Tweets in Karelian: from data collection to the content analysis

Ilia Moshnikov
Karelian Institute, University of Eastern Finland, Finland
ilia.moshnikov@uef.fi

Eugenia Rykova
University of Eastern Finland, Finland
Catholic University of Eichstätt-Ingolstadt, Germany
eugenryk@uef.fi

Paper Abstract

The internet in general and social media in particular offer a new domain for the use of minority languages, which is important from the perspective of language vitality and language revitalisation. In our presentation we focus on the visibility of the Karelian language on X (formerly known as Twitter). Karelian is an endangered minority language spoken in Russia and Finland. According to the latest census and research, the total number of Karelian speakers is roughly about 20,000 people (Sarhimaa 2017; Federal State Statistics Service 2021).

We present our data collection strategy based on the use of language-related keywords and hashtags. The data was scraped from X (Twitter) using the Postman API software (Postman, 2023). The multilingual dataset combines many different languages, with Finnish dominating. Our final data consists of 2625 entries written entirely or partially in Livvi, South and Viena Karelian. The language and Karelian dialects were labelled manually by the first author of the study, who is a native Livvi-Karelian speaker. The visibility of Karelian on X has increased significantly in recent years, with Livvi-Karelian being the most prominent dialect (Moshnikov and Rykova 2023). Automatic language detection (Jauhiainen et al. 2022) identified Livvi-Karelian (or a mix of dialects including Livvi-Karelian) as such with 99.7% sensitivity, and South Karelian and Viena Karelian as Livvi-Karelian with 90% and 73.8% sensitivity, respectively.

We also analysed the topics of Twitter (X) entries written in Karelian. Ten main topics were identified manually by close reading each entry. Since the data was collected using keywords and hashtags related to the Karelian language itself, most of the entries are related to the language and vocabulary in sense of translation or language learning. However, personal tweets are the most numerous among the original entries. Tweets about the status of the Karelian language and the process of language revitalisation are particularly interesting from a research perspective as well as individual use of the language. In our data it is also possible to analyse tweets about the Karelian language written in Finnish and Russian.

Keywords: automatic language recognition, data scraping, Karelian, minority languages, X (Twitter).

REFERENCES

Federal State Statistics Service. (2021). *Vserossijskaja perepis' naselenija 2020 [Russian Census 2020]*. https://rosstat.gov.ru/vpn/2020.

Jauhiainen, T., Jauhiainen, H., & Lindén, K. (2022). HeLI-OTS, Off-the-shelf language identifier for text. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3912–3922. Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.416/.

Moshnikov, I. & Rykova, E. (2023). Little Big Data: Karelian Twitter Corpus. *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023), 14–15 September 2023, University of Mannheim, Germany*, pp. 142–147. https://doi.org/10.14618/1z5k-pb25.

Postman. (2023). *Postman API Tool*. https://www.postman.com/.

Sarhimaa, A. (2017). *Vaietut ja vaiennetut. Karjalankieliset karjalaiset Suomessa [Silent and being forced to be silent: Karelian-speaking Karelians in Finland]*. Tietolipas 256. Helsinki: Finnish Literature Society.