From Words to Action: A National Initiative to Overcome Data Scarcity for the Slovene LLM

Špela Arhar Holdt (1), Špela Antloga (1, 2), Tina Munda (1), Eva Pori (1), Simon Krek (1, 3)

[1] University of Ljubljana, Slovenia
[2] University of Maribor, Slovenia
[3] Institut "Jožef Stefan", Slovenia

1) The project ***PoVeJMo—Adaptive Natural Language Processing with Large Language Models*** aims to develop large language models for the **Slovene language**.

2) LLMs rely on vast and diverse datasets, which poses challenges for languages with limited resources, such as Slovene, due to its **smaller speaker base and restricted data availability**.

3) To meet this challenge, we have commenced a **national initiative for a large-scale collection of Slovene texts.**

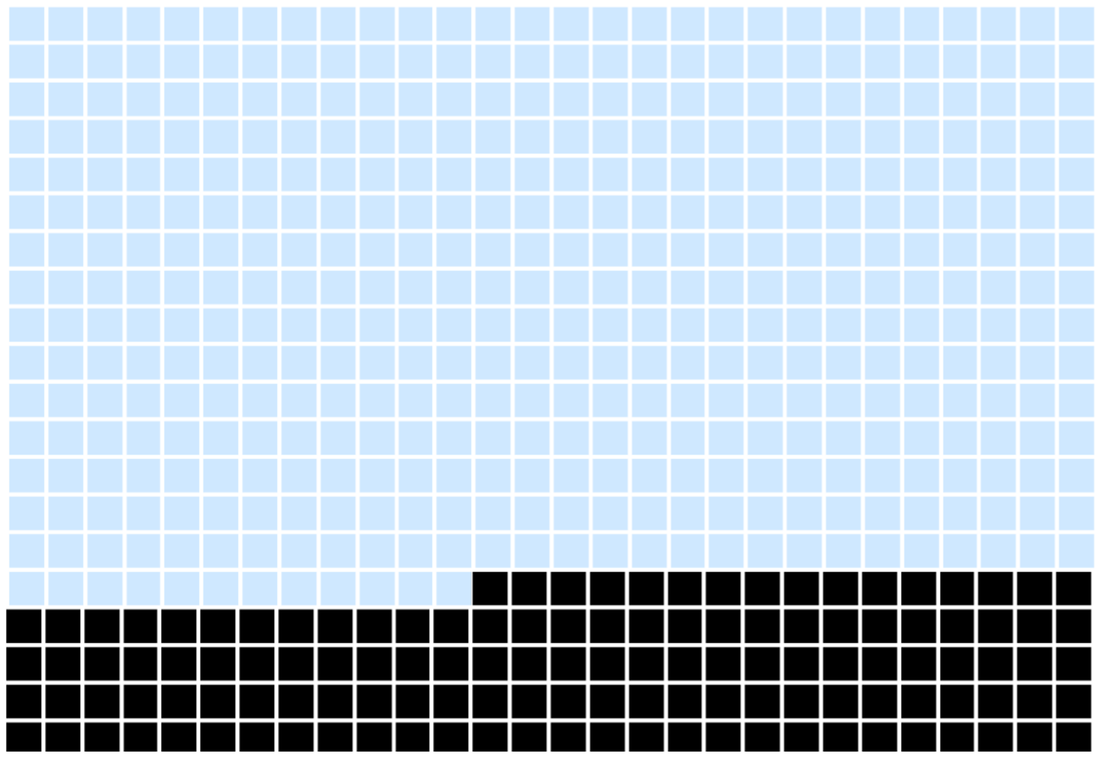
4) To overcome legal constraints, we have taken a more direct approach by **actively seeking permission from copyright holders** to use their texts.

5) The text collection campaign operates along **two main strategies:**

- We engage large-scale text providers such as national libraries, publishers, media organizations and government ministries.
- We engage individuals, inviting them to donate their own texts.

6) The data is received either via network connections or on portable storage devices. We have implemented a comprehensive protocol for **secure data management** at all stages of its lifecycle.

40 billion words



Web portal for text submission available at: zbiranje.povejmo.si

CURRENT STAGE

9,2 BILLION WORDS

Figure 1: The section of the web portal where the interested participants can provide their texts (https://zbiranje.povejmo.si/).

Figure 2: The section of the web portal where the interested participants can test the existing model (https://povejmo.si/klepet/).

Engaging institutions:

- Slovene language must remain comparable and competitive with other similar languages in the digital age.
- A collective effort.
- Independence from foreign corporations.

Engaging individuals:

- Radio, TV and social media presence.
- Web portal **povejmo.si** for text submission.
- Interactive section where users can try the current version of the language model. Try it at https://povejmo.si/klepet/

This poster was presented at the RESOURCEFUL-2025 Workshop at the NoDaLiDa/Baltic-HLT 2025 Conference (Tallinn, Estonia, 2. 3. 2025). Please refer to the corresponding short paper for further information, references, and contact details.

Affiliations:

FRI UNIVERSITY OF LJUBLJANA Faculty of Computer and Information Science

Institut "Jožef Stefan" Ljubljana, Slovenija

Univerza v Mariboru

Acknowledgements:

The research program Language Resources and Technologies for Slovene (P6-0411) and the projects Adaptive Natural Language Processing with Large Language Models and Large Language Models for Digital Humanities (GC-0002) are funded by the Slovenian Research and Innovation Agency.