The Balanced Corpus of Modern Latvian (LVK)

Kristīne Levāne-Petrova
kristine.levanepetrova@lumii.lv
The Institute of Mathematics and Computer Science, University of Latvia

Design principles of LVK

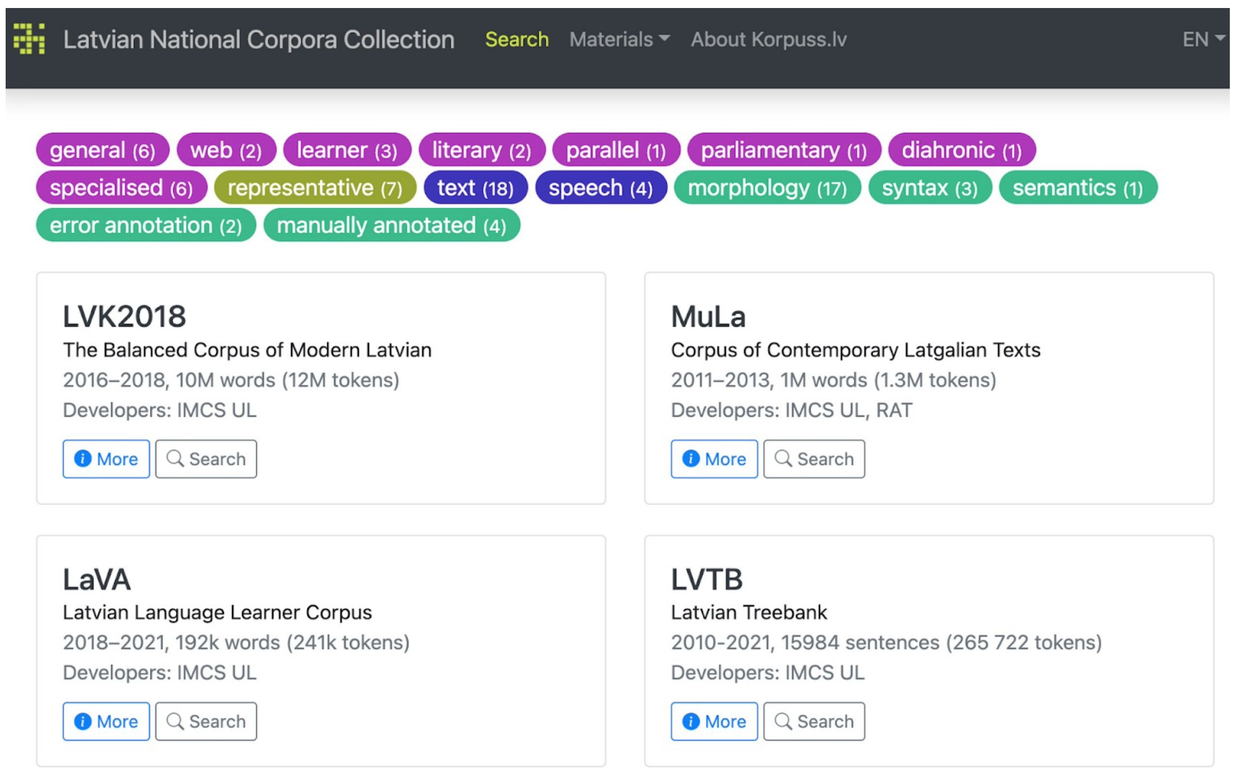
LVK has been designed since 2007 based on the on the Latvian Language Corpus Conception (2005).

LVK2018 is designed as a general-language, representative and publicly available cor- pus. It is a monolingual, fully morphologically and partly syntactically and semantically annotated corpus. Presently, it consists of 10 million tokens. Characteristics of LVK2018:

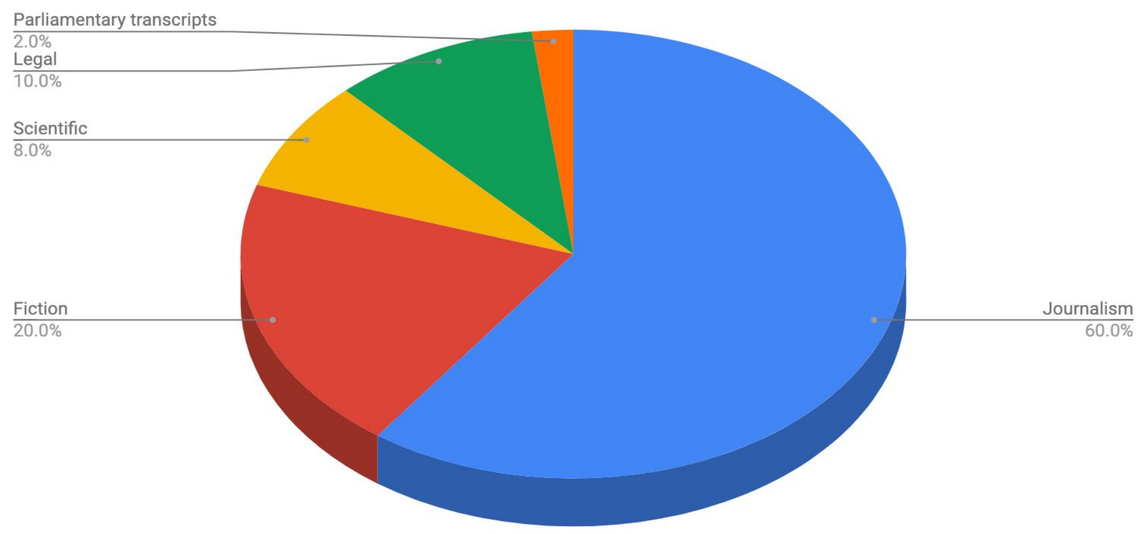
- **General** – the corpus includes sources from different domains, styles, genres, etc.
- **Balanced** – the corpus that aims to cover the variety of existing texts in estimated proportions.
- The corpus represents the **synchronic state** of the language.
- **Originality** – the corpus should only contain texts originally written in Latvian.
- The corpus is **representative**, it contains texts from all language styles, major domains and many subdomains.

Text selection criteria

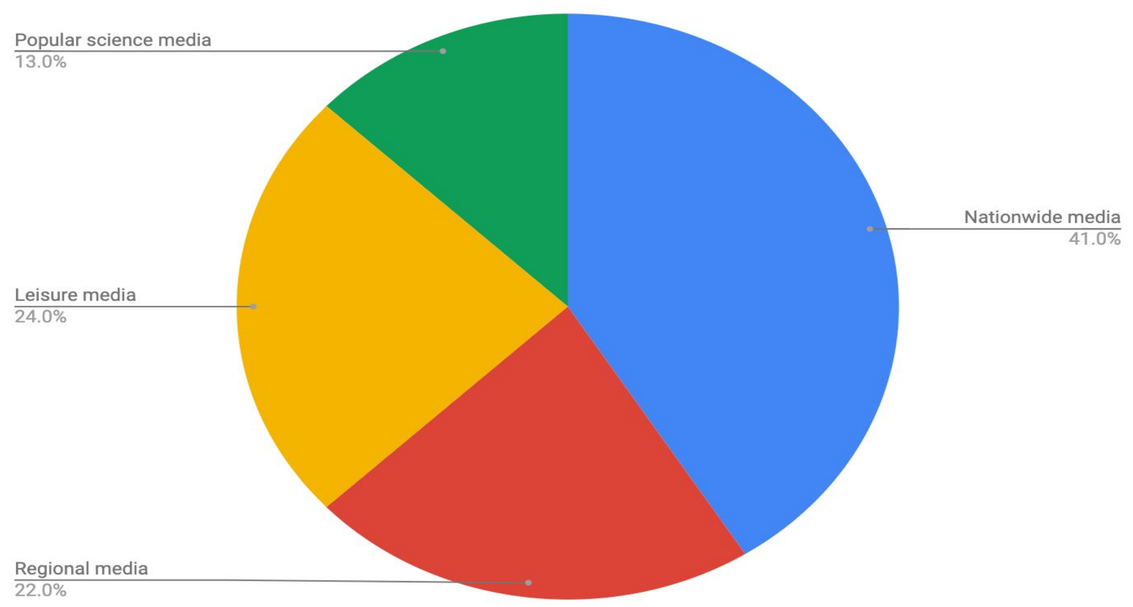
- **Time** – the corpus should contain texts created and published after 1991.
- The corpus should contain **full-text**.
- **Diversity** – texts should cover as wide range of topics as possible. The sample cannot exceed more than 5 % and 50,000 words of the particular section of the corpus to not dominate one author or domain in the corpus. If the text is longer than the limit, it should be cut from the end of the sample only.
- **Uniqueness** – the corpus sample should be represented in corpus just once.
- **Quality** – samples should only contain clean text written in literary language with appropriate usage of diacritics and punctuation in Latvian. Tables and other nontext parts should be removed.



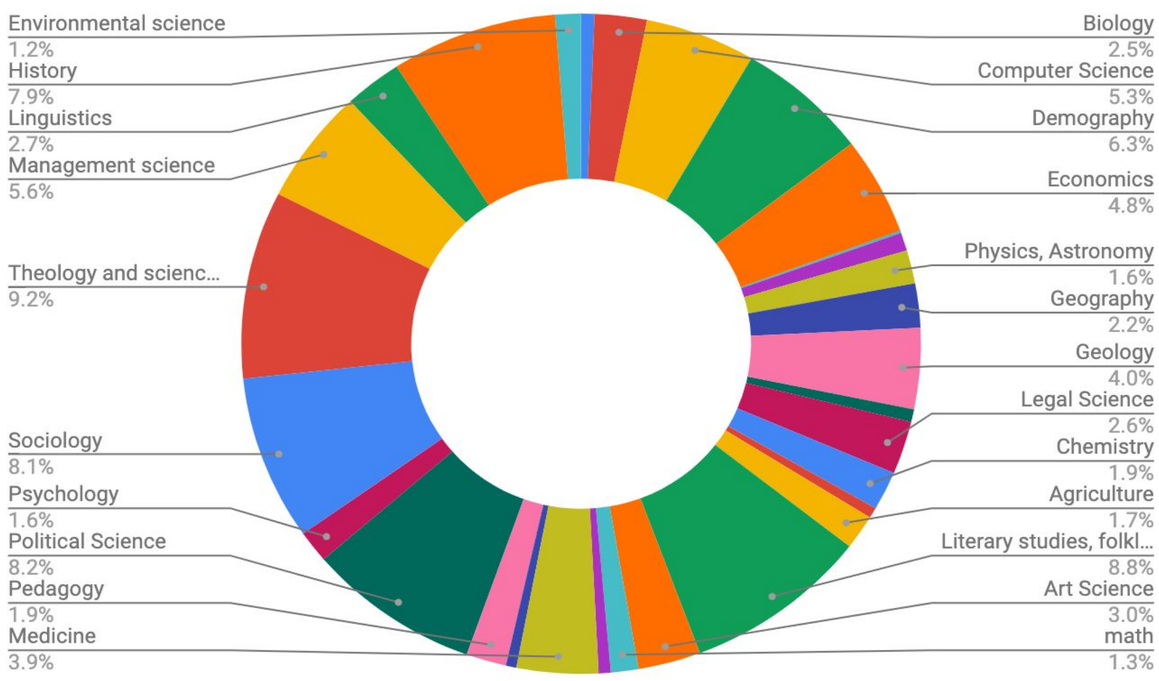
Composition of LVK2018



Composition of Journalism



Composition of Scientific texts



All LVK corpora and aubcorpora have been released in the framework of Latvian National Corpus and registered at the CLARIN-LV repository. LVK2018 and other corpora are freely available via the corpus query interface *NoSketch Engine*.

LVK2022

Institute of Mathematics and Computer Science
University of Latvia
CLARIN-LV
NATIONAL DEVELOPMENT PLAN 2020
EUROPEAN UNION
European Regional Development Fund
INVESTING IN YOUR FUTURE

The Latvian National Corpora Collection is available:
http://www.korpuss.lv

This work has received support from the "Strengthening of the capacity of doctoral studies at the University of Latvia within the framework of the new doctoral model", identification No. 8.2.2.0/20/I/006