Lessons Learned from Creating a Balanced Corpus from Online Data

Roberts DARĢIS, Kristīne LEVĀNE-PETROVA, and Ilmārs POIKĀNS
Institute of Mathematics and Computer Science, University of Latvia

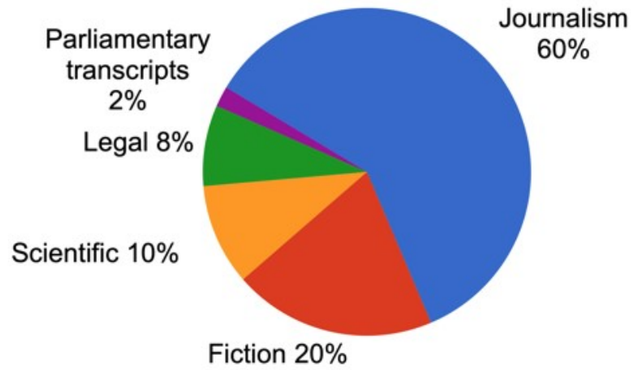
Abstract. Most of the new corpora are created from data obtained from various text holders, which requires cooperation agreements with each of the text holders. Reaching these cooperation agreements is a difficult and time consuming task. Developing a balanced corpus from various online sources do not require agreements with text holders, but it presents many more technical challenges, including text extraction, cleaning and validation.

Corpus Design Principles:

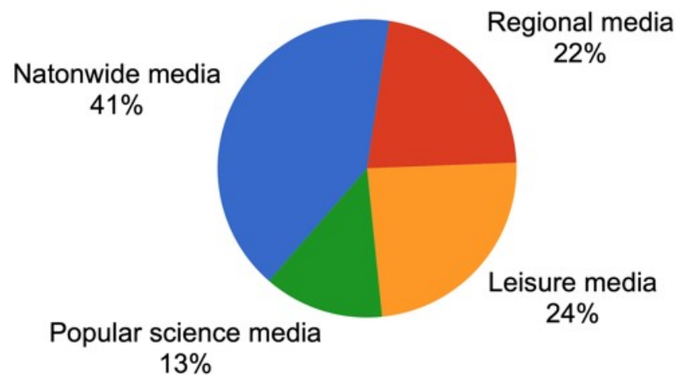
- **General** – the corpus includes sources from different domains, styles, genres, etc.
- **Balanced** – the corpus that aims to cover the variety of existing texts in estimated proportions.
- The corpus represents the **synchronic** state of the language. It covers sources as from the end of the last century until the present.
- **Originality** – the corpus should only contain texts originally written in Latvian. The obvious translations of the different texts into Latvian will not be included in LVK2018.
- The corpus is **representative**, it contains texts from all language styles, major domains and many subdomains.

www.korpuss.lv

Composition of LVK2018



Composition of the Journalism



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.
- **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 non-text parts should be removed.

Latviešu valodas aģentūra