Korpuss.lv – a versatile platform for digital humanities

Roberts Darģis, Baiba Saulīte

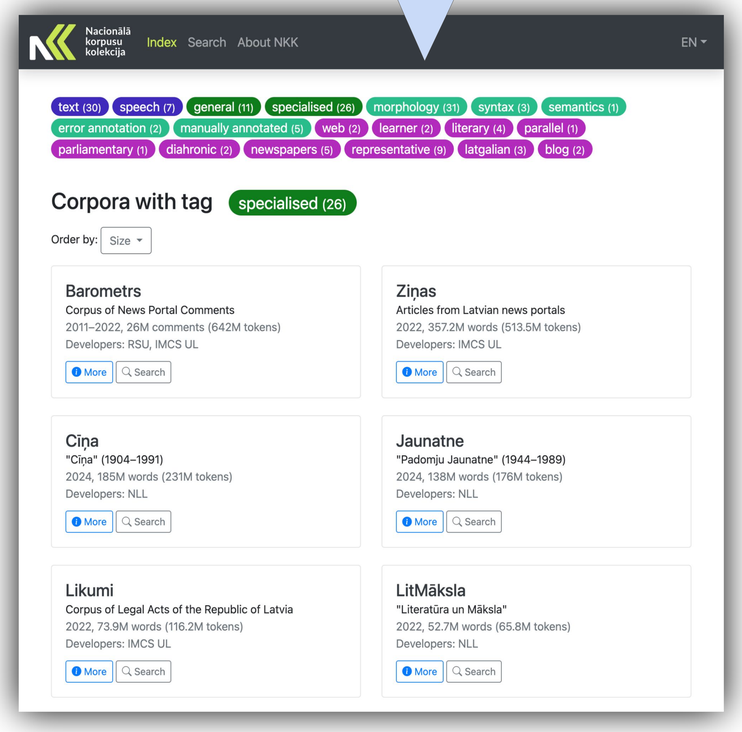
AILab, IMCS, University of Latvia

Latvian National Corpora Collection (LNCC) is an extensive and diverse collection comprising over 35 text and spoken corpora (totaling 2.8 billion tokens), from two corpus indexer instances (endpoints) maintained by Institute of Mathematics and Computer Science, University of Latvia, and National Library of Latvia. This collection represents a wide range of text types, such as news, blogs, scientific texts, parliamentary debates, colloquial texts, essays and early texts. Notably, almost all corpora in LNCC are re-annotated with a uniform morpho-syntactic annotation scheme, enabling federated search and consistent linguistic analysis across different types and genres.

korpuss.lv

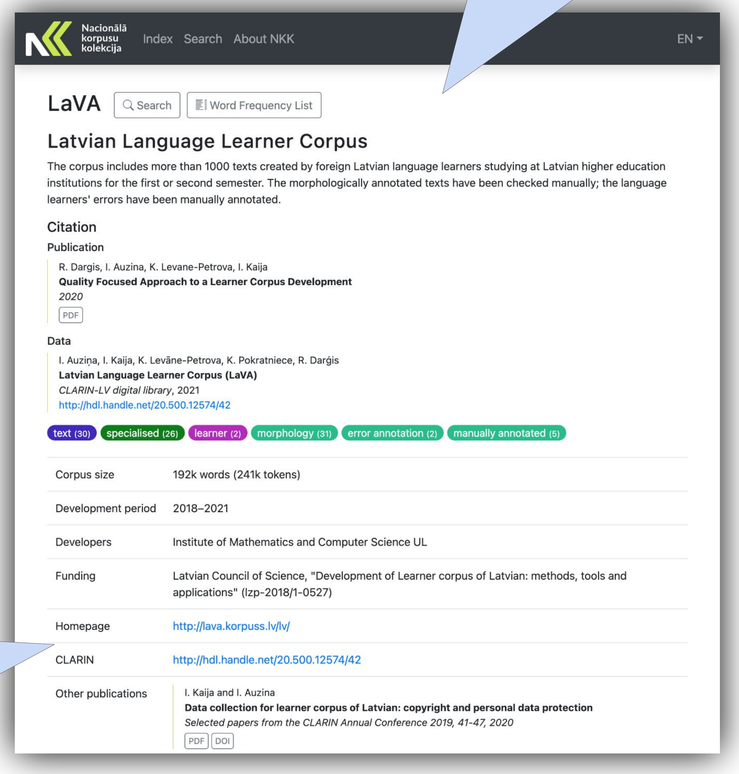
- A specialized corpora website and query platform hosting LNCC
- Facilitates efficient work with multiple corpora, providing easy access and detailed information on a diverse array of Latvian and Latgalian language corpora

Filter corpora by **tags** and sort them

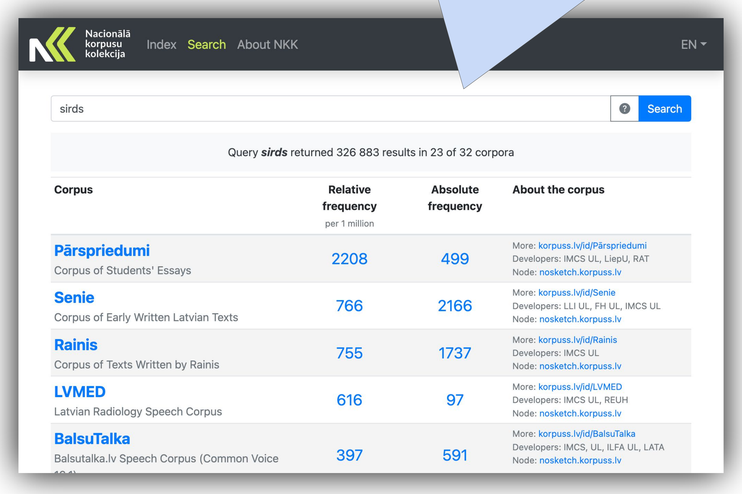


Information about the corpus and more:

- Extensive **metadata**
- References for **citation**
- Relevant **publications**



Find relevant corpora, term relative and absolute frequency by searching in all corpora at once, using **Federated Content Search (FCS)**

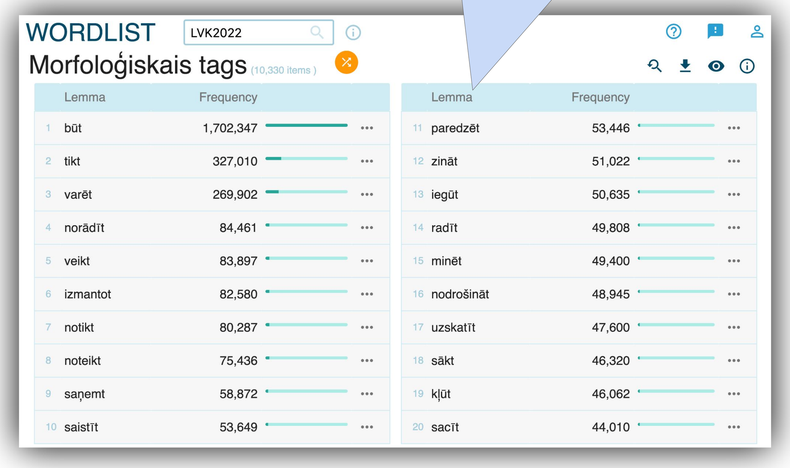


Most of the corpora can be **downloaded** from corpus homepage or CLARIN repository

NoSketch Engine

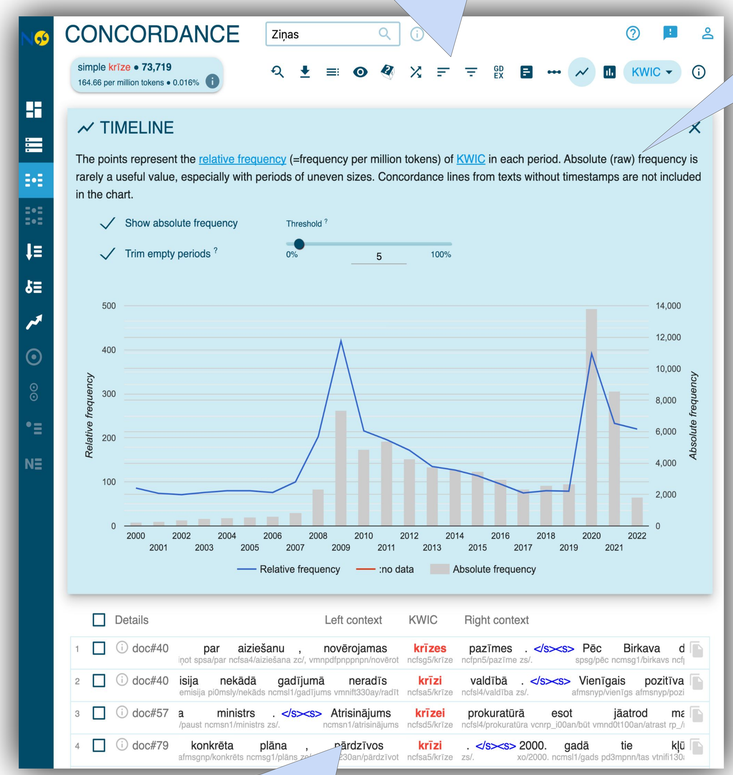
- A corpus management platform facilitating extensive corpus analysis: from frequency lists to condordanes and timelines
- Almost all corpora are morphologically annotated

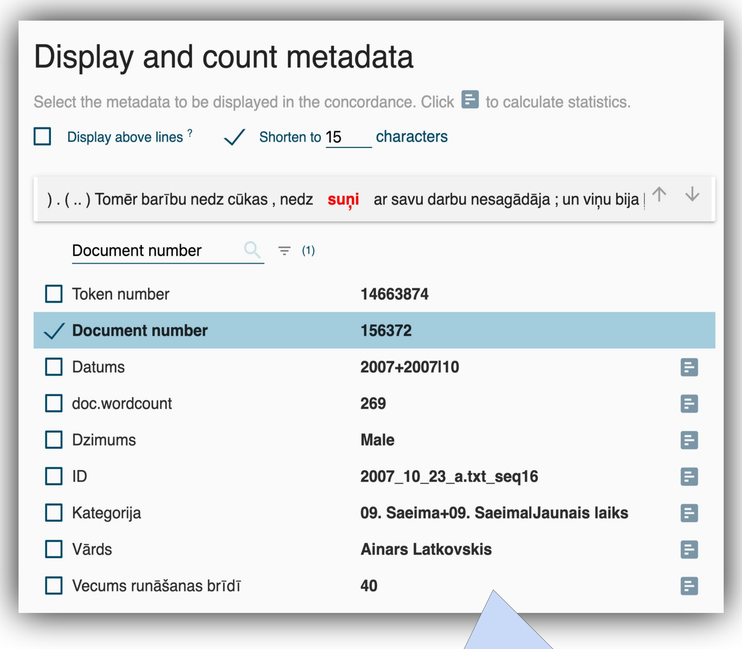
Calculate and download **word frequency** list



Configure and **download** result

Timeline of relative and absolute frequencies over years for supported corpora





Concordances with lemmas, morphological and syntactic annotations if available

Reference and other **metadata** about each concordance

Institute of Mathematics and Computer Science University of Latvia
LU Literatūras, folkloras un mākslas institūts
RTA Rēzeknes Tehnoloģiju akadēmija
Rīgas Stradiņa universitāte
LNB Latvijas Nacionālā bibliotēka
Latvijas Valodas institūts
Latvijas Universitāte Humanitāro zinātņu fakultāte
Liepājas Universitāte
Vytauto Didžiojo Universitetas
LETA nacionālā informācijas aģentūra
aslimnīca
TILDE
LATA

2027 Nacionālais attīstības plāns
Finansē Eiropas Savienība NextGenerationEU
EU Recovery and Resilience Facility project "Language Technology Initiative" (2.3.1.1.i.0/1/22/I/CFLA/002; 2023–2026)

Nacionālais attīstības plāns 2020
Eiropas Savienība Eiropas Reģionālās attīstības fonds
Ieguldījums tavā nākotnē
European Regional Development Fund (1.1.1.5/18/I/016; 2018–2020)
European Regional Development Fund (1.1.1.1/16/A/219; 2017–2019)

VPP Valsts pētījumu programma
National Research Programme "Digital Humanities" (VPP-IZM-DH-2022/1-0002; 2022–2025)
National Research Programme "Digital Resources of the Humanities" (VPP-IZM-DH-2020/1-0001; 2020–2022)

Latviešu valodas aģentūra
Funding for the development of the corpus conception, Balanced Corpus of Modern Latvian, etc. (2005–2022)