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

Inguna Skadiņa

Institute of Mathematics and Computer Science

University of Latvia

Mākslīgā intelekta laboratorija
LU MII

NATIONAL DEVELOPMENT PLAN 2020

EUROPEAN UNION
European Regional Development Fund

INVESTING IN YOUR FUTURE

Multi-word expressions

Multi-word expressions (MWEs) are lexical items that

can be decomposed in **multiple lexemes**

+

display lexical, syntactic, semantic, pragmatic and/or statistical **idiomaticity**

Often called as “pain in neck” (Sag et al. 2002: 1) :

raining cats and dogs
New York
tablet computer
part of speech
go on
make mistake

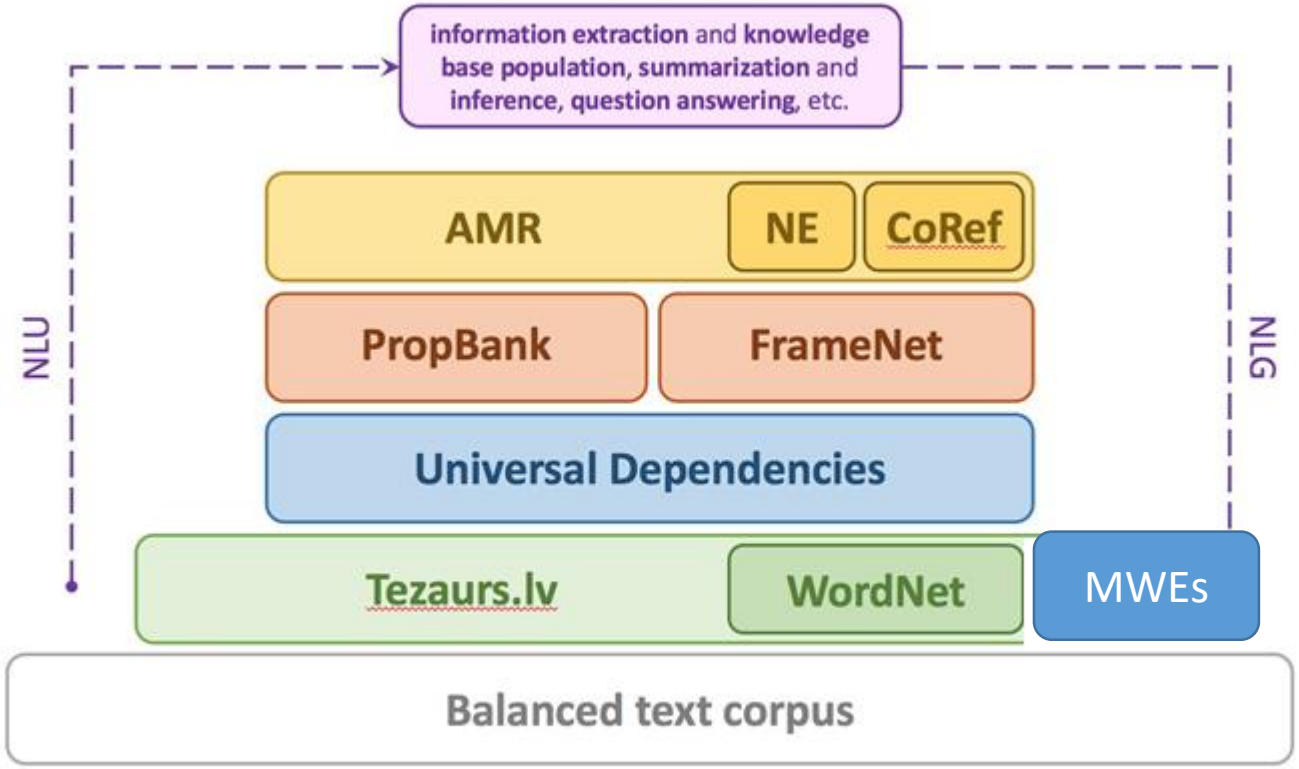
(Baldwin and Kim 2010: 269)

Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian

The project aims to create **multi-layered semantically annotated language resources** for Latvian,

anchored in widely acknowledged multilingual representations (AMR, PropBank, FrameNet, Universal Dependencies, Grammatical Framework),

that are required for the development of natural language understanding and generation applications.



Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian

NATIONAL DEVELOPMENT PLAN 2020

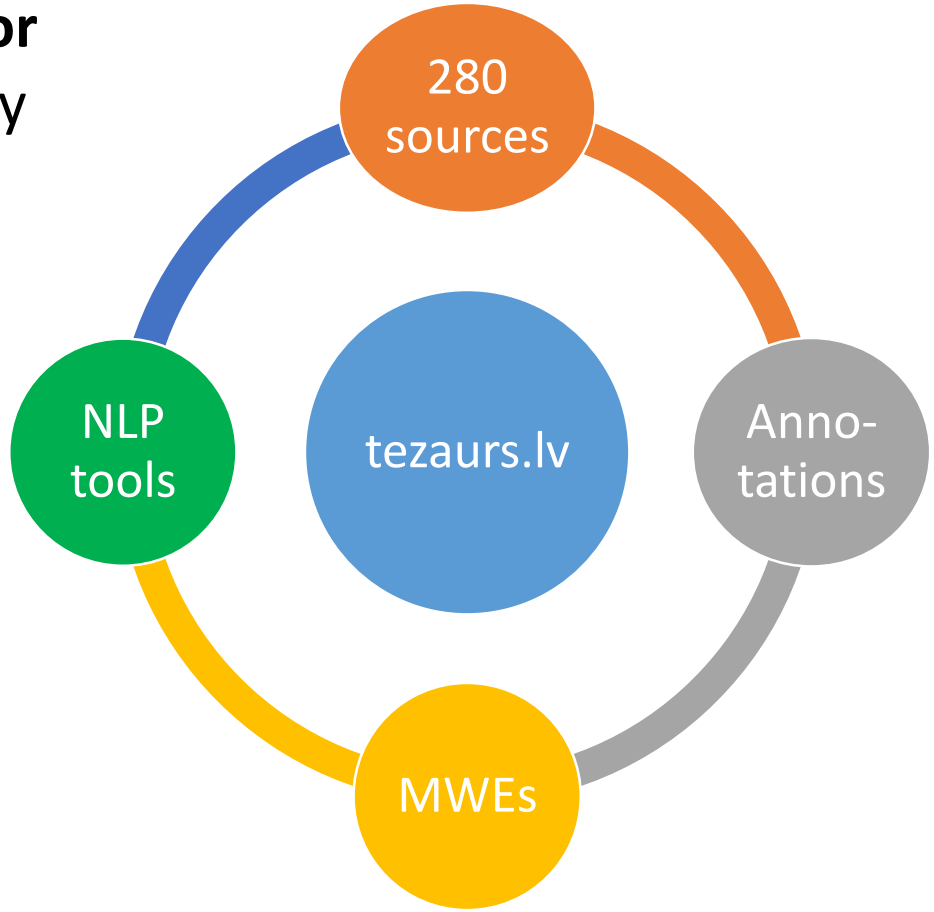
EUROPEAN UNION
European Regional Development Fund

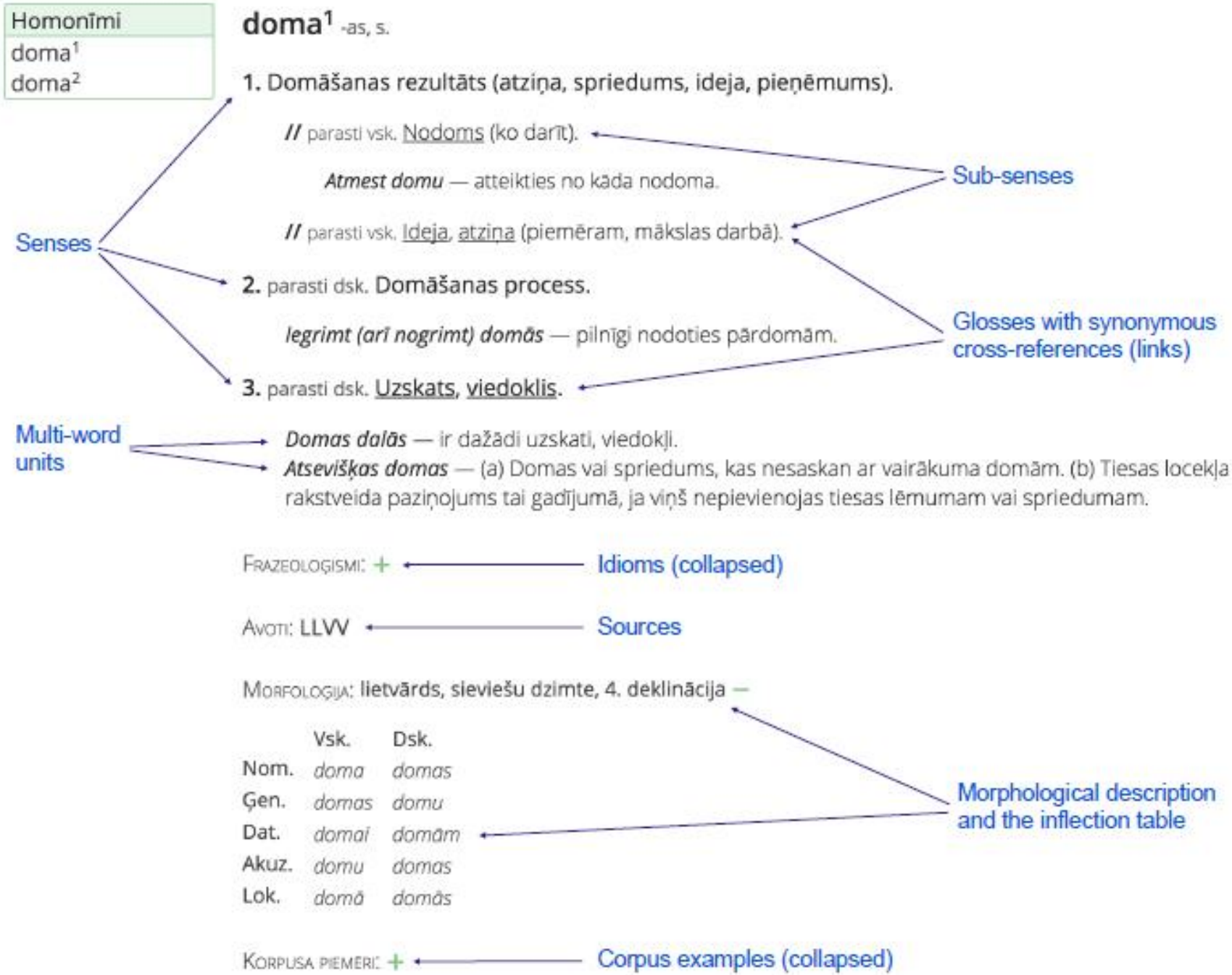
INVESTING IN YOUR FUTURE

Tezaurs.lv - the largest open lexical database for Latvian

Tezaurs.lv aims to be the **central computational lexicon for Latvian**, bringing together all Latvian words and frequently used multi-word units and allowing for the integration of other LT resources and tools:

- currently contains 295 760 lexical entries
- compiled from more than 280 sources
- the dictionary is
  - enriched with phonetic, morphological, semantic and other annotations
  - enhanced with language processing tools allowing generation of inflectional forms and selection of corpus examples on the fly
- more than 2000 requests per day





Spektors et al, 2016

The aim of this study

The aim is to extract lists of **good quality MWE candidates**, which

- can be delivered as open experimental MWE lexicon for Latvian

and

- after manual inspection could be added to the largest Latvian open lexical database *tezaurs.lv*

Strategies for MWE identification and extraction

MWE identification and extraction usually consists of two main steps:

- at first, morpho-syntactic patterns are applied to extract initial list of MWE candidates
- then, the initial list of MWE candidates is filtered by means of statistical measures

In this research:

- we start with application of statistical measures, allowing to identify wide range of MWE categories
- then apply linguistically motivated filters (patterns) to clean the list of initially extracted MWE candidates

Limitation: rather small amount of data

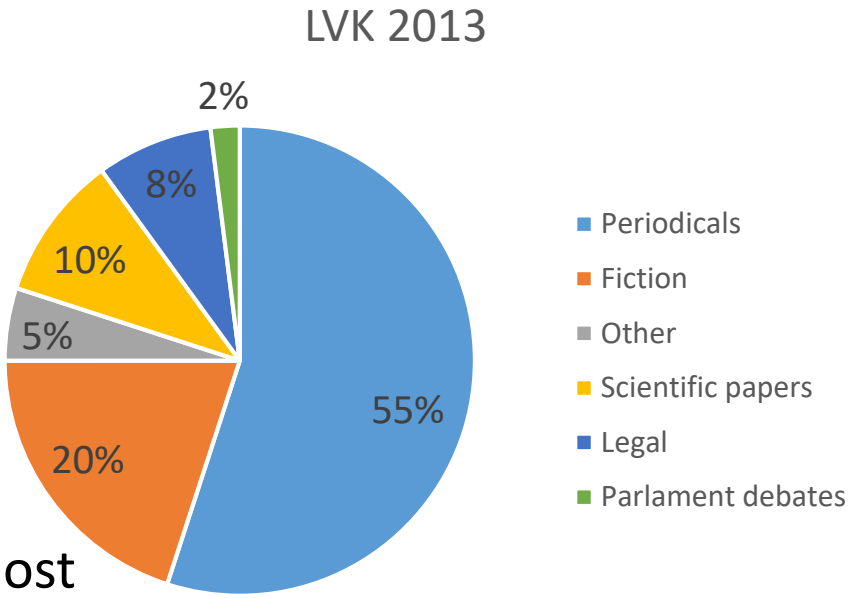
Corpus	Size			
	Sentences (thousands)	Tokens (million)	Unique tokens (thousands)	Unique lemmas (thousands)
Balanced Corpus of the Modern Latvian language	148	5,54	408,01	111,59
Latvian-Lithuanian parallel corpus	223	3,24	307,53	87,88
Open Subtitles corpus	454	2,37	117,01	56,44

Application of statistical measures

- **Frequency** - frequency usage of n-gramms
- **mi-score** (mutual information score): measures the strength of association (frequency of co-occurrence vs. separate occurance). mi-score of 3 or higher is usually considered to be significant
- **t-sore**: measures the confidence of association and could be applied also for low frequency words, the t-score of 2 or higher is considered to be statistically significant

Application of statistical measures

- The Balanced Corpus of the Modern Latvian Language contains **only modern literal Latvian language texts**
- Sentences:148 K, tokens 5,54 (million), unique tokens: 408,01 K, unique lemmas: 111,59 K
- Length of word sequences - up to 5
- Experiments:
  - at first mi-score and t-score were used individually and the most frequent MWE candidates were investigated
  - then we applied t-score as initial filter and afterwards sorted results by mi-score and frequency
- The threshold for t-score was set to 2.5, but for mi-score - 3



Application of Statistical Measures

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
very high (more than 70) mi-score, no MWE candidates	**4 MWE candidates, short, but stable phrases, e.g., multi-word conjunctions**	**4 MWE candidates**	**4 MWE candidates, short, but stable phrases, e.g., multi-word conjunctions**	**4 MWE candidates, complex noun phrases**	**7 MWE candidates, verbal constructions or nouns followed by relative clause**
"(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)	**pensiju shēmas līdzekļu pārvaldītāju reģistrā** (register of the pension scheme asset managers)	**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.)	**Ministru kabineta** (Cabinet of Ministers)

Lemmatization

- When the MWE candidate list is ordered by mi-score, **four named entities** and **five terms** are among top 10 MWE candidates.
- When the list is ordered by frequency, **3 complex function words** and **2 frequent MWEs** are included.

Word sequences with highest mi-score (9 candidates)	Most frequent word sequences (5 cand.)
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)

Filtering MWE Candidates

Statistical filters (thresholds):

- high frequency (and mi-score between 4 to 11) is better signal that the string could be MWE, than high mi-score and low frequency (e.g. below 10)
- in case of t-score – high frequency together with high t-score is a signal for a good MWE candidate
- if t-score is used as initial filter and mi-score is used as the second filter,
  - most of MWE candidates will be frequent and have mi-score value be between 10 and 35,

  or

  - will have high mi-score and low frequency.

Linguistic filters

- Simple regular expressions - filters out ungrammatical word sequences, e.g.:
  - *un* (*and*), *vai* (*or*) as the last word
  - *būt* (*to be*) at the beginning or end of string consisting of two words (e.g. *būt ļoti* (*to be very*)
- Morpho-syntactic filters:
  - filters out word sequences that are ungrammatical

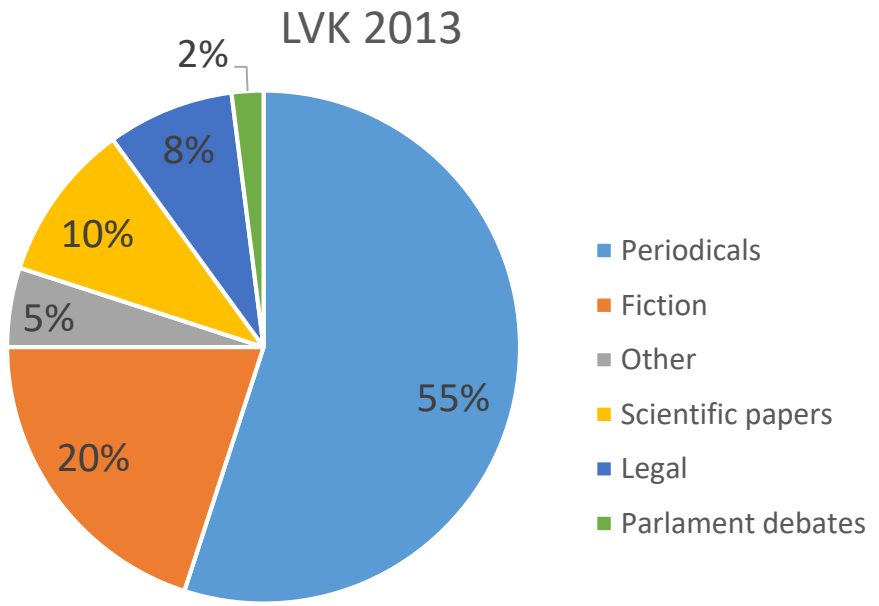
    ir ļoti (is very) , however *stājas spēkā* (*comes into force)*
  - extracts specific categories of MWEs
- Overlapping MWE candidates

  stājas spēkā (comes into force), stājas spēkā ar (comes into force from), kas stājas spēkā (which comes into force)

choice of the most appropriate MWE needs to be made by lexicographer.

Results: Balanced Corpus of the Modern Latvian Language

- The corpus is balanced and contains **only modern literal Latvian language texts**, about 4,5 million words in total.
- Hypothesis: this corpus is a good source to identify different types of MWEs that occur in Latvian rather frequently.
- Findings: simple statistical measures applied on this corpus allow us mainly to identify good MWE candidates for the legal domain (e.g., *stājas spēkā - comes into force*)



Limitation: 2-3 tokens

- When MWEs were identified by mi-score, all top 10 word sequences are MWEs. However, most of MWE candidates are named entities
- In case when MWEs were identified by t-score, 5 strings are named entities, 4 are terms and one (JP NVO RV) is string of characters.

mi-score (10 candidates)	t-score (9 candidates)
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)

t-score as measure for term extraction

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)

The previous 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

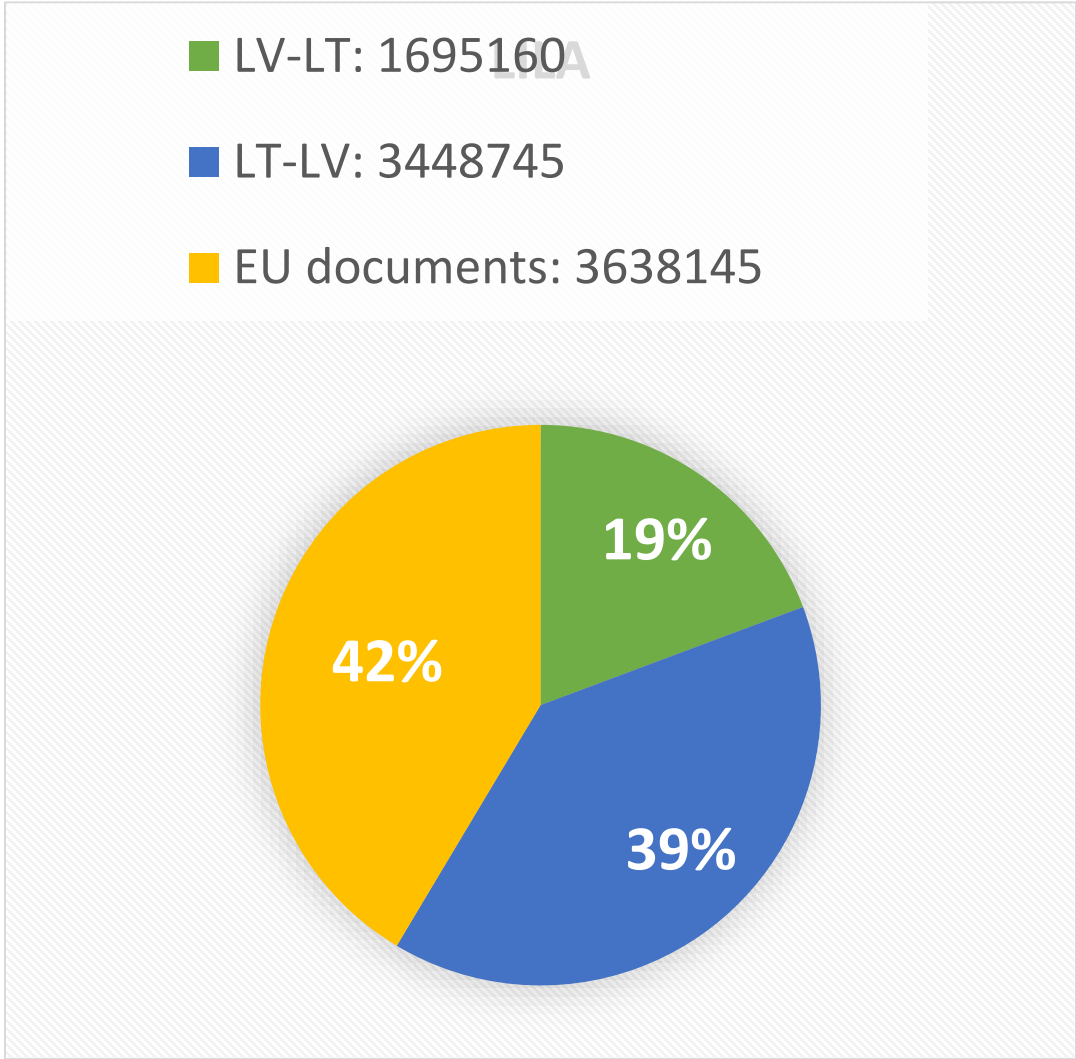
Extraction of verbal phrases

Freq.	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)

From top 10 MWE candidates, **7** could be accepted as MWEs: all are fixed in Latvian-English dictionary (Veisbergs 2005) and four of these MWEs are included in *tezaurs.lv.*

Latvian-Lithuanian Corpus LiLa

- LiLa contains more than half of the texts that are originally written in Latvian (19,3%) or Lithuanian (39,3%)
- 41,4% of corpus are EU legal texts that were chosen due to lack of translations between Latvian and Lithuanian.
- The corpus contains 8,7 million tokens in total; 3,24 million tokens (no legal documents) were used in this experiment
- Texts originally written in Baltic languages represent domains:
  - **modern fiction (86%),**
  - periodicals (5,9%),
  - popular literature (5,6%)



Latvian-Lithuanian Corpus

- Hypothesis: LiLa contains more frequently used fixed phrases and idiomatic expressions than the Balanced Corpus of Modern Latvian
- Results (ordered by mi-score): 7 MWE candidates are named entities, 1 fixed phrase, 1 part of longer phrase, 1 is a character string
- Results (ordered by frequency): 4 terms, 1 named entity, 1 complex function word, 4 other parts of longer phrases

mi-score (7 named entities)	Frequency (6 MWE candidates)
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)

Open Subtitles Corpus

- Sentences: 454 K, tokens - 2,37 million, unique tokens - 117,01 K, unique lemmas - 56,44 K
- We used lemmatized corpus and applied t-score for identification and mi-score as the second filter. In addition, MWE candidates were filtered by frequency (in previous experiments only threshold 5 was used)
- The outcome of this experiment differs from the previous ones – besides named entities, different idiomatic expressions are identified

- We' il help you.

- Viktor, come on.

I owe you so much.

- Please let me help you.

- This is your friend talking.

We' re your family now.

Goodbye.

Open Subtitles Corpus

- The outcome of this experiment differs from the previous ones – besides named entities, different **idiomatic expressions are identified**
- Similarly to the previous experiments, many (4 when frequency is at least 5 and 6 if frequency is at least 10/ 15) of extracted MWEs are named entities.
- However, different idiomatic expression are identified as well

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)

Conclusion

- In case of small amount of general domain (balanced) data automatic methods allow finding good MWE candidates – terms or named entities
- However, finding idiomatic expressions in small, general domain corpora is looking for the needle in haystack: only larger or more expressive corpus could help in identification process
- In case of small general 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
- If the aim of the MWE identification is to identify idiomatic expressions that recently have appeared in a language, then the corpus needs to represent more everyday language and have to be rather big, because idiomatic expressions are rare in balanced corpora that represent literary language and carefully edited texts.

Thank you!

This work is supported by the European Regional Development Fund under the grant agreements No. 1.1.1.1/16/A/219 (*Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian*)