lithuanian parallel corpus. Although it is also a rather small corpus, it contains more general domain texts than the Balanced Corpus of Modern Latvian, including modern fiction and news texts. Our hypothesis was that such a corpus could contain more frequently used fixed phrases and idiomatic expressions than the Balanced Corpus of Modern Latvian. However, as shown in Table 5, the results obtained are similar to the previous ones. When MWE candidates are ordered by mi-score, seven MWE candidates are named entities, one is part of a longer phrase (*Ventspils peldbaseina relaksācija – Ventspils swimming pool of relaxation*), one is a fixed phrase (*peldbaseins relaksācija komplekss – complex of swimming pools for relaxation*) and one is a character string (W/m). If MWE candidates are ordered by frequency, then four MWEs are terms (*apkure katls - central heating boiler, apkure iekārta – central heating boiler, sāls istaba – salt room, relaksācijas komplekss – complex of relaxation*), one is a named entity (*Ventspils peldbaseins – Ventspils swimming pool*), one is a complex function word (*ne tikai – not only*), while four other MWE candidates are parts of longer phrases.

Table 5: Top 10 MWE candidates extracted from the Latvian-Lithuanian parallel corpus and ordered by mi-score and frequency (MWEs are in bold).

mi-score	Frequency
SIA “ AD BALTIC “” (company)	**apkure katls** (central heating boiler)
Ventspils peldbaseins relaksācija komplekss	apkure iekārta (heating system)
izstāde “ Tech Industry ” (event)	**Ventspils peldbaseins** (Ventspils swimming pool)
“ Tech Industry ” (event)	koksne granula (wooden pellet)
“ AD BALTIC ” (named entity)	katls iekārta (boiler equipment)
SEALEY POWER products (named entity)	granula apkure (pellet heating)
W / m	informācija par (information about)
peldbaseins relaksācija komplekss (swimming pool relaxation complex)	sāls istaba (salt room)
Ventspils peldbaseins relaksācija (Ventspils swimming pool relaxation)	relaksācija komplekss (relaxation complex)
REN TV Baltija (named entity)	**ne tikai** (not only)

In contrast to the previous experiment, the Latvian-Lithuanian parallel corpus was too small to obtain good terminological units when a higher threshold for the t-score was set. Our hypothesis that we could identify idiomatic expressions in this corpus was thus not supported, and perhaps idiomatic expressions are quite rare in this collection.

3.3 Open Subtitles Corpus

As idiomatic expressions were not found in two previously used corpora, we turned to the last on our list – the Open Subtitles corpus. As in the previous experiments, we used a lemmatized corpus and applied the t-score for identification and mi-score as the second filter. In addition, MWE candidates were filtered by frequency (in the previous experiments a threshold of only five was used). The results of these experiments are summarized in Table 6, and they differ from those of the earlier ones – besides named entities, different idiomatic expressions are also identified.

Similar to the previous experiments, many (four when the frequency is at least five and six if it is at least 10 or 15) of the extracted MWEs are named entities (e.g., *viesnīca “dižena Budapešta “- hotel “great Budapest”, Bārts Šērmens – Bart Shermen, “zēns un ābols” - “boy and an apple* (painting)). However, different idiomatic expressions are identified too (e.g., *dzīvot laimīgi līdz mūžs gals – live happily to the end of his days, gulēt saldi – sleep well, daudz laimes dzimšanas diena – happy birthday; ar tas nebūt nekāds sakars – nothing to do with this,*).

Table 6: List of MWE candidates extracted from the Open Subtitles corpus, with the idiomatic expressions in bold.

Freq>=5	Freq>=10	Freq>=15
it ‘ s not going	it ‘ s not going	Bārts Šērmens (named entity)
viesnīca “ dižena Budapešta ” (hotel “Great Budapest”)	Bārts Šērmens (named entity)	“ Pearson Hardman
dzīvot laimīgi līdz mūžs gals (live happily till end of life)	dzeršana no zābaks (drink from boot)	“ Folsom foods ”
misis boss (named entity)	paskriet , paostīt , sarauties (run, sniff, cringe)	“ SouthJet ” 227
Bārts Šērmens (named entity)	Vašingtona māksla noziegums nodaļa (Washington Arts Crime Division)	“ Delta psi ” (named entity)
Rikijs Pontings (named entity)	laimīgi līdz mūžs gals (happily till end of life)	“ mežonīgs vepris ” (”wild hog” - name of bar)
dzeršana no zābaks (drink from boot)	“ SouthJet ” 227	pakārt viņš (hang him)
paskriet , paostīt , sarauties (run, sniff, cringe)	“ zēns ar ābols ” (“Boy with Apple” – painting)	daudz laime dzimšana diena (happy birthday)
gulēt saldi (sleep well)	“ Wayne enterprise ”	ar tas nebūt nekāds sakars (nothing to do with it)
kosmoss kuģis (spaceship)	“ Pearson Hardman ”	dzeršana no zābaks (drink from boot)

4 Conclusion

In this paper we discussed possible strategies for extraction of MWE candidates from different corpora – a balanced, parallel corpus that contains mainly fiction and a corpus of a specific genre (subtitles). We demonstrated that in case of a small amount of general domain (balanced) data ,automatic methods can be used to find good MWE candidates – terms or named entities. However, finding idiomatic expressions in small, general domain corpora is looking for a needle in a haystack: only a larger or more expressive corpus could help in the identification process.

In the case of a small parallel corpus, the most reliable results are obtained for named entities. Terms and complex function words could be also identified, but in this case more careful manual inspection is necessary. Therefore, our next task is to investigate the possibility of applying an automatically extracted bilingual dictionary as an additional filter to improve the precision of MWE candidates.

If the aim of the MWE identification is to identify idiomatic expressions that have recently appeared in a language, then the corpus needs to represent more everyday language and to be rather large, because idiomatic expressions are rare in balanced corpora that represent literary language and carefully edited texts.

References

Baldwin, T. and Kim, S.N. (2010). Multiword Expressions. In *Handbook of Natural Language Processing*, pp. 267-292.

Barlow. M. (2004). Collocate 1.0: Locating collocations and terminology.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, (2017). A.: Multiword expression processing: a survey. In *Computational Linguistics*, 43(4), pp. 837-892.

Deksne, D., Skadins, R. & Skadina, I. (2008). Dictionary of Multiword Expressions for Translation into Highly Inflected Languages. In: *Proceedings of the International Conference on Language Resources and Evaluation LREC 2008*, pp. 1401-1405.

Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., Paikens, P. (2018). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC),* pp. 4506-4513.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, pp. 923-929.

Manning, C. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing, MIT Press. Cambridge.

Paikens, P., Rituma, L. & Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*.

Levāne-Petrova K. (2012). Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. In *Baltistica VIII priedas*, Vilnius, pp. 89-98.

Pinnis, M. & Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In *Proceedings of the Fifth International Conference Baltic HLT 2012*, pp. 176-184.

Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *Recent Advances in Natural Language Processing,* pp. 562-570.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M. & Gornostay,T. (2012). Term Extraction, Tagging and Mapping Tools for Under-Resourced Languages. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework.* Theory and Applications of Natural Language Processing series XIV, Springer.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: a pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'02*, pp. 1-15.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegaard, G., Parra, C., Waszczuk, J., Constant, M., Osenova, P., Sangati, F. (2015) "PARSEME – PARSing and Multiword Expressions within a European multilingual network". In *Proceedings of the 7th Language & Technology Conference (LTC 2015)*, pp. 27-29.

Skadiņa, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I. & Rudzīte, A. (2012). *Latvian Language in the Digital Age*. Springer.

Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L. Rituma, L. & Saulite, B. (2016). Tezaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC),* pp. 2568-2571.

Utka, A., Levane-Petrova, K., Bielinskiene, A., Kovalevskaite, J., Rimkute, E. and Vevere, D. (2012). Lithuanian-Latvian-Lithuanian parallel corpus. In *Proceedings of the Fifth International Conference Baltic HLT 2012*, pp. 260-264.

Veisbergs, A. (2005). *Jaunā latviešu-angļu vārdnīca*. Zvaigzne ABC.

Acknowledgements

This work was supported by the European Regional Development Fund grant No. 1.1.1.1/16/A/219 "Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian".