The Infrastructure Behind Latvian National Corpora Collection

Roberts Dargis[1, 2], Baiba Valkovska[2]
University of Latvia, Raina bulvaris 19, Riga, Latvia
Institute of Mathematics and Computer Science, University of Latvia, Raina bulvaris 29, Riga, Latvia
{roberts.dargis, baiba.valkovska}@lumii.lv

Abstract

The rapid advancement of digital humanities and Natural Language Processing (NLP) necessitates centralized access to high-quality, large-scale language resources. This paper presents the technical infrastructure and evolving ecosystem of Korpuss.lv, the central access platform for the Latvian National Corpora Collection (LNCC). The LNCC consolidates 42 corpora developed by 14 institutions, comprising 2.8 billion tokens of written and spoken Latvian across diverse genres and annotation layers. Korpuss.lv has evolved from a simple metadata index into a comprehensive digital infrastructure that enhances corpus discoverability, accessibility, and usability for researchers in linguistics, digital humanities, and natural language processing. The platform integrates noSketchEngine as its primary corpus analysis tool and extends its functionality with custom modules, including a metadata-driven Corpora Explorer, a client-side Federated Content Search system, and precomputed UD-based Word Sketches. The ecosystem is further supported by CLARIN DSpace repositories for persistent storage and citation management, as well as a federated academic authentication architecture built on SATOSA and Keycloak via the CLARIN Service Provider Federation. The paper outlines architectural decisions, integration strategies, and future development plans.

Keywords: corpus infrastructure, federated authentication, corpus engine, CLARIN, Latvian corpora

1. Introduction

The rapid advancement of Natural Language Processing (NLP) and data-driven research in the digital humanities relies heavily on the availability of high-quality, large-scale language resources. For linguists, lexicographers, and AI developers, access to diverse and well-annotated corpora is essential for studying language variation, training models, and conducting quantitative textual analysis. However, language resources are often developed by independent institutions within separate, short-term projects. This frequently results in a fragmented digital ecosystem in which valuable data remain isolated, difficult to discover, or inaccessible due to incompatible formats and complex licensing restrictions.

To address these challenges, national and international infrastructures have emerged to consolidate language resources into centralized, standardized, and easily searchable platforms. For languages with smaller speaker populations and limited resources, such as Latvian, combining efforts within a unified national infrastructure is particularly important in order to maximize the impact and visibility of available data. A centralized ecosystem not only improves the discoverability of corpora but also provides standardized, user-friendly tools for querying and analyzing texts, reducing the need for extensive technical expertise among end users while properly managing academic access rights.

The Latvian National Corpora Collection (LNCC) is a diverse collection of corpora representing both written and spoken language. The collection contains 2.8 billion tokens of high-quality data totaling nearly 10 GB and covers a wide range of text types and genres, including news articles, social media posts, blogs, books, scientific texts, debates, and essays. The LNCC is a multi-institutional and multi-project initiative supported by the digital humanities and language technology communities in Latvia. Currently, it includes 42 corpora developed by 14 institutions.

A website called Korpuss.lv was developed to facilitate access to LNCC metadata and related services. Over time, Korpuss.lv has expanded in functionality, been extended with additional software components, and integrated with external resources. This paper focuses on the technical aspects of Korpuss.lv and its related ecosystem.

2. Related Work

Several projects are similar in scope. The most comparable are the CLARIN DSpace repositories (Straňák et al., 2020) established in various countries. CLARIN stands for Common Language Resources and Technology Infrastructure. A standard CLARIN DSpace installation hosts metadata records describing language resources and tools, and many of these records also provide access to the underlying data.

Some national CLARIN initiatives have expanded further or merged with related national activities, such as LINDAT/CLARIAH-CZ (Hajič et al., 2022), the Language Bank of Finland (The Language Bank of Finland, 1996), and Språkbanken CLARIN (Borin

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 44–48. Palma, Mallorca. May 11, 2026.

et al., 2012). These infrastructures have a broader scope, providing not only data but also computational infrastructure and tools for working with both hosted and user-supplied data.

The scope and focus of Korpuss.lv are to provide a curated collection of corpora together with browser-based tools preloaded with data, supporting the most common research use cases in linguistics and digital humanities. For additional use cases, links to download sites are provided for corpora that are available for download, allowing users to process the data with their own tools.

One of the most important components of the Korpuss.lv ecosystem is the corpus engine. Several open-source web-based corpus engines exist, each with its own strengths and limitations, such as noSketchEngine (Kilgarriff et al., 2014), Kontext (Machálek, 2020), and Korp (Borin et al., 2012). We use noSketchEngine because, when it was first deployed in 2017, it was the most feature-rich open-source web-based corpus engine available. Although alternative solutions have since matured, our user community is familiar with noSketchEngine, and migrating to a different platform is currently not a viable option. In the future, we may consider running an additional corpus engine in parallel if it offers distinctive features not available in noSketchEngine and demonstrates clear demand among users.

3. Korpuss.lv

Korpuss.lv[1] is a centralized digital infrastructure that hosts the LNCC. Initially developed as a simple index linking to external corpus platforms, it has evolved into an ecosystem comprising multiple interconnected components designed to improve corpus discoverability within Latvian research communities. The platform provides a user-friendly interface with filtering and sorting tools, federated content search across multiple corpora, and UD-based Latvian word sketches.

Our development strategy prioritizes the integration of established software solutions rather than building systems from scratch. However, as new use cases emerged for which no suitable ready-made tools were available, we developed additional modules to complement Korpuss.lv.

Korpuss.lv is implemented using Django[2], a high-level Python web framework. Python is widely used among NLP researchers, which facilitates the integration of existing NLP libraries when needed.

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

[2] Django framework – https://www.djangoproject.com/

3.1. Corpora Explorer

The Corpora Explorer serves as the main entry point to Korpuss.lv and presents the complete LNCC index through a filterable and sortable interface. Metadata-based filtering operates along three orthogonal dimensions: modality, distinguishing written text from speech; corpus type, distinguishing general-purpose from domain-specific corpora; and annotation level, including morphological, syntactic, error, manual, or diachronic annotation layers. Additionally, corpora with shared thematic provenance are grouped under labels such as historical, literary, or newspaper collections. This classification scheme is informed by conventions established by the Czech National Corpus project (Machálek, 2020) and the CLARIN resource family taxonomy (Fišer et al., 2018). The interface supports sorting chronologically by earliest data, reverse chronologically by most recent data, and by date of last update.

Each corpus is represented by a card displaying its identifier code, full name, and developing institution or institutions. The individual corpus page provides extended metadata, including associated publications, recommended citation formats derived from persistent identifiers assigned by the CLARIN DSpace repository (Section 5), and links to external download locations for corpora distributed under open or academic licenses.

3.2. Federated Context Search (FCS)

The Federated Content Search component enables simultaneous querying across all registered noSketchEngine corpus endpoints, returning both absolute and relative frequencies of the search term for each corpus. This functionality is particularly useful for identifying corpora that contain rare linguistic phenomena before performing detailed concordance analysis.

FCS is implemented as a client-side JavaScript application that sends requests directly from the user's browser to individual corpus engine endpoints, bypassing the Korpuss.lv application server. This architecture eliminates backend load and reduces latency associated with server-side proxying, providing a responsive asynchronous user experience as results arrive incrementally. The architectural trade-off is that all participating endpoints must expose permissive Cross-Origin Resource Sharing (CORS) headers. Endpoints without CORS support would require proxying through an intermediary server. All noSketchEngine instances within the Korpuss.lv ecosystem are configured accordingly. Result aggregation is performed in the browser. For each endpoint, the application issues a CQL query, parses the JSON response, and renders per-corpus frequency statistics in a unified table.

3.3. Word Sketches

The Word Sketch service[3] provides collocation and grammatical relation profiles for lemmas occurring in UD-parsed corpora. Unlike the dynamic word sketch computation available in the commercial Sketch Engine, the Korpuss.lv implementation relies on precomputed data structures. This design choice enables near-instantaneous query responses and minimizes runtime computational load.

Sketches are derived from dependency parses produced according to the UD annotation scheme (de Marneffe et al., 2021). The underlying data model is organized hierarchically across three levels: word (lemma), relation (typed syntactic dependency such as `nsubj`, `obj`, or `amod`), and collocation (a co-occurring lemma within a given relation). Sparsity and annotation noise are mitigated through a two-stage pruning strategy. Collocations with fewer than 10 occurrences are removed, and any relation node without remaining collocations, as well as any lemma node without remaining relations, is recursively removed. The resulting data are serialized and indexed for efficient lookup by lemma and part-of-speech tag.

4. Corpus Engine

The primary analytical tool within the Korpuss.lv ecosystem is the corpus engine. We use noSketchEngine, an open-source corpus management and analysis platform designed to support exploration of large text collections through a web interface. It represents a limited version of the commercial Sketch Engine system and is built on core components including Manatee for indexing and fast retrieval, Bonito for the graphical interface, and Corpus Query Language (CQL) for advanced search operations. The platform enables complex concordance searches, frequency list generation, and timeline-based analysis of language change. Users can perform simple searches by word form, lemma, or part-of-speech tag, or construct more advanced CQL queries combining multiple linguistic attributes and metadata filters.

Although it does not include automated word sketches or dictionary-building tools available in the commercial version, noSketchEngine provides robust support for investigating syntactic structures, semantic relationships, discourse patterns, and diachronic variation, making it well suited for research in linguistics, lexicography, and digital humanities.

We provide access to various types of corpora through our noSketchEngine instance, including learner corpora, parallel corpora, and speech event corpora. To ensure a consistent user experience, we aim to morphologically annotate all corpora that are not manually annotated using the same morphological annotator (Paikens et al., 2024). We have also automatically morphologically annotated a phonetic corpus containing word-level phonetic annotations (Auziņa et al., 2024). noSketchEngine uses tab-separated vertical files, a format also supported by the morphological annotator. For corpora containing manually annotated layers such as phonetic transcriptions, we convert the original data into vertical format using a custom processing pipeline and then generate morphological feature columns using the annotator. In rare cases, annotation accuracy may be slightly reduced when tokenization does not fully match the internal tokenization of the morphological annotator. However, this limitation is outweighed by the benefit of integrating morphological and manually annotated layers within the same corpus.

We are also exploring the use of Universal Dependencies (de Marneffe et al., 2021) and are gradually applying UD parsing to the corpora (Znotiņš, 2026). Although noSketchEngine does not natively support querying over dependency relations, the inclusion of UPOS and deprel layers provides valuable information for linguistic research, even for languages with relatively free word order such as Latvian.

noSketchEngine does not allow users to view entire documents when they exceed the maximum context window. This limitation makes it possible to provide broad access even when corpora are distributed under academic licenses with copyright restrictions. For corpora strictly restricted to academic users, we operate a separate noSketchEngine instance protected by academic authentication.

We plan to develop a dedicated document viewer that will be linked from corpus metadata within noSketchEngine. This viewer will expand research possibilities, especially for multimodal corpora containing images, audio, or video, such as speech, sign language, or OCR corpora. Academic authentication will also be supported in the document viewer for restricted corpora.

5. CLARIN

CLARIN (Common Language Resources and Technology Infrastructure)[4] is a European digital infrastructure providing sustainable access to language data and tools. It connects certified centers across multiple countries and offers repository services as well as federated authentication mechanisms. Within the Korpuss.lv ecosystem, CLARIN fulfills two essential technical roles: persistent and citable

[3] LVK2022 Word Sketches – https://korpuss.lv/skices

[4] CLARIN – https://www.clarin.eu

storage of language resources, and controlled access to restricted datasets through a federated identity framework.

The Service Provider Federation connects CLARIN-registered service providers to national identity federations across EU member states. This model allows academic users to access password-protected resources using their home institution credentials without registering separate accounts for each service. For providers, it ensures that access can be restricted to verified academic users while offering single sign-on convenience to end users through their existing institutional login.

CLARIN data repositories serve as the persistent storage and cataloging backbone of the infrastructure. They host structured metadata records describing corpora and tools, assign persistent identifiers to datasets, and often distribute data files directly. Persistent identifiers are essential for scholarly reproducibility because they enable unambiguous citation of specific corpus versions. When data are distributed under academic licenses, access is mediated by the federated authentication framework, which verifies institutional affiliation before granting download permissions. The most widely used repository platform within the CLARIN network is CLARIN DSpace[5].

Latvia operates its own national CLARIN DSpace instance (Skadiņa et al., 2020). The Korpuss.lv backend harvests metadata and associated persistent identifiers to automatically generate standardized citation recommendations on corpus detail pages. If data are hosted in the repository, a download button linking to the repository is also displayed.

6. Academical Authentication

Several corpora within the LNCC are restricted to academic users affiliated with recognized institutions. To enforce these restrictions, we required a federated authentication solution compatible with eduGAIN[6], the interfederation service linking national research and education identity federations. Direct membership in eduGAIN is not available to individual organizations, and participation must be mediated through a national federation. For this purpose, we use the CLARIN Service Provider Federation.

Typically, each web application requiring federated access must be registered independently as a service provider, which involves administrative overhead and manual review. To avoid registering each Korpuss.lv component separately, we implemented a single identity proxy registered once with the CLARIN Service Provider Federation that forwards authentication to downstream applications. For the application-facing layer, we use Keycloak[7], an open-source identity and access management platform responsible for session management and token issuance. However, Keycloak does not natively support SAML discovery services and cannot automatically ingest metadata from external identity provider registries, requiring manual configuration of each upstream provider.

To address this limitation, we introduced an intermediate proxy between Keycloak and eduGAIN that supports dynamic discovery and presents itself to Keycloak as a single unified identity provider. We evaluated SimpleSAMLphp[8] and SATOSA[9], selecting SATOSA primarily because it is implemented in Python, aligning with the broader Korpuss.lv technology stack and reducing operational complexity. In this architecture, SATOSA manages SAML discovery and federates outward to eduGAIN, while presenting a single SAML identity provider endpoint to Keycloak, which secures individual Korpuss.lv services. This layered design allows a single federation registration while maintaining flexibility to add new protected services with minimal configuration effort.

7. Conclusion

This paper presented the technical infrastructure and evolving ecosystem of Korpuss.lv, the central access point for the Latvian National Corpora Collection. By consolidating diverse corpora from multiple institutions into a unified platform, Korpuss.lv has significantly lowered the barrier to entry for researchers in linguistics, digital humanities, and natural language processing. Its evolution from a simple directory to a comprehensive digital ecosystem ensures that Latvian language resources are discoverable, accessible, and sustainable.

The development of Korpuss.lv demonstrates the value of integrating robust open-source solutions such as noSketchEngine and CLARIN DSpace with custom modules tailored to specific research needs. The implementation of an academic authentication gateway based on SATOSA and Keycloak effectively bridges institutional identity federations such as eduGAIN with secure access to academically licensed corpora.

The steady growth in user engagement and increasing academic citations referencing Korpuss.lv and the LNCC highlight the platform's importance within the research landscape. The next major development step is the creation of a dedicated document viewer with academic authentication support,

[5] CLARIN DSpace – https://github.com/ufal/clarin-dspace

[6] eduGAIN – https://edugain.org/

[7] Keycloak – https://www.keycloak.org/

[8] SimpleSAMLphp – https://simplesamlphp.org/

[9] SATOSA – https://github.com/IdentityPython/SATOSA

enabling full-text access and improved support for multimodal corpora containing audio, video, and images. Continued refinement of the Korpuss.lv ecosystem will further empower researchers and contribute to safeguarding the digital future of the Latvian language.

8. Acknowledgements

This work was supported by the EU Recovery and Resilience Facility project Language Technology Initiative (2.3.1.1.i.0/1/22/I/CFLA/002) in synergy with the State Research Programme Letonika – Fostering a Latvian and European Society project Digital Resources and AI Technologies for the Sustainability of the Latvian Language (DigiLATE) (VPP-IZM-LETONIKA-2025/1-0004).

9. Bibliographical References

Ilze Auziņa, Normunds Grūzītis, Roberts Darģis, Guna Rābante-Buša, Didzis Goško, Jānis Vempers, Raivis Kivkucāns, and Artūrs Znotiņš. 2024. Recent latvian speech corpora for linguistic research and technology development. *Baltic Journal of Modern Computing*, 12(4):646–658.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp — the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 474–478, Istanbul, Turkey. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Hajič, Eva Hajičová, Barbora Hladká, Jozef Mišutka, Ondřej Košarko, and Pavel Straňák. 2022. Lindat/clariah-cz: Where we are and where we go. *CLARIN. The infrastructure for language resources. Berlin: De Gruyter*.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1:7–36.

Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Pēteris Paikens, Lauma Pretkalniņa, and Laura Rituma. 2024. A computational model of Latvian morphology. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 221–232, Torino, Italia. ELRA and ICCL.

Inguna Skadiņa, Ilze Auziņa, Normunds Grūzītis, and Artūrs Znotiņš. 2020. Clarin in latvia: From the preparatory phase to the construction phase and operation. In *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN)*, pages 342–350.

Pavel Straňák, Ondřej Košarko, and Jozef Mišutka. 2020. Clarin-dspace repository at lindat/clarin. *Grey Journal (TGJ)*, 16.

The Language Bank of Finland. 1996. Kielipankki – The Language Bank of Finland. https://www.kielipankki.fi. University of Helsinki and CSC – IT Center for Science.

Artūrs Znotiņš. 2026. Pretraining and benchmarking modern encoders for Latvian. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages*. ACL. To appear.