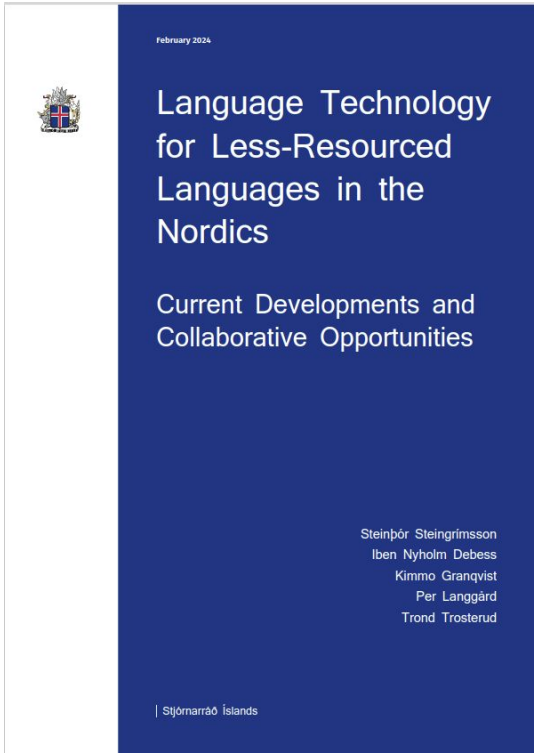
Árnastofnun

# Report on Language Technology for Less-resourced Languages in the Nordics: Current State and Prospects

Language Technologies and AI in the Nordic Countries
5.–6. November 2024

Steinþór Steingrímsson




- Language Technology was one of the focus areas of Iceland's presidency of the Nordic Council of Ministers:

  ***"During the Icelandic presidency, emphasis will also be placed on the development of a common Nordic policy on digital language technology."***

- A group was formed to compile a report on the status of less-resourced Nordic languages

- https://www.government.is/publications/reports/report/2024/02/02/Language-Technology-for-Less-Resourced-Languages-in-the-Nordics/

# Objectives of the group

- Report on the status of LT for less-resourced languages in the Nordics
  - Languages with fewer than 1 million speakers
  - Faroese, Greenlandic, Icelandic, Karelian, Kven, Meänkieli, Romani, Sámi languages

- We give an overview of developments after the ELE reports were written, as well as comments or additions

- We discuss opportunities for Nordic collaboration in advancing LT for these languages

- We propose recommendations on how this can be supported by policy

# Status of languages

- Each language discussed in a separate section
- There is great variety in LT support for these languages
  - Size of the LT community
    - Education locally
  - Large differences in public support
  - Large differences in available data
  - Large differences in available tools
  - Differences in intitutional language planning

- It can be hard to transfer actions or activities that work from one language to another… but people can work together

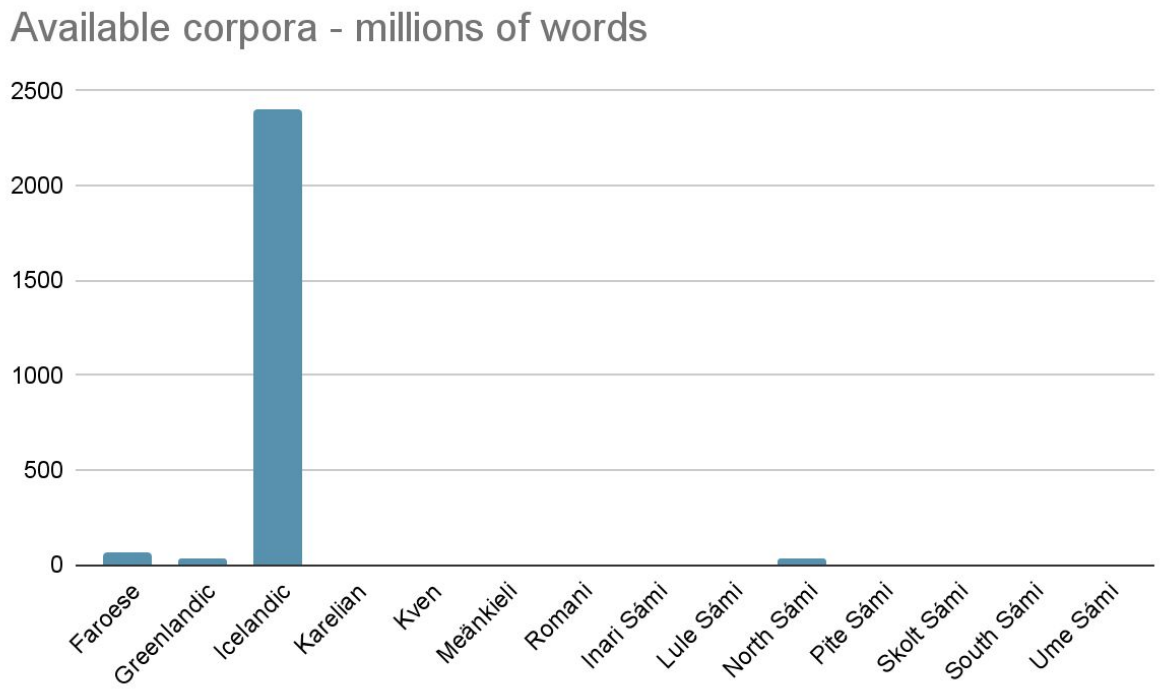
# Availability and maturity of selected language tools

| Language | MT | Grammar model, spell checker | TTS | ASR |
|---|---|---|---|---|
| **Faroese** | multilingual, beta | production | yes | yes |
| **Greenlandic** | beta | production | yes | no |
| **Icelandic** | multilingual | production | yes | yes |
| **Karelian** | alpha | beta | no | no |
| **Kven** | alpha | production | no | no |
| **Meänkieli** | alpha | beta | no | no |
| **Romani** | - | alpha for some | no | no |
| **Inari Sámi** | alpha | production | no | no |
| **Lule Sámi** | alpha | production | alpha | no |
| **North Sámi** | alpha | production | yes | alpha |
| **Pite Sámi** | - | beta | no | no |
| **Skolt Sámi** | alpha | beta | no | no |
| **South Sámi** | alpha | production | no | no |
| **Ume Sámi** | - | - | no | no |

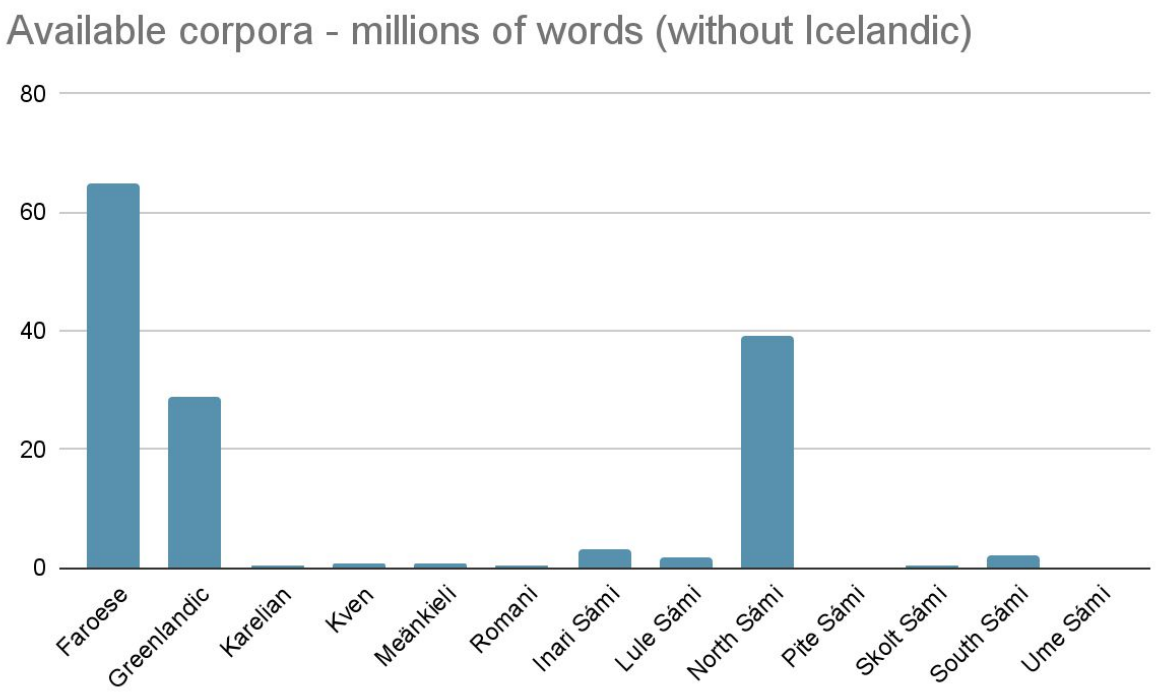
# Availability and maturity of selected language resources

| Language | Lexical resources | Speech corpora | Parallel corpora, words L1 |
|---|---|---|---|
| **Faroese** | good | medium | medium |
| **Greenlandic** | medium | - | small |
| **Icelandic** | good | large | large |
| **Karelian** | low | - | - |
| **Kven** | low | - | medium |
| **Meänkieli** | medium | - | small |
| **Romani** | low | - | - |
| **Inari Sámi** | medium | - | medium |
| **Lule Sámi** | low | medium | medium |
| **North Sámi** | medium | medium | large |
| **Pite Sámi** | low | - | - |
| **Skolt Sámi** | medium | small | small |
| **South Sámi** | low | - | medium |
| **Ume Sámi** | low | - | - |

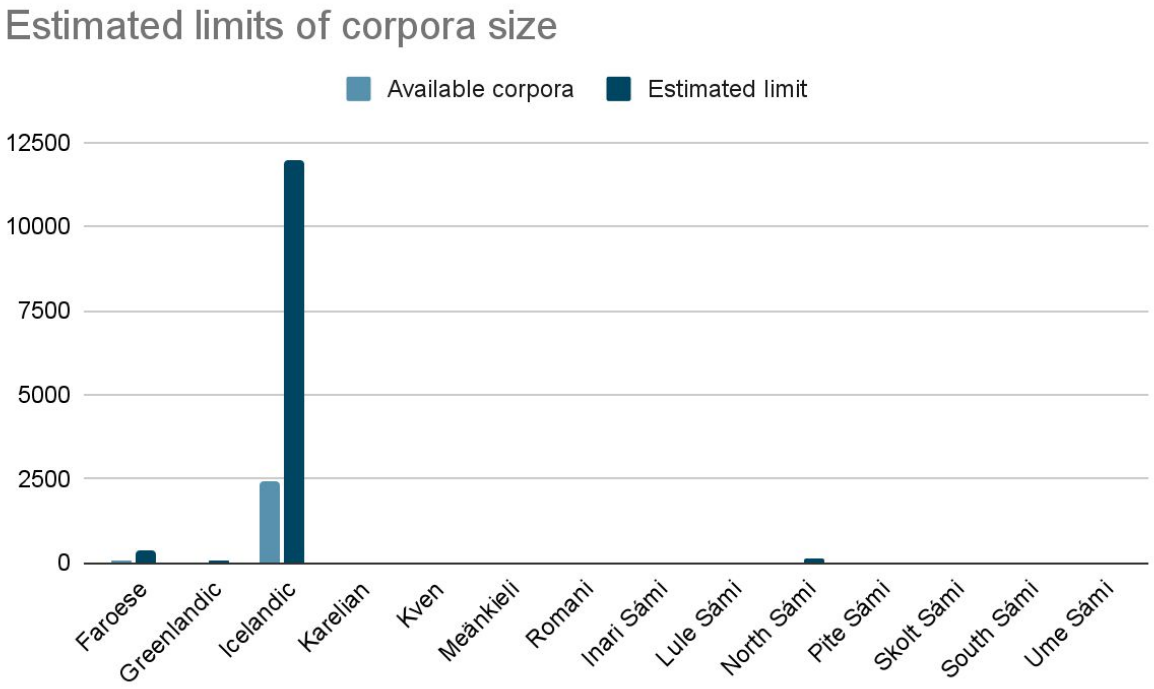
# Size of available text corpora



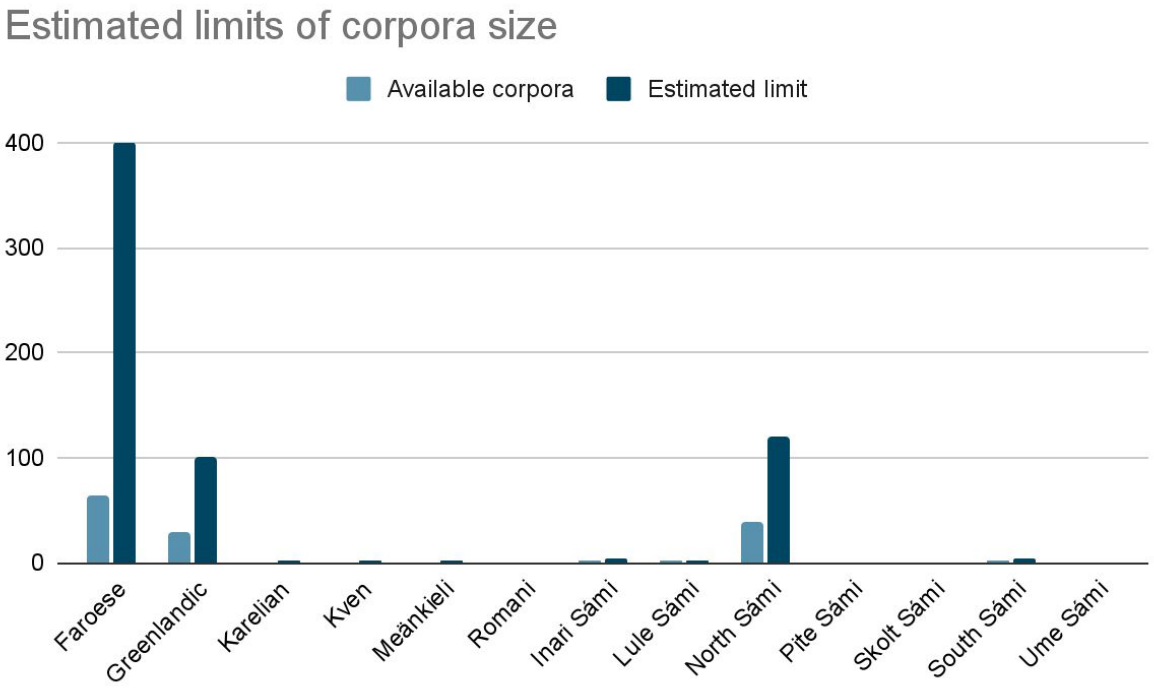
# Size of available text corpora



# Estimated possible size of monolingual corpora



# Estimated possible size of monolingual corpora



# Developments after the ELE reports

- For each language, there is a section on developments since the ELE reports were written in 2022
  - New data, tools and/or models for: Faroese, Icelandic, Greenlandic
  - Some developments for others:
    - Karelian: MT, keyboard layout for Karelian
    - Meänkieli Work on a spellchecker started
    - Romani Language revitalisation program (some LT there)
    - Sámi languages R&D group in UiT has grown; neural MT
    - Kven Dictionary work

# Methods in LT for low-resource languages

- The trend in LT has been to build larger and larger models
  - Powerful for some areas, such as MT, ASR and TTS – given that large enough resources are available
  - If large amounts of data are unavailable – we have to use other approaches
    - Statistical (data needed, but less than for neural)
    - Rule-based (no corpora needed)

# Opportunities for Nordic collaboration

- Education
- Infrastructure
- Exchanging knowledge and methods
  - Research collaboration
  - Network
- Collaboration on content and data

# Recommendations

A. **Continued focus and investment** in LT for small languages
B. **Initiate LT strategies** and implementation
C. **Collaboration on the political level** to advocate accessibility on major platforms
D. **Collaboration on research and development** be continued and intensified
E. **Collaboration on creating educational programmes/courses**
F. **Introduce legislation** that facilitates data collection

Árnastofnun

# Report on Language Technology for Less-resourced Languages in the Nordics: Current State and Prospects

Language Technologies and AI in the Nordic Countries
5.–6. November 2024

Steinþór Steingrímsson