Authors:
Anda Baklāne, Valdis Saulespurēns

LATVIAN PROSE COUNTER: From Digitized Books to Data Visualizations

The Latvian Prose Counter (LPC) is a multifaceted digital platform that showcases the potential of digital text analysis and visualization, provides comprehensive insights into Latvian novels from the 19th and 20th centuries, and serves as an experimental hub for full-text and metadata analysis of these novels. This initiative is a collaborative effort between the National Library of Latvia (NLL), the Institute of Literature, Folklore, and Art of the University of Latvia (ILFA), aiming to synergize the resources of both institutions to forge a comprehensive digital resource. The morphological and syntactical markup of texts is realized by using NLP tools created by the Institute of Informatics and Mathematics of the University of Latvia. New literary works and new works, functionalities, and research examples are regularly added to the platform.

1. DATA PROCESSING WORKFLOW

PREPROCESSING

- Digitization
- OCR finetuning
- Metadata
- Data warehouse

ADAPTATION

- Lemmatization
- POS tagging
- NER tagging
- Word2Vec

LV-PIPE

- Public: nlp.pipe.lv
- Mass tagging: local
- Partner institution: AiLab at IMCS UL

POSTPROCESSING

- Pandas DataFrame - JSON
- Local Python Jupyter Notebooks

- Exploratory data analysis
- Aggregation, validation
- Archive DataFrame – Parquet

READY TO SERVE VISUALIZATIONS

Interactive
Plotly diagrams
https://proza.lnb.lv/

JUPYTER NOTEBOOKS FOR REUSE

- Library's private collection
- Public selection externally at GitHub

DATA EXPORT FOR REUSE:

- JSON
- tsv

2. DATA HYDRATION IN PROZA.LNB.LV

DATA

Prepared
JSON, TSV

EDITING

Markdown templates for content editing

VISUALIZATIONS

Premade Plotly visualization as embeddable HTML fragments

Back-end architecture

- Static 300+pages
- Hugo based
- Refreshed daily

- NLL hosted
- Ubuntu 22.04
- Traefik/NGINX server

Front-end technologies

- Tailwind for CSS
- Minimal jQuery DataTables plugin
- Plotly (d3 based) visualizations

PARTNERS

- Image assets
- Descriptions
- Additional metadata
  Partner institution: ILFA UL

„Dahrgais tehws, mehs uſwarejàm!"

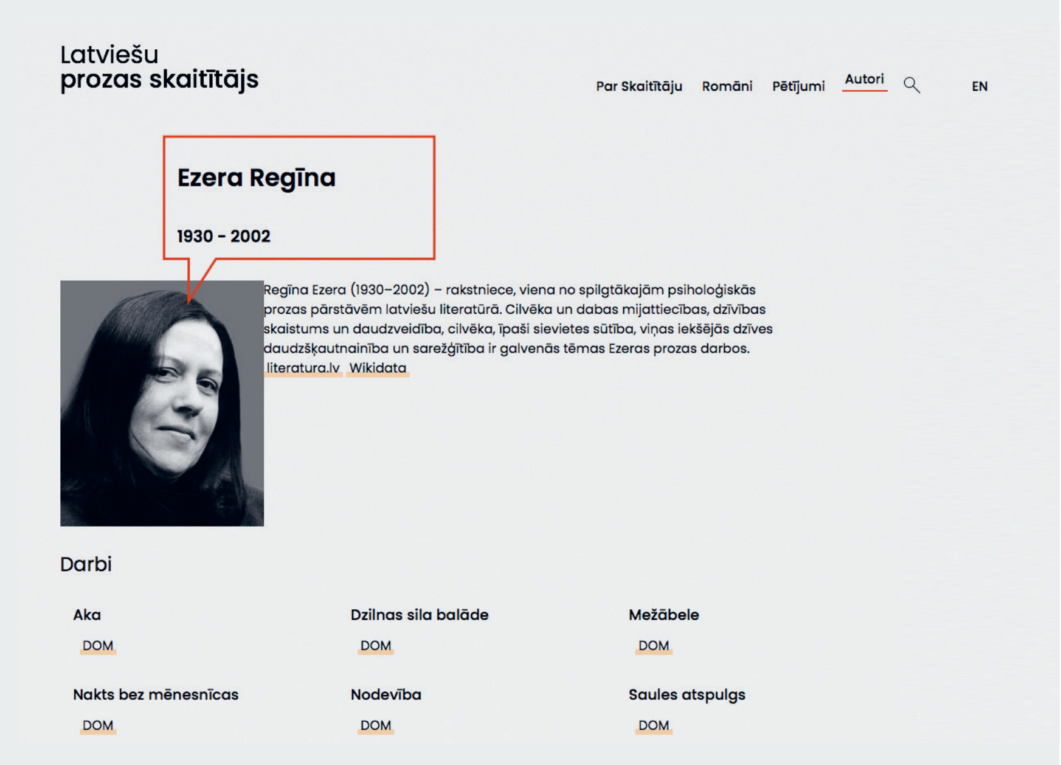
„Gods un ꞩlawa jums!" ꞩirmais kungs eeꞩauzàs.

„Lai dſihwo augſti muhꞩu wadons un warons!" wiꞩi bruņineeki ꞩauza, pazeldami augſtu gaiꞩâ ꞩawus ſobenus.

„Dārgais tēvs, mēs uzvarejām!"

„Gods un slava jums!" sirmais kungs iesaucās.

„Lai dzīvo augsti mūsu vadons un varons!" visi bruņinieki sauca, paceldami augstu gaisā savus zobenus.



Latviešu
prozas skaitītājs

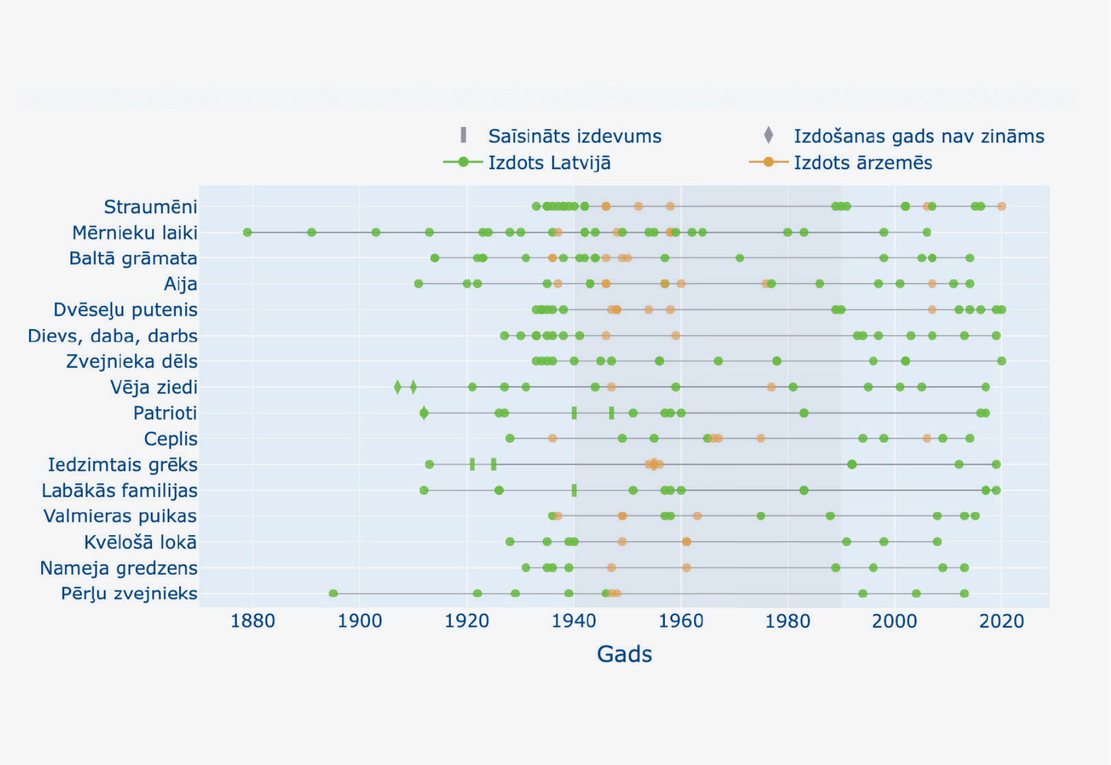
Ezera Regīna

1930 - 2002

Regīna Ezera (1930–2002) – rakstniece, viena no spilgtākajām psiholoģiskās prozas pārstāvēm latviešu literatūrā. Cilvēka un dabas mijattiecības, dzīvības skaistums un daudzveidība, cilvēka, īpaši sievietes sūtība, viņas iekšējās dzīves daudzšķautnainība un sarežģītība ir galvenās tēmas Ezeras prozas darbos.
literatura.lv Wikidata

Darbi

Aka DOM	Dzilnas sila balāde DOM	Mežābele DOM
Nakts bez mēnesnīcas DOM	Nodevība DOM	Saules atspulgs DOM



Novels

The Latvian Prose Counter displays data on Latvian novels of the 19th and 20th centuries. The most prominent part of the collection is LatSenRom (1879-1940) - a comprehensive corpus of Latvian long prose fiction, covering all novels released in Latvian as books from 1879 to 1940 (approx. 460 works when it will be finalized). LatSenRom encompasses several iterations: original text, typeface-normalized text, and grammar-normalized text. In addition to that, it is available in a structured form with morphological, syntactic, and NER markup. The corpus is available to users as an individual dataset, or it can be explored through the interfaces - the corpus analysis platform (korpuss.lnb.lv) and a website Latvian Prose Counter.

Authors

The "Authors" section of the website features profiles of individual writers. Currently, the main objects of interest are diagrams that display the most frequently used words with a parts of speech filter, as well as statistics on sentence length and lexical diversity. The portrait and short description in the header are sourced from the literatura.lv database, which holds information on prominent figures in the cultural field. Links to the entries of the novels in the digital library are also provided. https://proza.lnb.lv/en/authors/.

Research

The "Research" section showcases research projects based on Latvian novels. It includes links to published academic papers and presents engaging visual explorations. Featured research examples include an exploratory diagram that displays data points of authors' lifespans alongside the publication dates of their novels. Other project examples are a study focused on lexical diversity within Latvian novels and an analysis of reprint data from early Latvian novels.

Acknowledgements
The project "Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia", implemented within the framework of the National Research Programme "Digital Humanities" (No: VPP-IZM-DH-2022/1-0002)

Institute of Literature, Folklore, and Art of the University of Latvia
Artificial Intelligence Laboratory (AiLab) at IMCS, University of Latvia

Contact: dh@lnb.lv