Baltic J. Modern Computing, Vol. 12 (2024), No. 4, pp. 636–645
https://doi.org/10.22364/bjmc.2024.12.4.23

Korpuss.lv - a Versatile Platform for Digital Humanities

Roberts DARĢIS, Baiba SAULĪTE

Institute of Mathematics and Computer Science, University of Latvia
29 Raiņa boulevard, Riga, LV-1459, Latvia

{roberts.dargis, baiba.saulite}@lumii.lv

ORCID 0000-0001-9375-6410, ORCID 0000-0002-7300-8674

Abstract. The Latvian National Corpora Collection (LNCC), accessible through Korpuss.lv, is an extensive and diverse collection of about 40 text and spoken corpora, totalling 2.8 billion tokens. These corpora represent a wide range of text types, such as news articles, blogs, scientific texts, parliamentary debates, and essays. Importantly, almost all the corpora in the LNCC have been re-annotated with a uniform morpho-syntactic annotation scheme, enabling federated search and consistent linguistic analysis across different text types and genres. This feature is especially valuable for computational linguistics and language technology development, offering objective data for studies in lexicography, terminology, grammar, semantics, and language learning. Thus, Korpuss.lv emerges as a critical tool in the digital humanities, helping to develop and refine language technologies and research methodologies.

Keywords: corpus linguistics, corpora collection, federated search, noSketch Engine, timeline

1 Introduction

Language research seeks to understand the structure, function, and use of language through systematic study. Corpora, which are large, structured collections of textual or spoken language data, are invaluable in this endeavour (McEnery and Hardie, 2012). They provide a robust empirical basis for both qualitative and quantitative analysis, enhancing the accuracy and depth of linguistic studies. By using corpora, researchers can achieve greater objectivity and reproducibility in their studies. As language continues to evolve, the importance of corpora in documenting and analyzing these changes only grows, making them an indispensable resource in linguistic research.

Despite the numerous advantages, several challenges limit the effective use of corpora. Many corpora are not readily accessible to researchers due to complex formats or the need for specialized software to prepare the data for analysis. This often requires advanced technical skills, creating barriers for researchers without such expertise. Additionally, running annotation tools can be computationally demanding.

This paper outlines our efforts to promote corpus-based research in the digital humanities and linguistics by improving the accessibility of corpora. We began by introducing the Latvian National Corpora Collection (LNCC), a curated list of corpora, with their metadata hosted on Korpuss.lv. The list of corpora included in the LNCC is carefully curated to maintain the representativeness and quality of the LNCC. The minimum requirement (apart from general content quality requirements) for the inclusion of a Latvian language corpus in the LNCC is open access to the corpus, even if it is not available as open data (Saulite et al., 2022).

We then focused on supporting common research scenarios by providing the corpora in a unified annotation format via the noSketch Engine platform. We also provide training materials[1] and organize workshops for teachers, researchers and students on how to use corpora efficiently. The use of the developed corpora and tools has been integrated into courses such as Computational Linguistics, Contemporary Latvian: Phonetics and Lexicology.

Currently, 39 corpora, developed by 13 institutions, are available within the LNCC. Most of these corpora are annotated with a uniform morpho-syntactic annotation schema (Paikens et al., 2024) and included in the federated search, which combines multiple corpora from two corpus indexing instances (endpoints) maintained by the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) and the National Library of Latvia. The federated search includes 34 corpora, comprising 2.8 billion tokens[2].

2 Korpuss.lv

Korpuss.lv serves as the main access point for the LNCC. (Saulite et al., 2022) The primary goal of the website's user interface (UI) and user experience (UX) design is to facilitate the discovery of various corpora. The site is available in both Latvian and English. The homepage (Figure 1) contains a list of corpus information cards with filtering and sorting functionalities that help researchers identify the most relevant corpora for their studies.

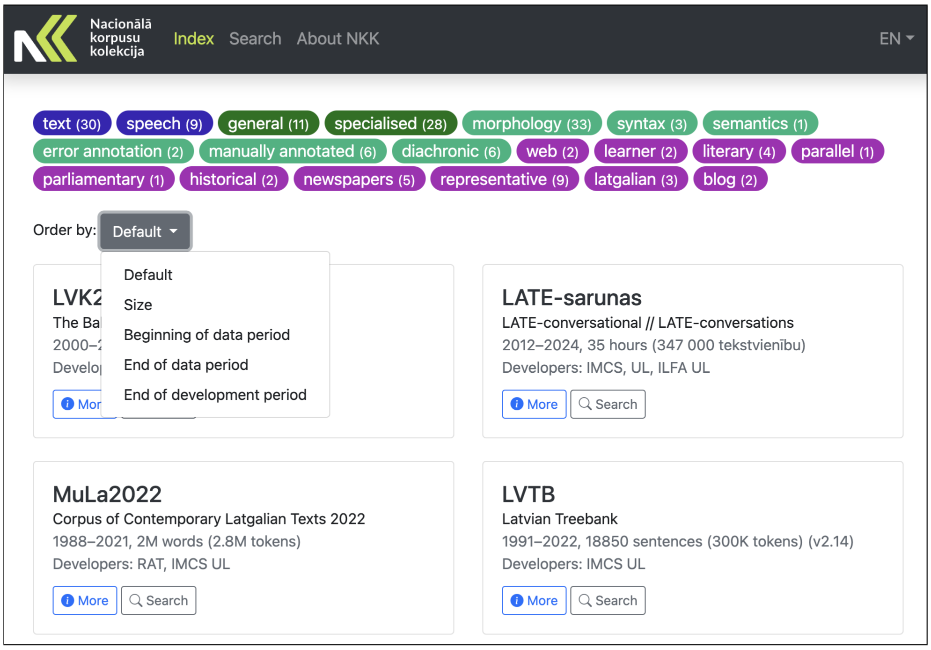
At the top of the homepage, a tag cloud displays the number of corpora associated with each tag, allowing researchers to refine the corpus list based on specific categories. This classification is inspired by the CLARIN resource families[3] and the Czech National Corpus project[4]. The LNCC platform provides filtering by meta-tags: type of data included in the corpus (written and speech), type of corpus (general and specialised), type of the annotation level or metadata (morphology, syntax, error annotation, manually annotated, diachronic). Some corpora are also grouped by other common features, such as historical, literary or newspapers.

[1] https://korpuss.lv/docs

[2] https://korpuss.lv/en/about

[3] https://www.clarin.eu/resource-families

[4] https://wiki.korpus.cz/doku.php/en:cnk:uvodd



Nacionālā korpusu kolekcija Index Search About NKK EN

text (30) speech (9) general (11) specialised (28) morphology (33) syntax (3) semantics (1) error annotation (2) manually annotated (6) diachronic (6) web (2) learner (2) literary (4) parallel (1) parliamentary (1) historical (2) newspapers (5) representative (9) latgalian (3) blog (2)

Order by: Default

Default
Size
Beginning of data period
End of data period
End of development period

LVK2
The Ba
2000–2
Develo
More

LATE-sarunas
LATE-conversational // LATE-conversations
2012–2024, 35 hours (347 000 tekstvienību)
Developers: IMCS, UL, ILFA UL
More Search

MuLa2022
Corpus of Contemporary Latgalian Texts 2022
1988–2021, 2M words (2.8M tokens)
Developers: RAT, IMCS UL
More Search

LVTB
Latvian Treebank
1991–2022, 18850 sentences (300K tokens) (v2.14)
Developers: IMCS UL
More Search

Fig. 1. Homepage of Korpuss.lv

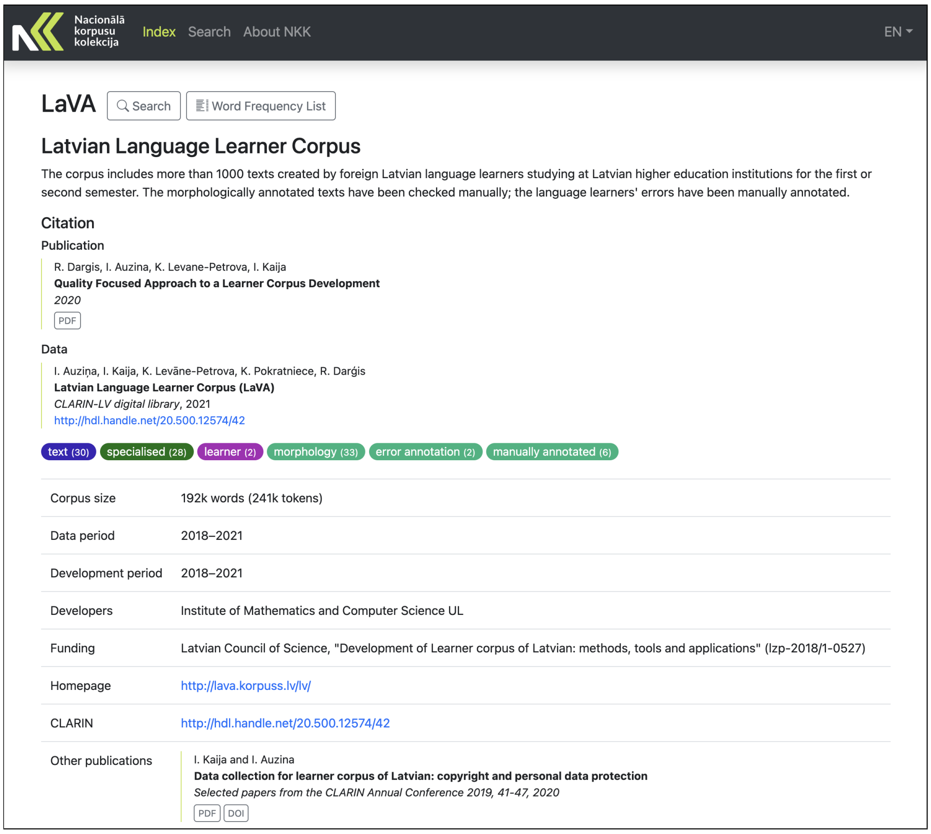
By default, four highlighted corpora picked by editorial team are prominently displayed at the top, while the remaining corpora are listed alphabetically. Researchers can also sort the corpora based on several criteria:

- Size: Corpora are sorted by token count, with the largest displayed first.
- Beginning of data period: Corpora are listed with the oldest data at the top, useful for finding historical corpora or newer data when scrolling down.
- End of data period: Corpora are sorted by the most recent data, helping researchers find the most contemporary datasets or older ones.
- End of development period: Corpora are sorted by their latest updates, with the most recent on top. Useful for identifying recently concluded or ongoing projects.

To maintain simplicity, the interface does not include an option to reverse the sort order. Since all corpora are displayed simultaneously (without pagination), researchers can scroll through the list from the bottom up if needed.

Each corpus card provides key details such as the corpus code, full name, data period, size, developers, and a search button if an online search is available. Clicking on a card leads to a detailed corpus information page (Figure 2).

The corpus information page provides comprehensive details and related resources. This includes links to the corpus homepage, search functionalities, word frequency lists, metadata tags associated with the particular corpus, and the CLARIN repository. These



LaVA

Latvian Language Learner Corpus

The corpus includes more than 1000 texts created by foreign Latvian language learners studying at Latvian higher education institutions for the first or second semester. The morphologically annotated texts have been checked manually; the language learners' errors have been manually annotated.

Citation

Publication

R. Dargis, I. Auzina, K. Levane-Petrova, I. Kaija
Quality Focused Approach to a Learner Corpus Development
2020
PDF

Data

I. Auziņa, I. Kaija, K. Levāne-Petrova, K. Pokratniece, R. Darģis
Latvian Language Learner Corpus (LaVA)
CLARIN-LV digital library, 2021
http://hdl.handle.net/20.500.12574/42

text (30) specialised (28) learner (2) morphology (33) error annotation (2) manually annotated (6)

Corpus size	192k words (241k tokens)
Data period	2018–2021
Development period	2018–2021
Developers	Institute of Mathematics and Computer Science UL
Funding	Latvian Council of Science, "Development of Learner corpus of Latvian: methods, tools and applications" (lzp-2018/1-0527)
Homepage	http://lava.korpuss.lv/lv/
CLARIN	http://hdl.handle.net/20.500.12574/42
Other publications	I. Kaija and I. Auzina **Data collection for learner corpus of Latvian: copyright and personal data protection** *Selected papers from the CLARIN Annual Conference 2019, 41-47, 2020* PDF DOI

Fig. 2. Information page for the LaVA corpus

resources simplify access for researchers, facilitating the exploration of the corpus in their scholarly work.

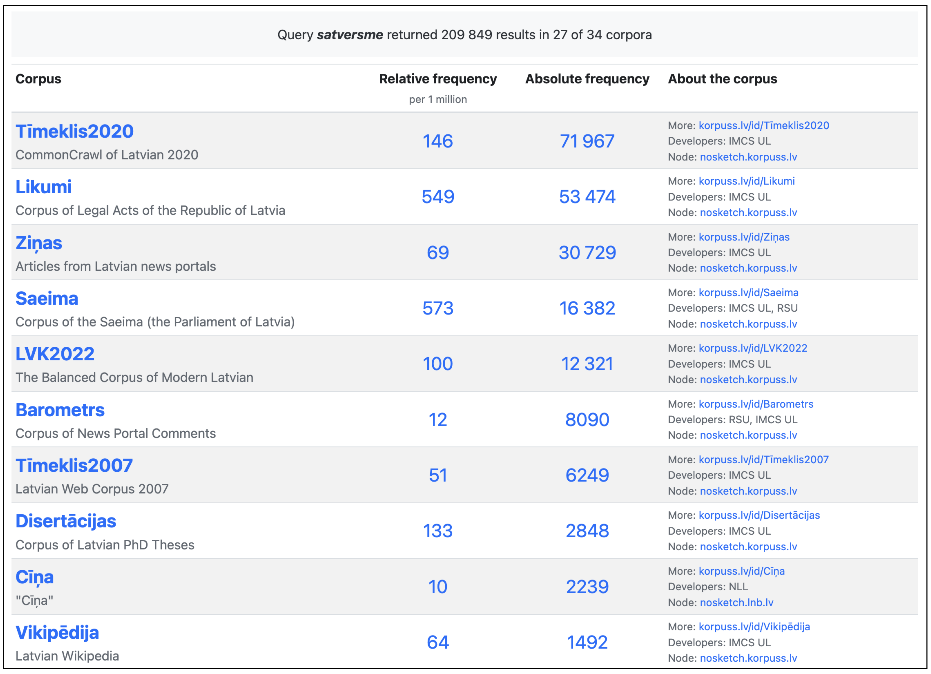
The CLARIN Repository is a central component of the Common Language Resources and Technology Infrastructure (CLARIN), a European initiative designed to provide sustainable access to a wide array of digital language resources and tools. The CLARIN repository hosts metadata and, for some corpora, also provides the actual data for download.

Additionally, the corpus information page lists key publications associated with the corpus. These publications provide insights into the corpus development, methodologies, and significant findings. Such references are critical for understanding the research context of the corpus. Furthermore, the page provides detailed citation information to ensure proper acknowledgment in academic work. This includes one highlighted publication and a data reference if the corpus is published in the CLARIN repository.

Correct citation is essential for the authors of the corpus, as it enhances their citation-based metrics and indexes. Corpora are often developed as part of funded research projects, where various key performance indicators (KPIs) are used to evaluate the impact of the project. One such KPI is the number of studies utilizing the corpus. Proper citation helps to identify and document these studies, thereby demonstrating the relevance and impact of the corpus. This, in turn, aids authors in securing further funding.

3 Federated Content Search

Federated content search (FCS) is an efficient way to locate relevant corpora for a study. FCS allows researchers to search for specific words or phrases across multiple corpora simultaneously. Currently, 34 out of the 39 corpora in the LNCC are searchable via FCS, covering 2.8 billion tokens.



Query ***satversme*** returned 209 849 results in 27 of 34 corpora

Corpus	Relative frequency per 1 million	Absolute frequency	About the corpus
Tīmeklis2020 CommonCrawl of Latvian 2020	146	71 967	More: korpuss.lv/id/Tīmeklis2020 Developers: IMCS UL Node: nosketch.korpuss.lv
Likumi Corpus of Legal Acts of the Republic of Latvia	549	53 474	More: korpuss.lv/id/Likumi Developers: IMCS UL Node: nosketch.korpuss.lv
Ziņas Articles from Latvian news portals	69	30 729	More: korpuss.lv/id/Ziņas Developers: IMCS UL Node: nosketch.korpuss.lv
Saeima Corpus of the Saeima (the Parliament of Latvia)	573	16 382	More: korpuss.lv/id/Saeima Developers: IMCS UL, RSU Node: nosketch.korpuss.lv
LVK2022 The Balanced Corpus of Modern Latvian	100	12 321	More: korpuss.lv/id/LVK2022 Developers: IMCS UL Node: nosketch.korpuss.lv
Barometrs Corpus of News Portal Comments	12	8090	More: korpuss.lv/id/Barometrs Developers: RSU, IMCS UL Node: nosketch.korpuss.lv
Tīmeklis2007 Latvian Web Corpus 2007	51	6249	More: korpuss.lv/id/Tīmeklis2007 Developers: IMCS UL Node: nosketch.korpuss.lv
Disertācijas Corpus of Latvian PhD Theses	133	2848	More: korpuss.lv/id/Disertācijas Developers: IMCS UL Node: nosketch.korpuss.lv
Cīņa "Cīņa"	10	2239	More: korpuss.lv/id/Cīņa Developers: NLL Node: nosketch.lnb.lv
Vikipēdija Latvian Wikipedia	64	1492	More: korpuss.lv/id/Vikipēdija Developers: IMCS UL Node: nosketch.korpuss.lv

Fig. 3. Federated search result for the search term *satversme* (constitution)

Search queries can be made on the token or lemma layers. Instructions for query formulation are available under a question mark icon. The search query supports wildcard symbols for single and multiple characters, as well as the *OR* operator.

The search results display both absolute hit frequency and relative frequency per million tokens (see Figure 3). By comparing relative frequencies across different corpora, researchers can quickly assess term usage in various domains, such as news, legal texts, or novels. Absolute frequencies help identify which corpus contains the most occurrences of rare terms.

Results can be sorted alphabetically or by relative/absolute frequency by clicking on the column headers. Clicking a result leads to a concordance view, and a direct link to the corpus info page is also provided.

4 noSketch Engine

noSketch Engine[5] is a robust corpus management and analysis tool widely used by linguists, lexicographers, and digital humanities researchers (Kilgarriff et al., 2014). It facilitates the exploration of large text corpora by enabling complex queries to extract detailed insights into word usage, syntactic patterns, and semantic relationships. The platform provides essential features such as concordances, collocations, frequency lists, and timelines, making it a valuable resource for linguistic and cultural analysis.

At the time of publishing the initial corpora, noSketch Engine was selected for its maturity and comprehensive functionality, which were unmatched by other available tools. However, recent advancements in the field have led to the development of alternative platforms, including KonText (Machálek, 2020) and Korp (Borin et al., 2012), among others. These tools offer additional functionalities that may complement or expand upon the capabilities of noSketch Engine.

Future development plans include exploring the possibility of hosting additional corpus management platforms alongside noSketch Engine. Of particular interest are features tailored to the analysis of semantically annotated corpora, which could significantly enhance the range of analytical methods available to researchers.

4.1 Concordance

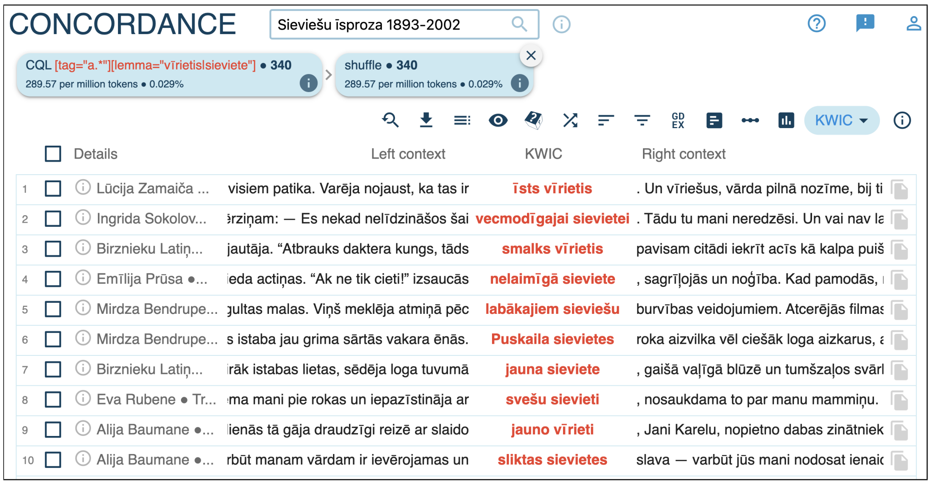
A concordance is a list of occurrences of a word or phrase in a corpus, shown within its immediate context. This feature allows researchers to examine how a specific word or expression is used across different texts, time periods, or genres. In digital humanities, concordances are valuable for studying the nuanced meanings of words, tracking shifts in discourse, and understanding the cultural context in which language is used.

A basic search allows to generate concordances of words or phrases by wordform, lemma or tag. Concordances can be further filtered down by text types (corpus metadata fields), which vary across corpora. A corpus query language (CQL) can be used for more advanced use cases, such as filtering by more than one attribute (for example lemma and part of a tag together). These filters can also be combined to create advanced search criteria for a phrase - all adjectives followed by the lexemes *sieviete* (woman) or *vīrietis* (man). (see Figure 4).

4.2 Frequency List

Frequency lists rank words or phrases by their occurrence in a corpus. These lists are useful for identifying key themes or topics within large datasets, and for comparing linguistic features across text types, genres, or periods.

[5] http://www.sketchengine.eu


CONCORDANCE
Sieviešu īsproza 1893-2002
CQL [tag="a.*"][lemma="vīrietis|sieviete"] • 340
289.57 per million tokens • 0.029%
shuffle • 340
289.57 per million tokens • 0.029%
GD EX
KWIC

	☐	Details	Left context	KWIC	Right context
1	☐	ⓘ Lūcija Zamaiča ...	visiem patika. Varēja nojaust, ka tas ir	īsts vīrietis	. Un vīriešus, vārda pilnā nozīme, bij ti
2	☐	ⓘ Ingrida Sokolov...	erziņam: — Es nekad nelīdzināšos šai	vecmodīgajai sievietei	. Tādu tu mani neredzēsi. Un vai nav la
3	☐	ⓘ Birznieku Latiņ...	jautāja. “Atbrauks daktera kungs, tāds	smalks vīrietis	pavisam citādi iekrīt acīs kā kalpa puiš
4	☐	ⓘ Emīlija Prūsa •...	ieda actiņas. “Ak ne tik cieti!” izsaucās	nelaimīgā sieviete	, sagrīļojās un noģība. Kad pamodās,
5	☐	ⓘ Mirdza Bendrupe...	jultas malas. Viņš meklēja atmiņā pēc	labākajiem sieviešu	burvības veidojumiem. Atcerējās filmas
6	☐	ⓘ Mirdza Bendrupe...	s istaba jau grima sārtās vakara ēnās.	Puskaila sievietes	roka aizvilka vēl ciešāk loga aizkarus, a
7	☐	ⓘ Birznieku Latiņ...	irāk istabas lietas, sēdēja loga tuvumā	jauna sieviete	, gaišā vaļīgā blūzē un tumšzaļos svārl
8	☐	ⓘ Eva Rubene • Tr...	ma mani pie rokas un iepazīstināja ar	svešu sievieti	, nosaukdama to par manu mammiņu.
9	☐	ⓘ Alija Baumane •...	ienās tā gāja draudzīgi reizē ar slaido	jauno vīrieti	, Jani Karelu, nopietno dabas zinātniek
10	☐	ⓘ Alija Baumane •...	rbūt manam vārdam ir ievērojamas un	sliktas sievietes	slava — varbūt jūs mani nodosat ienaic

Fig. 4. Concordance lines for the CQL `[tag="a.*"][lemma="vīrietis|sieviete"]`

A frequency list can be generated using the word list option in the main menu or from concordances, which provides more flexibility. For example, a researcher could search for adjectives and create a list of the most frequent adjectives, or search for a particular lemma and list its common wordforms. The search is not restricted to single words. For example, if the researcher is looking for adjectives followed by *sieviete* (woman) or *vīrietis* (man) (Figure 4), the frequency list offers the possibility to compare the difference in adjectives used to describe a man and a woman. The first two adjectives used to describe both women and men in the Corpus of Latvian Women Writers' Short Fiction (Kārkla and Matulis, 2022) are *jauns* (young) and *svešs* (strange).

Frequency lists can also be created from document types. For example, researchers can find in which research fields word *heart* is more common by checking distribution in PhD thesis or find which deputy most often mentions *inflation* in parliamentary debates.

4.3 Timeline

Word frequencies over time can be put in a graph to create timelines. Timelines allow researchers to track the usage of words or phrases over time, revealing trends in language change, the rise or fall of certain topics, or shifts in discourse. Timelines are particularly valuable for diachronic studies, where researchers aim to investigate how language evolves or how public attention to certain issues changes over the time. Figure 5 shows the use of word *krīze* (crisis) in the articles from Latvian news portals.

The timeline feature allows for the analysis of both absolute and relative frequencies. To avoid misleading conclusions, researchers can filter out periods with limited data. Corpora with available timeline functionality are tagged as diachronic on Korpuss.lv. These corpora include news, parliamentary debates, legal acts, PhD thesis, and others.

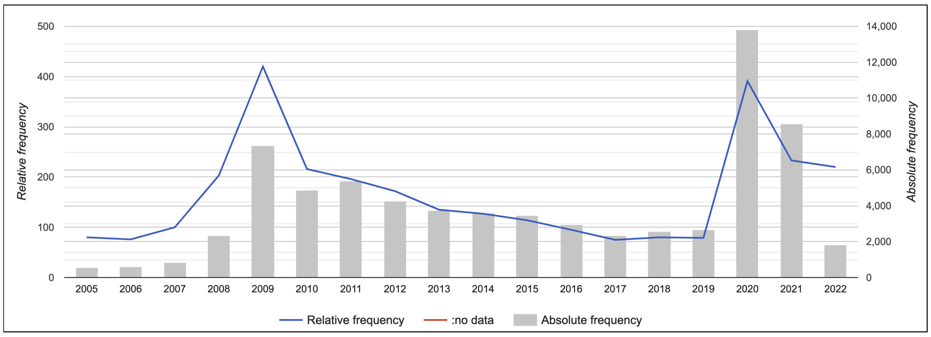


Fig. 5. Timeline function: the frequency change over time for word *krīze* (crisis)

5 Impact

The domain name Korpuss.lv was initially registered in 2007 to provide information about the first version of the Balanced Corpus of Modern Latvian. In May 2018, the platform evolved into an index of multiple corpora, launching with a collection of ten corpora. The name Latvian National Corpora Collection (LNCC) was officially adopted in November 2022. The first corpus within the CLARIN repository was published in July 2020, and citation guidelines were incorporated in January 2023. As of now, the LNCC hosts 39 corpora, 29 of which are published in the CLARIN repository. Over the past year, the LNCC platform has recorded 6,600 users and 33,000 page views.

To assess the academic impact of the LNCC, a systematic analysis was conducted using the Google Scholar search engine[6]. Search terms were enclosed in quotation marks to ensure exact matches. The analysis revealed that Korpuss.lv has been cited in over 200 scholarly works since 2020. Despite the relative novelty of the name, Latvian National Corpora Collection has been referenced in 18 English-language publications and 8 Latvian-language publications.

We recommend that authors cite LNCC resources using the designated publication or CLARIN data reference when available. However, instances of direct citation using Korpuss.lv URLs persist. Specifically, URLs linking to corpus information pages on Korpuss.lv have been cited in 37 scholarly works, while CLARIN URLs have been cited in 81 scholarly works.

The use of CLARIN URLs is particularly advantageous, as corpus codes and names in Latvian or English are often too generic and can lead to false positive matches in search results. Standardized citations ensure greater precision and facilitate the accurate identification and attribution of resources in scholarly contexts.

[6] Google Scholar: https://scholar.google.com/

6 Conclusion

In this paper, we have outlined our efforts to make corpora more accessible and useful for researchers in digital humanities and linguistics through the development of the Latvian National Corpora Collection (LNCC). By standardizing annotation formats, enhancing resource discoverability on Korpuss.lv, and enabling advanced functionalities such as federated content search and noSketch Engine integration, we have addressed several challenges that hinder corpus-based research. These initiatives lower technical barriers and promote the reproducibility and scalability of linguistic studies.

The availability of 39 corpora, including 2.8 billion tokens accessible through federated search, significantly broadens the scope for linguistic analysis. Tools such as frequency lists, concordances, and timelines empower researchers to explore language patterns and historical trends in greater detail. Our continued curation of the LNCC, alongside educational efforts such as seminars and learning materials, will further support corpus-based research across diverse academic disciplines.

Moving forward, our focus will be on expanding the collection, improving platform usability, and providing resources that allow researchers to fully harness the potential of corpus-based studies in language research.

Acknowledgements

This work was supported by the State Research Programme project DHELI (grant agreement No. VPP-IZM-DH-2022/1-0002) in synergy with the EU Recovery and Resilience Facility project Language Technology Initiative (grant agreement No. 2.3.1.1.i.0/1/22/I/CFLA/002).

References

Borin, L., Forsberg, M., Roxendal, J. (2012). Korp – the corpus infrastructure of språkbanken, *Proceedings of LREC 2012. Istanbul: ELRA*, p. 474–478.

Kārkla, Z., Matulis, H. (2022). Corpus of latvian women writers' short fiction. CLARIN-LV digital library at IMCS, University of Latvia.
http://hdl.handle.net/20.500.12574/69

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The sketch engine: ten years on, *Lexicography* **1**, 7–36.

Machálek, T. (2020). KonText: Advanced and flexible corpus query interface, *in* Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 7003–7008.
https://aclanthology.org/2020.lrec-1.865

McEnery, T., Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*, Cambridge University Press.

Paikens, P., Pretkalnina, L., Rituma, L. (2024). A computational model of latvian morphology, *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, p. 221.
https://aclanthology.org/2024.lrec-main.20

Saulite, B., Dargis, R., Gruzitis, N., Auzina, I., Levane-Petrova, K., Pretkalnina, L., Rituma, L., Paikens, P., Znotins, A., Strankale, L., Pokratniece, K., Poikans, I., Barzdins, G., Skadina, I., Baklane, A., Saulespurens, V., Ziedins, J. (2022). Latvian national corpora collection – korpuss.lv, *13th Language Resources and Evaluation Conference (LREC)*, pp. 5123–5129. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.548.pdf

Received December 2, 2024 , accepted December 10, 2024