International Conference
"Scientific, Administrative and Educational Dimensions of Terminology (2025)"

www.mruni.eu

EU Parallel Migration Corpus in English and Lithuanian for Terminology Analysis

Olga Ušinskienė
Mykolas Romeris University, Lithuania
olga.usinskiene@mruni.eu

Sigita Rackevičienė
Mykolas Romeris University, Lithuania
sigita.rackeviciene@mruni.eu

Mykolas Romeris University

www.muni.eu

The Aim and Objectives of the Paper:

Aim:

To present the **EU parallel migration corpus** (English-Lithuanian), compiled for bilingual migration terminology analysis, to explore the opportunities and challenges of working with the Lithuanian data in Sketch Engine, and to outline its availability for future research.

Objectives:

- Presenting the building stages of the corpus and its integration into Sketch Engine
- Describing the composition of the corpus
- Highlighting new opportunities provided by Sketch Engine for Lithuanian data processing and discussing the remaining challenges
- Presenting availability of the corpus and its adherence to FAIR data principles

www.muni.eu

The Relevance of the Paper:

- Growing need to clarify migration-related terminology for diverse user groups.
- Heightened relevance due to Eastern European crises (hybrid attacks, war in Ukraine, refugee flows).
- Domain governed by international and national documents, requiring terminology harmonisation.

www.muni.eu

I. Building Stages of a Parallel Corpus

- **Discovering data** – Identifying relevant sources in EUR-Lex
- **Acquiring metadata** – Collecting document information
- **Retrieving, converting & aligning texts** – Using LF Aligner to retrieve documents, convert them to plain text and align sentence pairs.
- **Uploading texts to corpus software** – Importing aligned files to Sketch Engine
- **Uploading metadata** – Linking metadata for advanced search

www.muni.eu

Discovering data

EUR-Lex - official online access point to European Union (EU) law and other public documents maintained by the Publications Office of the European Union

Legally binding documents:

- Primary legislation (treaties of the EU)
- Secondary legislation (directives, regulations, decisions, etc.)
- Case law (judgments and opinions from the Court of Justice of the EU)
- International agreements involving the EU

Legally non-binding documents:

- Preparatory acts (such as legislative proposals, reports, and opinions)
- Parliamentary questions and other official documents
- European Commission Communications, reports, etc.

Available in all 24 official EU languages and allows multilingual search.

Acquiring metadata

Collection of metadata of selected documents:

Form of document, CELEX number (unique identifier assigned to each EU legal document), Title, Author (Institution), Publishing year, Word count, Home Website, URL

Form of document	CELEX	Title of document	Author of the document (individual or collective)	Publishing year	Word count	Home Website	URL
Council decision	32023D2420	Arrangements between the European Union, on the one part, and, respectively, the Republic of Iceland, the Kingdom of Norway, the Swiss Confederation and the Principality of Liechtenstein, on the other part, as regards the modalities of the participation by those countries in the European Asylum Support Office, in order for them to be able to participate in the European Union Agency for Asylum	the Council of the European Union	2023	597	eur-lex.europa.eu	https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202302420
Implementing regulation	32023R0169	Commission Implementing Regulation (EU) 2023/169 of 25 January 2023 establishing the template for the annual performance reports on the Asylum, Migration and Integration Fund for the 2021-2027 programming period, pursuant to Regulation (EU) 2021/1147 of the European Parliament and of the Council	European Commission, Directorate-General for Migration and Home Affairs	2023	2,828	eur-lex.europa.eu	https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32023R0169
Directive	32024L1233	Directive (EU) 2024/1233 of the European Parliament and of the Council of 24 April 2024 on a single application procedure for a single permit for third-country nationals to reside and work in the territory of a Member State and on a common set of rights for third-country workers legally residing in a Member State (recast)	European Parliament, Council of the European Union	2024	9,941	eur-lex.europa.eu	https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401233

www.mruni.eu

Retrieving, converting & aligning texts:

- Using LF Aligner to retrieve documents, convert them to plain text, and align sentence pairs
- Producing TMX (Translation Memory Exchange) files for further processing

LF Alignment Editor 2.3 - aligned_32001L0040_en-lt.txt

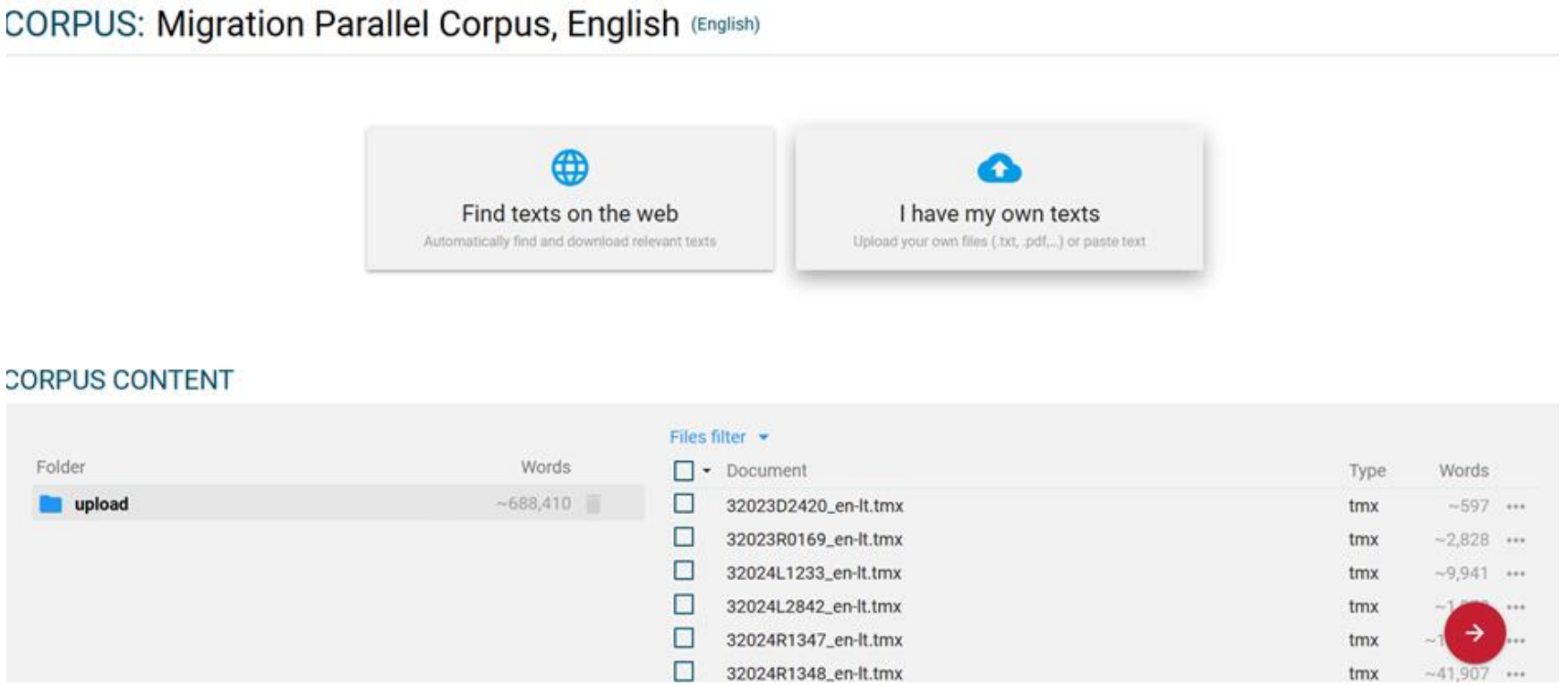
File Edit Help

1	Avis juridique important	Svarbus teisinis pranešimas	32001L0040_en
2	32001L0040	32001L0040	32001L0040_en
3	Council Directive 2001/40/EC of 28 May 2001 on the mutual recognition of decisions on the expulsion of third country nationals		32001L0040_en
4	Official Journal L 149 , 02/06/2001 P. 0034 - 0036	Oficialusis leidinys L 149 , 02/06/2001 p. 0034 - 0036	32001L0040_en
5	Council Directive 2001/40/EC	Tarybos direktyva 2001/40/EB	32001L0040_en
6	of 28 May 2001	2001 m. gegužės 28 d.	32001L0040_en
7	on the mutual recognition of decisions on the expulsion of third country nationals	dėl abipusio sprendimų dėl trečiųjų šalių piliečių išsiuntimo pripažinimo	32001L0040_en
8	THE COUNCIL OF THE EUROPEAN UNION,	EUROPOS SĄJUNGOS TARYBA,	32001L0040_en
9	Having regard to the Treaty establishing the European Community, and in particular Article 63(3) thereof,	atsižvelgdama į Europos bendrijos steigimo sutartį, ypač į jos 63 straipsnio 3 dalį,	32001L0040_en
10	Having regard to the initiative of the French Republic(1),	atsižvelgdama į Prancūzijos Respublikos iniciatyvą [1],	32001L0040_en
11	Having regard to the opinion of the European Parliament(2),	atsižvelgdama į Europos Parlamento nuomonę [2],	32001L0040_en
12	Whereas:	kadangi:	32001L0040_en
13	(1) The Treaty stipulates that the Council is to adopt measures on immigration policy within areas comprising conditions of entry and residence as well as illegal immigration and illegal residence.		32001L0040_en

Document before aligning

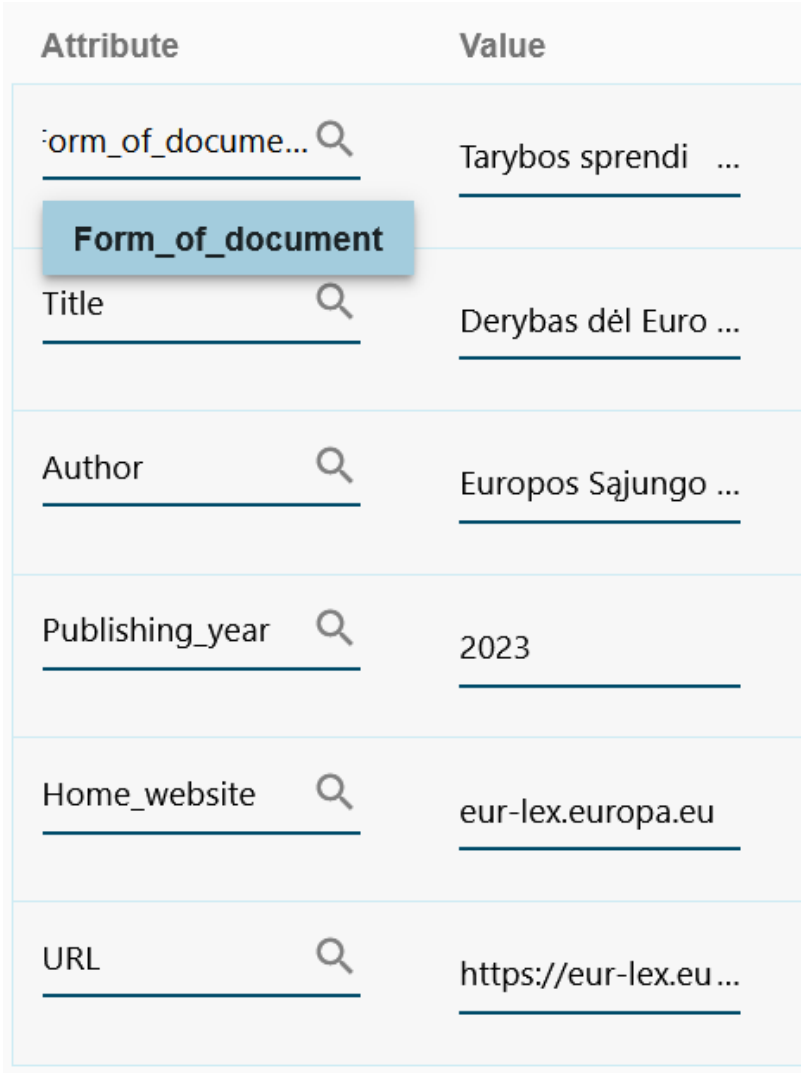
www.muni.eu

Uploading the corpus to Sketch Engine



www.muni.eu

Uploading the metadata to Sketch Engine



Role of Metadata

Enables **sorting texts and creating subcorpora** by:

- Document form
- Institution
- Publishing year

II. Composition of the parallel corpus

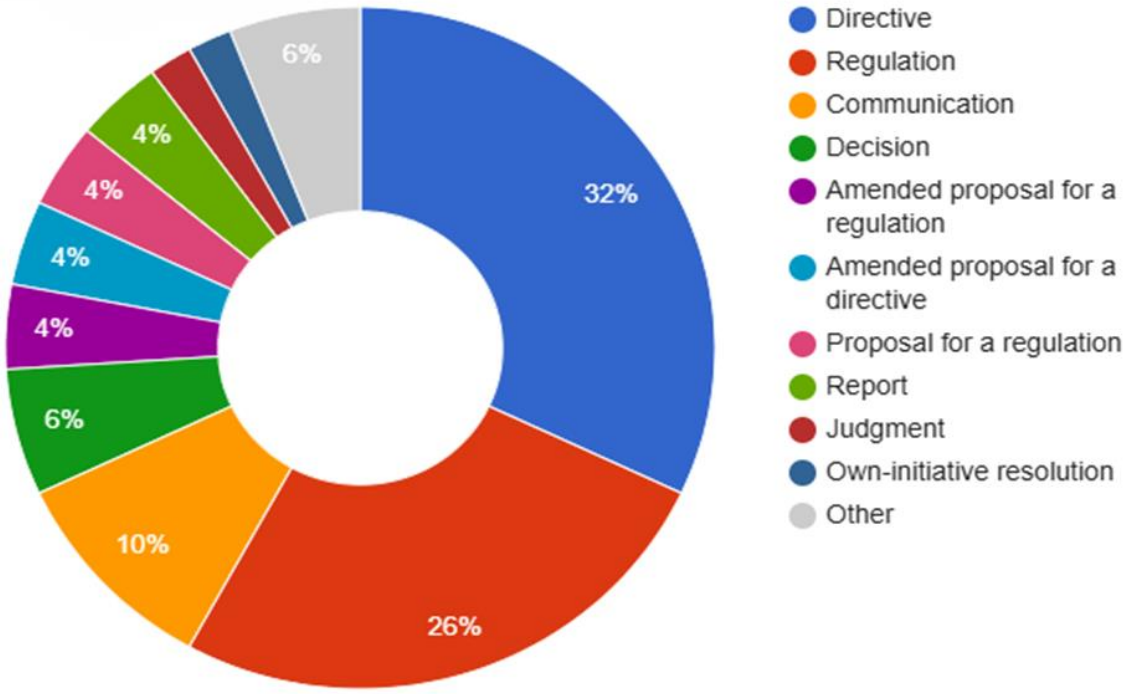
Total number of files: 50 tmx (EN-LT) files

Total number of words: EN 688,410, LT 534,940

- **Forms of the documents:**
- legally binding documents (regulations, directives, decisions, judgments)
- legally non-binding documents (resolutions, reports, communications, proposals, opinions)
- **Institutions that issued documents**
- European Parliament, European Commission, Council of the EU, Committees of the EU, etc.
- **Publication year:**
- From 1998 to 2024

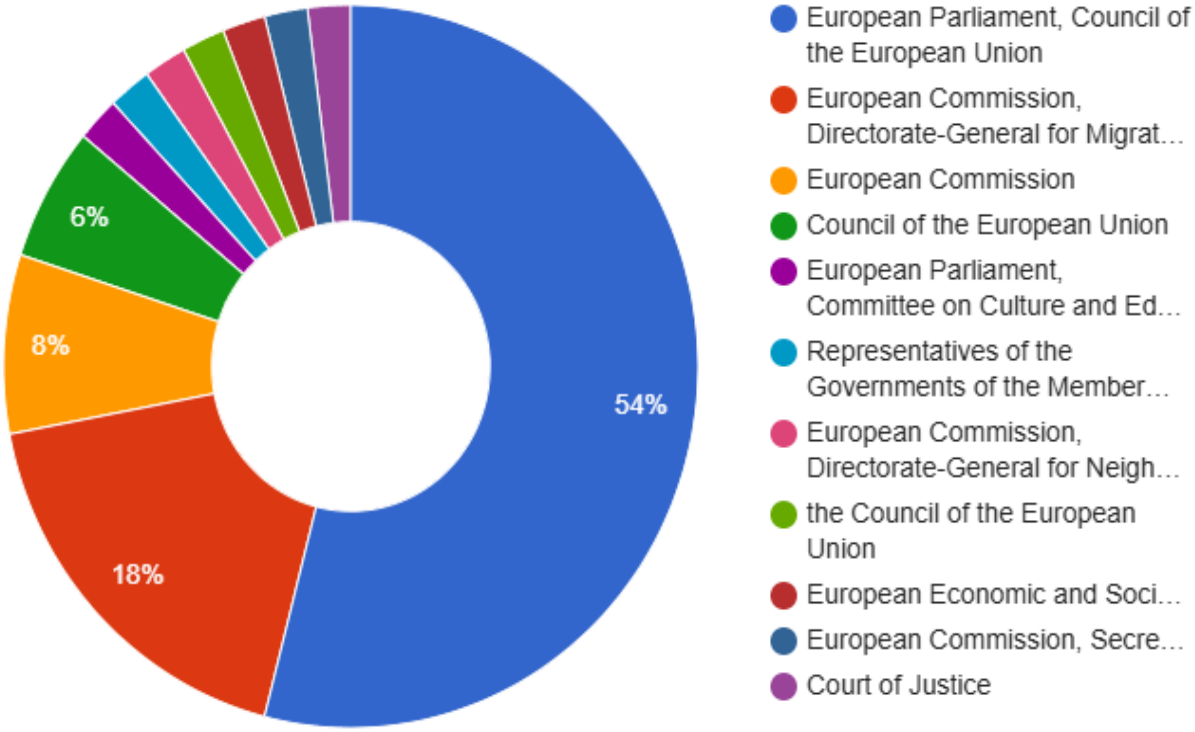
www.muni.eu

The number of documents by their form



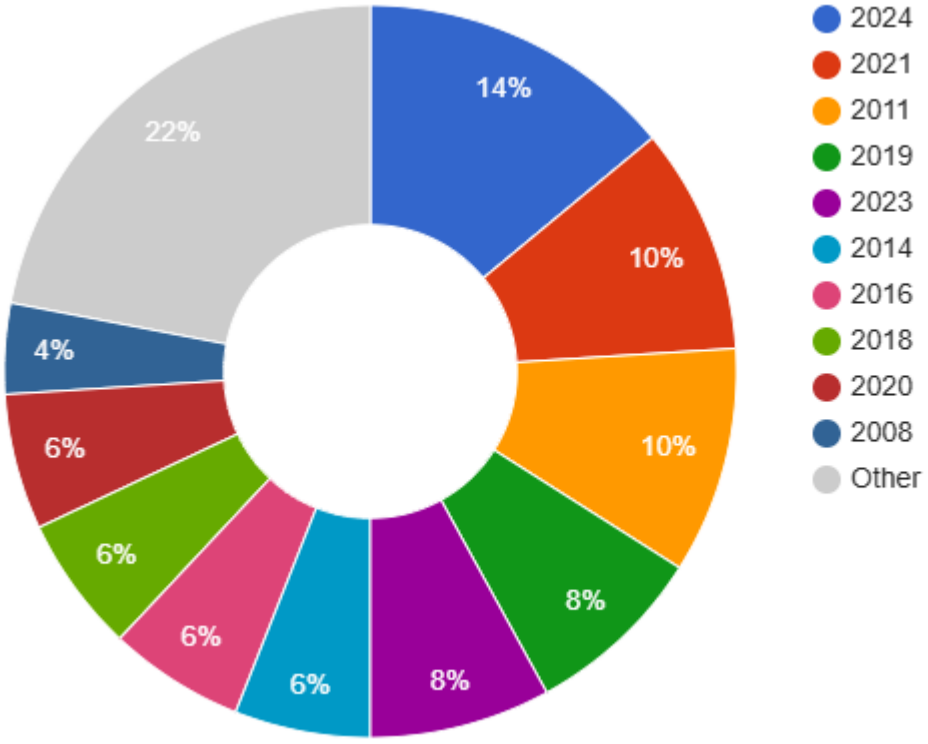
www.mruni.eu

The number of documents by the institution



www.mruni.eu

The number of documents by publishing year

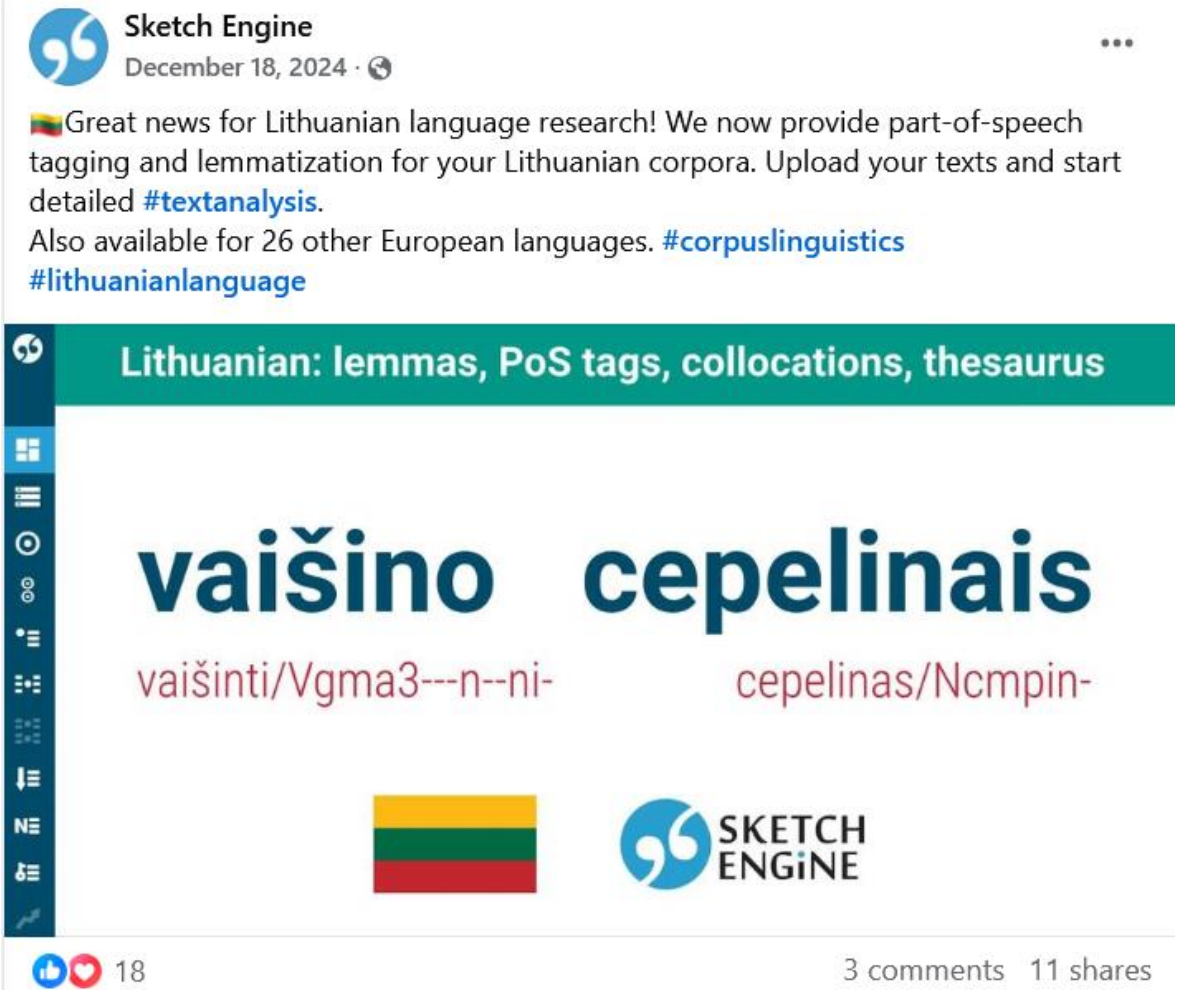


www.muni.eu

www.muni.eu

III. New Opportunities Provided by Sketch Engine for Lithuanian Corpus Analysis and Remaining Challenges

In **December 2024**, **Sketch Engine** released an updated toolkit specifically for **Lithuanian texts**, incorporating enhancements to support richer analysis of Lithuanian corpora.



https://www.facebook.com/SketchEngine/posts/pfbid0pjsxyD3Hu4wZSXtEoGjzGrF4msDb9dGPunZA7VhadTijzpV7eqbzDfSqg9ceoaPwl

www.mruni.eu

www.muni.eu

What does it mean in practice:

1. **Part-of-speech tagging** – tagging each word with its grammatical function
2. **Lemmatisation** – normalising words to their base forms for more accurate frequency analysis and comparison
3. **Word Sketches** – analysing collocations and typical word usage
4. **Keyword & key phrase extraction** – identifying key lexis in the corpus

All these tools enhance terminology discovery and extraction.

www.muni.eu

Part-of-Speech Tagging includes comprehensive tags indicating parts of speech and other grammatical features.

E.g. Noun Tags Include

Type – common or proper noun
Gender – masculine, feminine, common
Number – singular, plural, dual
Case – nominative, genitive, etc.
Reflexiveness – yes, no
Name type (for proper nouns) – first name, surname, geographical name, irrelevant



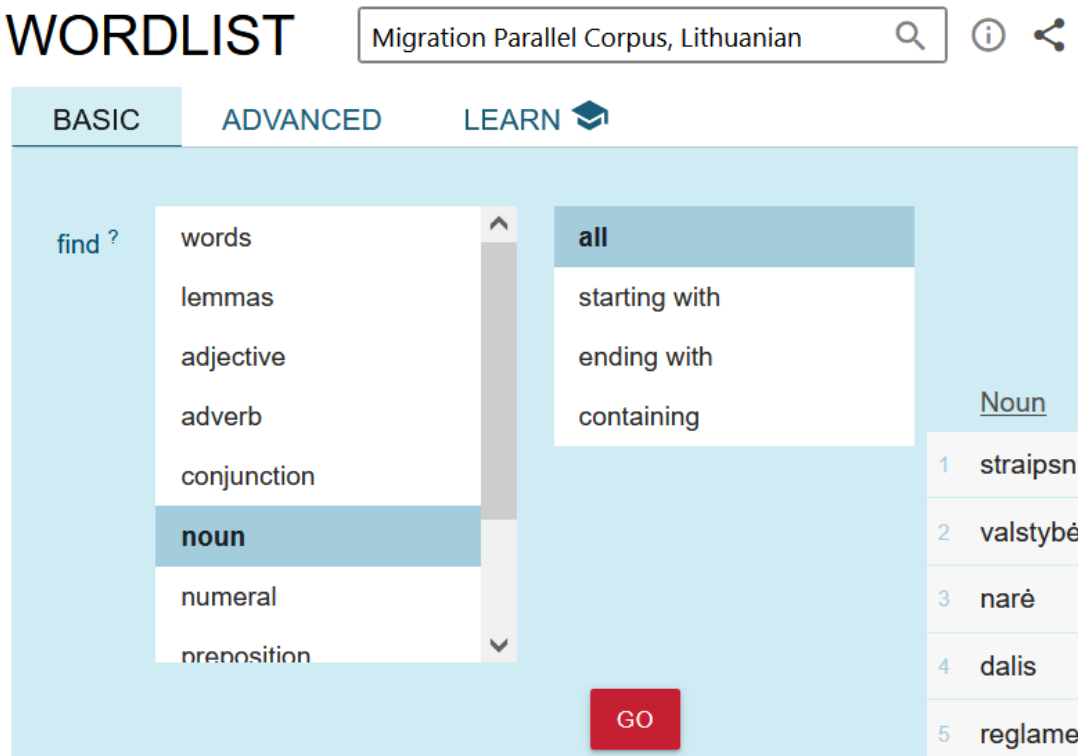
Lemmatisation

- Provides the base form of each word (**lemma**) or base form and part of speech (**lempos**)
- Enables searching for **all grammatical forms of a word in one query**



https://www.sketchengine.eu/lithuanian-multext-east-part-of-speech-tagset/

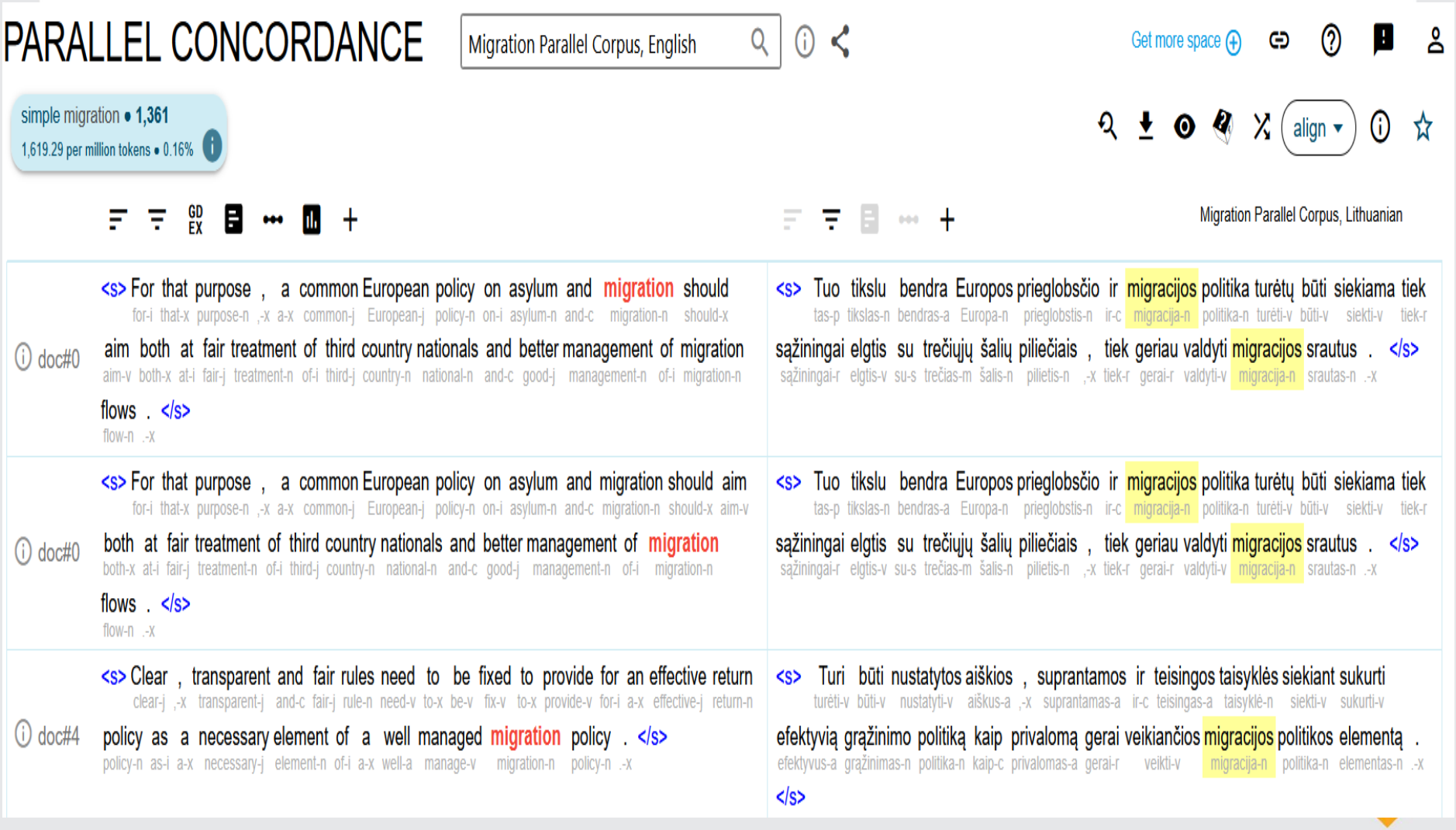
Wordlist tool enables to extract frequency lists of lemmatised words of different parts of speech



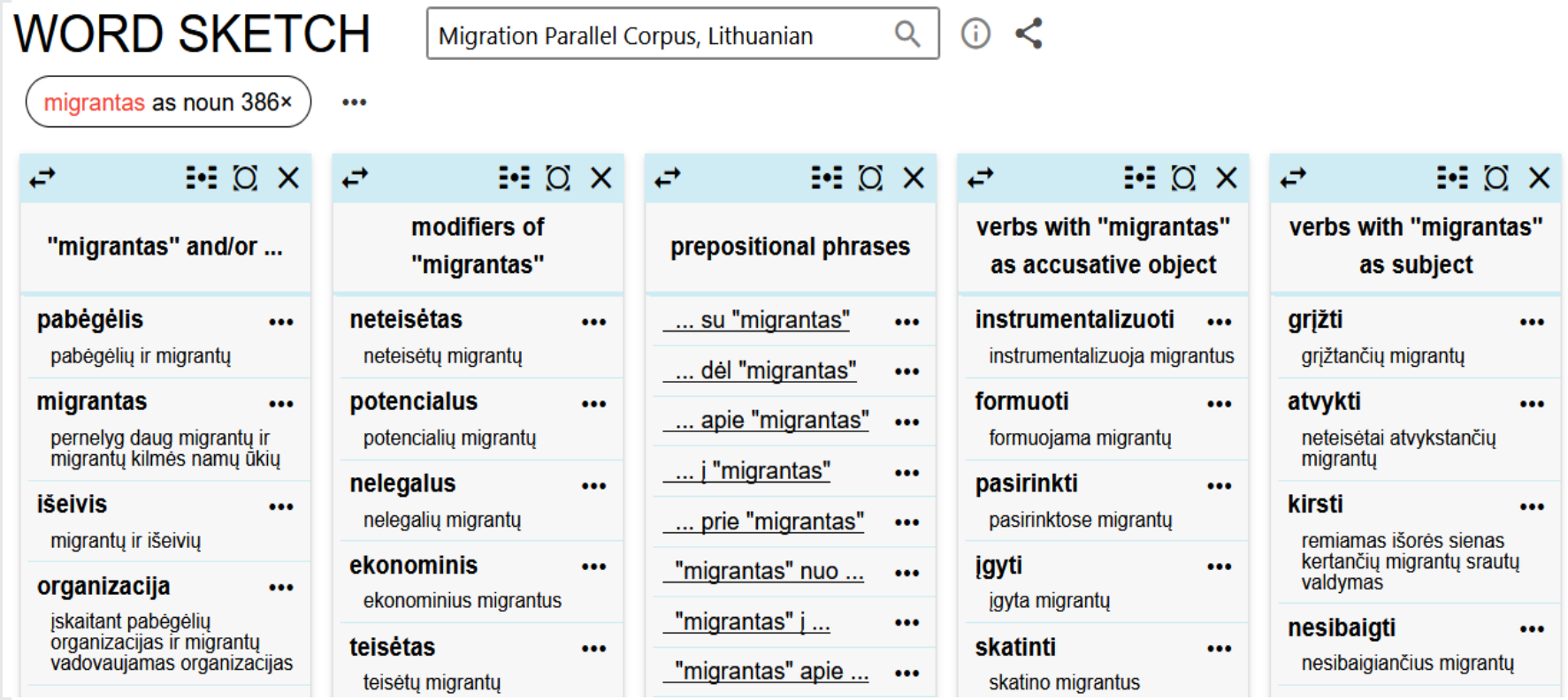
	Noun	Frequency ? ↓	DOCF ?	
1	straipsnis	10,109	49	•••
2	valstybė	9,702	48	•••
3	narė	7,699	47	•••
4	dalis	6,797	49	•••
5	reglamentas	4,109	43	•••
6	šalis	3,934	45	•••
7	europa	3,658	49	•••
8	asmuo	3,600	46	•••
9	teisė	3,573	45	•••
10	apsauga	3,354	46	•••

www.muni.eu

Parallel Concordance tool enables to search for Lithuanian equivalents of selected English terms and explore their lemmas and parts of speech.



Word-Sketch tool enables to extract collocations of selected words organised by lexico-grammatical patterns



www.muni.eu

Lithuanian Word-Sketch grammar needs expansion to include more lexico-grammatical patterns for searching diverse collocations.

Currently it supports the following functions for automatic search of nominal collocations:

```
Sketch Grammar for Lithuanian
lithuanian-multext-hm1-1.0.wsdef.m4

# coordination ="%w" and/or ...

# modifiers of nouns

# predicative adjective (with copula)

# subjects of verbs

# accusative objects of verbs

# prepositional phrases
```

www.mruni.eu

KEYWORD tool extracts key words and phrases of the focus corpus by comparing it to another (reference) corpus

KEYWORDS

Migration Parallel Corpus, Lithuanian

SINGLE-WORDS ✓ MULTI-WORD TERMS ✓

reference corpus: Lithuanian Web 2021 (ltTenTen21)

	Term	Frequency[?] Focus	Reference	Score [?]
1	tarptautinis apsauga	1,600	516	1,865.6
2	apsaugos prašymas	517	49	724.1
3	tarptautinės apsaugos prašymas	509	30	718.7
4	tarptautinis apsaugos prašymas	509	30	718.7
5	valstybės narė	3,688	21,276	511.2

Lithuanian Term Grammar for Single Words needs further development to reduce false-positive keywords and improve accuracy of term extraction.

KEYWORDS | Migration Parallel Corpus, Lithuanian

SINGLE-WORDS ✓ | MULTI-WORD TERMS ✓

reference corpus: Lithuanian Web 2021 (ltTenTen21)

	Lemma	Frequency[?] Focus	Reference	Score [?]
1	prašytojas	2,023	4,864	924.2
2	eurodac	406	90	559.0
3	ð	396	311	498.9
4	ï	302	115	411.7
5	sąveikumas	320	559	368.1
6	eu-lisa	245	7	349.9
7	ol	783	5,095	346.7
8	nelydėti	373	1,502	322.2
9	acquis	266	576	304.3
10	etias	181	29	256.3

www.muni.eu

Lithuanian term grammar for multi-words needs modifications in order to more accurately extract multi-word combinations. The following issues are observed:

Mismatched Adjective-Noun Gender Agreement:

adjective gender is not adapted to feminine nouns), e.g.

tarptautinis apsauga	(should be *tarptautinė apsauga*)
nacionalinis teisė	(should be *nacionalinė teisė*)
daugybinis tapatybė	(should be *daugybinė tapatybė*)

Missing Pronominal (Definitive) Forms of Adjectives:

pronominal (definitive) forms of adjectives are not recognised, e.g.

mėlynas kortelė	(should be *mėlynoji kortelė*)
mėlynas kortelės turėtojas	(should be *mėlynosios kortelės turėtojas*)

Incomplete Noun phrases:

extracted multi-word combinations are not complete noun phrases

atitinkamas Sąjunga	(should be *atitinkamos Sąjungos įstaigos / teisės sritys / priemonės*)
perkeliamo asmuo	(should be *perkeliamo asmens veikla / trumpalaikio judumo laikotarpis / leidimas*)

www.mruni.eu

www.muni.eu

III. Corpus Availability –

Corpus availability adheres to FAIR principles of data management:

F – Findable

A – Accessible

I – Interoperable

R - Reusable

Corpus Availability (1) – Sketch Engine

The corpus is stored on the **Sketch Engine** platform and can be shared upon request with other Sketch Engine users.

FAIR principles:

Findable: No

Accessible: Partially - available to **Sketch Engine users** upon request

Interoperable: **Yes** - downloadable in open standard formats (.txt, .tmx, .vert)

Reusable: **Yes** - files can be reused in research in Sketch Engine or other tools



Corpus Availability (2) – CLARIN-LT

The corpus will be deposited in the **CLARIN-LT (open research infrastructure) repository** to ensure long-term open access and preservation.

Deposited in open standard (non-proprietary) formats:

Translation Memory Exchange (.tmx) – aligned EN–LT texts

Plain text (.txt) – EN and LT texts separately

Vertical text (.vert) – EN and LT texts separately

Comma-Separated Values (.csv) – metadata for EN and LT texts

Depositing the corpus in CLARIN-LT will ensure **full FAIR compliance:**

Findable – easily findable in the repository, with a stable URL or persistent identifier

Accessible – publicly available

Interoperable – compatible with multiple tools and platforms

Reusable – includes metadata and documentation for future research

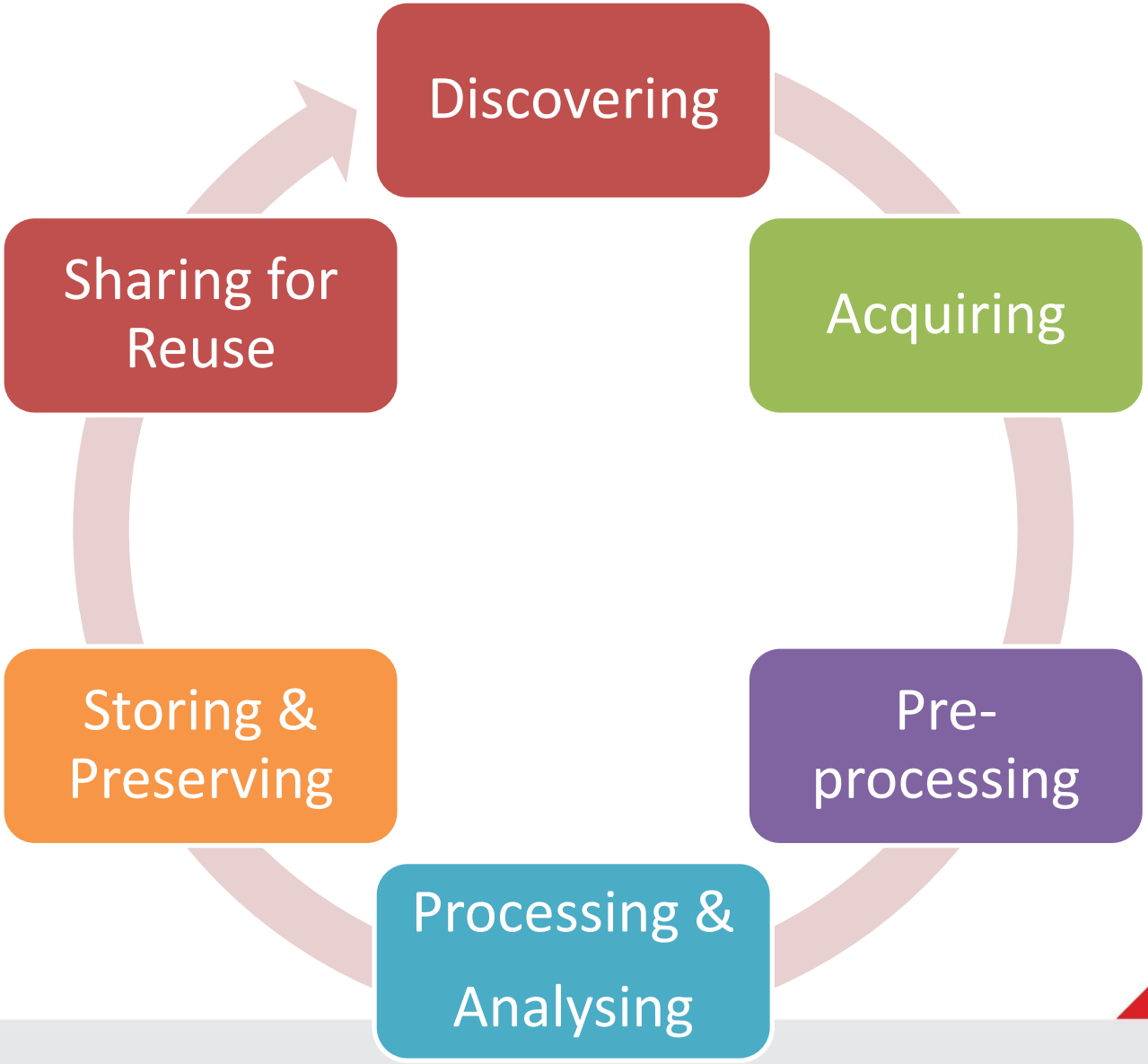
CLARIN-LT

www.mruni.eu

Conclusions

www.muni.eu

The **English–Lithuanian Parallel Migration Corpus** follows a corpus lifecycle adhering to FAIR data principles:



www.mruni.eu

Discovering: Identifying relevant EU documents (1998–2024).

Acquiring: Collecting texts and their metadata from EU institution repositories.

Pre-processing: Converting, aligning, lemmatising, and POS tagging the texts.

Processing & Analysing: Extracting word sketches, keywords and analysing them.

Storing & Preserving: Documenting the dataset and depositing it in an open research infrastructure repository for long-term access.

Sharing for Reuse: Making the corpus accessible to researchers, providing documentation to facilitate reuse, and ensuring compliance with FAIR principles.

www.muni.eu

www.mruni.eu

Thank you for attention!

Mykolas Romeris University