February 2024

Language Technology for Less-Resourced Languages in the Nordics

Current Developments and Collaborative Opportunities

Steinþór Steingrímsson
Iben Nyholm Debess
Kimmo Granqvist
Per Langgård
Trond Trosterud

Stjórnarráð Íslands

Publisher:

Stjórnarráð Íslands

Language technology for less-resourced languages in the Nordics: Current Developments and Collaborative Opportunities

February 2024

Authors:

Steinþór Steingrímsson

Iben Nyholm Debess

Kimmo Granqvist

Per Langgård

Trond Trosterud

Umbrot og textavinnsla:

Stjórnarráð Íslands

Table of Contents

Motivation for the report

In 2019, the Nordic Prime Ministers approved a new vision for Nordic cooperation: to make the Nordic countries the most sustainable and integrated region in the world by 2030. In order to fulfil this vision, three strategic focus areas were put forward: green Nordics, competitive Nordics and socially sustainable Nordics.

In the programme for Iceland's presidency of the Nordic Council of Ministers in 2023, special attention was paid to the three focus areas in the Council's action plan, with a particular emphasis on the Nordic countries moving forward in digital development and looking for ways to make new electronic solutions accessible to everyone. As noted in the programme, the Nordic countries have been at the forefront of innovation and implementation of technological solutions. As the ultimate goal is to create a continuous, inclusive region, the programme suggested developing a "common Nordic policy on digital language technology, which may contribute to the advancement of all Nordic languages in the digital world, so they will be accessible in communications with devices and in all data work."[1]

With the advent of artificial intelligence, globalisation and increased migration, the status of languages in the digital world has undergone enormous changes. The Iceland government's recent investment in language technology for Icelandic, with the project plan *Language technology for Icelandic 2018-2022*, has already yielded positive results. The project plan has been crucial for the status of the Icelandic language in an international context where minority languages and languages with relatively few native speakers are in a precarious position, not least due to the spread of social media and overarching influence of a few tech giants. This is as true in the Nordics as it is around the world.

One of the initiatives under Iceland's presidency was thus to establish a working group made up of experts in the fields of languages and language technology. The group was to compile a report on the status of Nordic minority languages, and the languages of island nations in the Nordic region, in relation to language technology.

Due to their low number of native speakers, these languages are at increased risk of being overwhelmed by rapid technological changes. Some

[1] The Nordic Region – A Force for Peace. Programme for the Icelandic Presidency of the Nordic Council of Ministers 2023. Nordic Council of Ministers. 2022.

of the languages are official languages, while others are not. As Iceland's presidency of the Nordic Council of Ministers has a strong tradition of promoting West Nordic cooperation and drawing attention to the uniqueness of the region, it was decided that the project group would focus particularly on minority languages in the region, as well as other West Nordic languages with less than a million native speakers, instead of Danish, Norwegian, Finnish and Swedish.

„The languages that are spoken in the Nordic countries are a fundamental pillar of the cultural richness we enjoy in the region. Through our languages, people have told stories and tales throughout the centuries that bind us Nordics together. The fast technological changes we've witnessed in recent years pose a challenge to our language wealth that is of utmost importance to address for future generations. It is possible to have thriving Nordic languages, big and small, while these technical evolvements occur. To do so, we must invest in our languages and make them relevant and accessible in the digital world. This report will be helpful on that journey and yield positive results both in the short and long term for our beautiful languages and the upcoming generations of their speakers.“

Lilja D. Alfreðsdóttir, Icelandic Minister of Culture and Business Affairs.

1. Introduction

The working group, established to survey the status of Nordic languages with fewer than one million speakers in the context of language technology (LT), was composed of five experts in the field of LT. Four representatives were nominated by the Committee of Senior Officials for Culture (EK-K): Iben Nyholm Debess, coordinator of the Centre for Language Technology at the University of the Faroe Islands, Kimmo Granqvist, senior lecturer at the University of Helsinki, Per Langgård, head advisor at the Language Secretariat of Greenland, and Trond Trosterud, Professor at the Arctic University of Norway. Steinþór Steingrímsson, project manager at the Árni Magnússon Institute for Icelandic Studies, was nominated by the Icelandic Ministry of Culture and Business Affairs.

The group was tasked with reporting on the status of LT for the languages in question; whether any policies or programmes are in place or being planned for these languages, the use of LT for these languages in the public sphere, and how collaboration and mutually beneficial projects could have a positive impact on these language communities.

The languages discussed in the report are Faroese, Greenlandic, Icelandic, Karelian, Kven, Meänkieli, Romani and the Sámi languages. These are the official Nordic languages spoken by less than one million people, as well as many of the minority languages spoken in the Nordic countries. The main Greenlandic dialects (East Greenlandic, Inuktun and West Greenlandic) are not addressed individually in the report because the written variety of Greenlandic is officially recognised as the single national standard for all dialects in written form.

Romani, Kven and Sámi have official status in Norway, and Romani and Meänkieli are officially recognised as minority languages in Sweden. Finland has no official national minority languages, but in addition to the indigenous Sámi languages there are four minority languages with a history dating back to pre-independent Finland. Two of these, Karelian and Finnish Romani, are included in the report while the other two, Yiddish and Tatar, are not included. The reason for this is that the orthographic and grammatical standardisation of these languages is (also) conducted outside the Nordic countries. Swedish in Finland is not included in this report, as it is defined as a national language, nor are the Nordic sign languages included here as they deserve their own report.

Some work in terms of LT has been carried out for all of the languages discussed in this report, but what has been done differs widely. Previous work on the languages is discussed in reports from the European Language

Equality (ELE) project, which was concluded in the spring of 2023.[2] The project included reports on all official EU languages as well as some minority languages. In February 2022, the ELE report on Icelandic was published,[3] followed by a report on the other languages discussed here in May 2022.[4] We will give an overview of developments after these reports were written, as well as providing comments or additions to the reports. Finally, we discuss opportunities for Nordic collaboration in advancing the field of LT for the languages discussed herein, and propose recommendations on how this can be supported by policy.

Recommendations (Discussed in more detail in Section 5)

Language technology for small languages should be a continued focus of Nordic language policy in the years ahead.

Collaboration on the political level to advocate for accessibility on major platforms.

Initiate language technology strategies and implementation for the languages that currently lack such measures.

Support for collaboration within language technology for Nordic national and indigenous languages should be continued and intensified.

Collaboration on creating educational programmes/courses for developing and maintaining local expertise.

Introduction of legislation to facilitate the collection of corpus material.

[2] https://european-language-equality.eu

[3] *European Language Equality. D1.19. Report on the Icelandic Language*

[4] *European Language Equality. D1.38. Report on the Nordic Minority Languages*

2. Languages

2.1 Faroese

Introduction

Faroese is spoken by an estimated 70,000 people. The Faroese language is the national language of the Faroe Islands, and serves as the primary medium of communication across domains. Faroese exhibits dialectal variations, primarily manifested at the phonetic level, all coherently represented through a standardised orthography.

Status

The 2021 European Language Equality Report on Nordic Minority Languages provides information about the status of accessibility and implementation of language technology in Faroese. In short, the ELE Report reports language technology support for Faroese on a basic level. Refer to the report for details.

Shortcomings

Access to the major platforms is a highly relevant issue for Faroese. While Faroese is acknowledged as a language on these platforms, the lack of language tools or functionalities remains a challenge.

Despite the language being recognised and digital keyboard compatibility being implemented across all major operating systems, Faroese lacks support as a system-wide localised language within any of them (Windows, macOS, Linux, Android, iOS). The same applies to large software suites (Microsoft Office, Google Docs, Libre Office). While certain components of these systems and software show signs of localisation efforts for Faroese, such instances are sporadic and secondary (e.g. the names of emojis are Faroese, but all major and primary functions are not). Given the widespread implementation of these operating systems and software suites in both educational and professional sectors, Faroese speakers are forced to choose a linguistic interface in a non-native language across all personal or professional devices.

Another concern in the Faroese LT environment is the development and cultivation of local expertise. As an example, the Faroe Islands have experts in language and experts in technology, but, until recently, no local specialists in the intersecting field of LT. This situation is slowly improving,

with Nordic cooperation laying the groundwork for developing the necessary skills locally. Measures should also be taken in the near future to ensure sustained education in the field.

Microsoft included Faroese in their Translator application, leveraging a locally developed parallel dataset. The development of the application itself is not local, however, and the linguistic quality, and thereby usability, is not ideal. This exemplifies the importance of local autonomy and influence, as well as a native evaluation framework for Faroese in order to assess the quality of externally developed data, applications and tools.

Recent developments

The field of Faroese language technology has grown in recent years, marked by successful completion of projects and initiation of new ones as well as growing international scholarly interest. As mentioned in the ELE Report, the Ravnur Project published a Basic Language Resource Kit targeted at speech recognition. The project and the BLARK dataset has resulted in the development of several applications for speech recognition. Concurrently, external researchers' interest in the field has resulted in novel datasets designed for diverse tasks, as well as the development of new tools and smaller language models.

One major development in recent years has been the establishment of the first publicly funded institution dedicated to LT research and development. In 2023, the University of the Faroe Islands founded the Centre of Language Technology, a modest research centre under the Department of Language and Literature. This institutional initiative recognises the importance of formalising the work in LT and ensuring stable progress. The centre is currently involved in various research and development projects in the fields of speech recognition, text corpus compilation, machine translation, language modelling, OCR of handwritten texts, spell checking, and generation of specific datasets (e.g. Semantic Textual Similarity).

Recent initiatives have also resulted in a variety of new resources, datasets and tools. Some resources have been developed from local initiatives (e.g. the BLARK) and some by neighbouring collaborators (e.g. BÍN and FoBERT).

Advancements in neural techniques and artificial intelligence in general, and the new generations of large language models and the typically multilingual approach, have also opened up new possibilities for Faroese. For example, Faroese was included in the large multilingual Meta: NLLB (translation) and MMS (speech technology) projects. OpenAI's GPT4 demonstrates competence in understanding, translation and generation of Faroese (with room for improved linguistic quality, though). The GPT-SW3

from AI Sweden performs well when translating Faroese to English, despite not being trained with Faroese data. The Faroese proficiency of these models seems to be the result of transfer learning via data from the other Nordic languages, especially Icelandic. In and of themselves, these models (the translation models, speech models and language models) are not yet at a linguistic level or accessible enough to be useful to Faroese speakers in everyday use. However, their performance in Faroese shows great promise for development. The capabilities of the models can facilitate the increase in data volumes via data augmentation for further development of models and tools.

While acknowledging the positive inclusion of a minority language in projects with major corporations, it is important to be aware of the cultural biases and other issues that arise in these contexts.

In summation, the recent developments in Faroese language technology have been positive. The number of resources is expanding and new resources are generally accessible. It should be noted, however, that much of the work remains in developmental and research stages, and broad applicability and functionality are yet to be obtained.

Resource overviews and updates will be available at mtd.setur.fo and borealium.org.

Ongoing or planned LT initiatives

The strongest initiatives in recent years have been the now completed Ravnur Project, and the new language technology research centre at the University of the Faroe Islands. No further large initiatives are planned at this point.

LT in the public sector

The public sector in the Faroe Islands uses LT to some extent. Faroese text-to-speech, a long-standing feature, has been implemented in relevant institutions and is mainly used to assist people with reading impairments. Its implementation spans the entire educational system. Additionally, many websites (news, public information) have applied the text-to-speech functionality. Contrastingly, speech recognition has not been implemented as an operational tool in the public sector. The recent introduction of speech recognition capabilities in Faroese necessitates a gradual assimilation. Applicable settings in the public sector could be e.g. the Parliament (as seen in other countries). The health care sector has expressed interest in implementing speech recognition in their upcoming system, but as of yet this is very preliminary. Another applicable setting for

speech recognition as a writing aid is in the educational system. The Faroese spell checker is widely used, both privately and professionally.

The local private sector has not been active in the field of Faroese language technology. Any private initiatives to develop solutions come from companies abroad (Norway and Denmark).

Recommendations for next steps

As the ELE Report also states, the Faroe Islands lack a general LT strategy. The importance of LT development and infrastructure will only increase. A persistent recommendation is the development and formulation of a political strategy dedicated to the development of the Faroese language technology field and efforts to secure funding for executing and implementing the strategy.

The strategy should address these points:

- Clarifying the priority of data and tools - what novel developments are needed, what should be sustained and updated.
- Continuing and expanding regional collaboration between developers and researchers. Focus on e.g. methodology and best practice for low-resource languages.
- Engaging in political advocacy to promote accessibility on major digital platforms, necessitating regional collaboration on a political level.
- Creating an environment for the private sector to engage in this field and offer services.
- Securing and cultivating local expertise, recognising the historical dependence on foreign experts for project execution. While collaboration with external partners is crucial, these must be carried out in balance with local autonomy. Investing in education in language technology.
- Focusing on widespread implementation of relevant tools and software to ensure broad user accessibility, thereby promoting inclusivity in language technology utilisation.

2.2 Greenlandic

Introduction

Greenlandic is the biggest language in the family of Inuit languages, spoken by a little under 100,000 people across the Arctic, from the easternmost tip of Siberia to East Greenland. Greenlandic, or *Kalaallisut*, is the official language of Greenland and first language for approximately 75% of Greenland's total population of 56,600 people.

Status

Greenlandic language technology has almost exclusively been developed at Oqaasileriffik/the Language Secretariat since its beginnings in 2005. However, throughout this period, there have been sporadic attempts at developing machine translation (MT) using stochastic technology or artificial intelligence at foreign universities and by major commercial entities. However, until very recently, none of these have achieved functionalities that could be integrated into Greenlandic everyday life.

The developmental work of Oqaasileriffik has been particularly affected by the fact that no universities in the world offer education in LT that incorporates the very specific requirements necessary for NLP for a polysynthetic language. Consequently, with support from respective partners at the Arctic University of Norway and the University of Southern Denmark, Oqaasileriffik has had to handle the qualification of the next generation of Greenlandic language technologists single-handedly. Internal education remains an activity that drains many of the already limited resources for the development of more advanced and improved Greenlandic language technology.

Following the conclusion of the kal-dan-kal MT project in late 2021, the language technology team was significantly reduced. During this time, 2-3 man-years have been allocated, spread across several staff members with a variety of obligations besides LT.

In 2022-23, efforts to make the fundamental Greenlandic resources robust enough for more advanced applications have been a top priority. Still, the alpha versions of Greenlandic MT (kal-dan and dan-kal) are continuously maintained and improved, although the progress only slowly manifests as enhanced quality in the actual translations. However, this work contributes significantly to strengthening the Greenlandic language model as evidenced in the Greenlandic parser.

Recent developments

The activities and initiatives outlined in the ELE report from the spring of 2022 have all been expanded and maintained since then and many new initiatives have been established.

Kal2eng gloss instrument: English L2 is rapidly gaining ground in all parts of Greenlandic society, especially after the introduction of the high-speed internet cable, which has made access to the internet fast and affordable. These developments have been so rapid that it is no longer uncommon to hear Greenlandic children speak English among themselves. Therefore, Oqaasileriffik assesses that there is an urgent need for modern English instruction tailored to Greenlandic conditions to replace the existing education based on Danish teaching materials and Danish traditions. Based on this, Oqaasileriffik has initiated a pilot project aimed at building a comprehensive Greenlandic-English semantically tagged lexical resource, which presently has a growth of approximately a thousand lemmas per year.

In conjunction with the development of the English lexical resource, a glossing tool was created and made accessible for free on Oqaasileriffik's website at the start of 2024. This glossing tool is based on the Greenlandic parser, incorporating automatic glosses retrieved from the lexical database. It is expected to contribute significantly to improving general language comprehension, particularly by highlighting similarities and differences between parts-of-speech, function words, etc., in the two languages. The glossing tool is also expected to benefit both Greenlanders learning English as an L2 and non-Greenlanders. This tool will help individuals acquire enough information to navigate through a Greenlandic text while waiting for a sufficiently robust Greenlandic-English MT to be developed.

Word prediction: Word prediction has broad application for use, including within the lexicon function in text processing and as an integral component of compensatory aids for weak readers and individuals with dyslexia.

Until now, it has been assumed that word prediction would not be feasible for the polysynthetic Greenlandic language as the choice of the next morpheme in the ongoing word construction is so flexible that predictability is almost inconceivable. Yet, the ongoing compilation of texts for the Greenlandic corpus and increasingly robust analyses of continuous Greenlandic text have created an opportunity to develop statistical models even at the morpheme level.

In 2023, Oqaasileriffik had the opportunity to initiate a word prediction project, thanks to a substantial grant from NORDPLUS Nordic Languages. The work is underway, but it is still too early to determine whether the prediction system will become robust enough to be integrated in live

applications. However, due to the substantial need, especially for individuals with dyslexia, it is imperative to explore this avenue as thoroughly as possible.[5]

Term base: Following political demands and with full funding, Oqaasileriffik's most significant task for both this year and the next is the development of several terminologies, with legal terminology being the largest among them. Construction of the terminology database creates a fundamental resource upon which current and future LT can be built.

Greenlandic L2 with technological support: The largest Greenlandic L2 system[6] is undergoing revision in 2023 and 2024. Greenlandic LT is directly involved in the revision, as several chapters include reception exercises with machine-generated words and sentences presented by the Greenlandic speech synthesis.

The introduction of this technology is expected to mark a paradigm shift in Greenlandic as a foreign language. For the first time, it will become possible to generate an infinite number of examples to train the endless combinations that can be formed from the student's constantly expanding linguistic knowledge. This will be achieved without involving unfamiliar roots, derivatives, or inflections for the learner, which could complicate or hinder the comprehension—and consequently, the acquisition—of familiar material.

As indicated in the ELE report and other sources, previous attempts to create translation systems to and from Greenlandic using statistical approaches have only shown limited results. However, the outcomes of a recent study from Aarhus University have been intriguing. With the latest advancements in artificial intelligence, it is increasingly well documented that data-driven technology has unequivocally surpassed rule-based technology, including in Greenlandic MT.

The Greenlandic online translators Kal-dan and Dan-kal remain operational and will continue to be maintained until the next generation of Greenlandic MT is ready. The exact timeline for this transition is currently unclear, but there are ongoing experiments with hybrid MT using texts tagged with rule-based technology as input aiming to partially compensate for the scarcity of parallel training data.

[5] See Tino Didriksen. 2023. *Morphemes as Predictive Text Units.*

https://oqaasileriffik.gl/d/papers/maptu.pdf

[6] https://learngreenlandic.com/online/

The integration of artificial intelligence in MT does not signify the complete abandonment of rule-based translation. On the contrary, it is anticipated that rule-based translation will be retained and continuously corrected. This form of translation is deemed to have far greater pedagogical potential, both for native language instruction in Greenlandic and as supportive material in Greenlandic L2.

Crucial shortcomings

Today, there is usable Greenlandic language technology for many societal needs. However, there is a crucial resource that hasn't even been initiated yet, and there are significant issues with the application and dissemination of the existing resources.

The lack of Greenlandic speech-to-text effectively impedes many initiatives that could otherwise enhance the Greenlandic community and language. Without speech recognition, there is no automatic subtitling for films and television, no advanced compensatory aids for weak readers and individuals with dyslexia, no elegant solutions for costly telephone settings, and no support programmes for cross-lingual oral communication, among many other missed opportunities for the Greenlandic language. Simply put, there are numerous possibilities for the Greenlandic language that cannot be realised due to lack of access to speech recognition.

The cost of developing Greenlandic speech recognition cannot be accurately determined at present. Estimates vary from low development costs due to the limited Greenlandic phonetic inventory, to high costs because the numerous sandhi rules complicate the isolation of di- and triphones, making unit selection almost impossible. Oqaasileriffik consequently invests much energy in trying to secure funds for a limited pilot project on speech recognition, which they hope will shed light on the expected development costs.

The second challenge in Greenlandic language technology lies in its application and dissemination. Even though Greenland more or less possesses the expertise to develop programs for most societal needs, these programs are not available on major platforms like Microsoft, Google, and Apple.

LT in the public sector

Greenlandic language technology tools and resources, such as writing aids, lexical resources or speech synthesis, are always open source and accessible online for everyone as soon as they are completed. They are widely used, both by individuals and public institutions, and the demand for more is felt daily.

There are two major impediments that reduce or delay Oqaasileriffik's possibilities of producing more of the services society asks for:

1. Tech giants
2. Lack of staff

Issues regarding services provided by the big tech companies are thoroughly examined in the ELE report and are not repeated here. The latter problem could be alleviated with more financial and human resources allocated to language technology development in Greenland and to educational programmes securing future staff with the skills needed for next generation Greenlandic language technology. As mentioned, only 2-3 full-time years of work are allocated to all language technological activities in Greenland, including language technological education. This clearly imposes many limitations on how ambitious the development can be.

Political anchoring is factual, but not formalised.

Greenlandic language technology has only limited departmental support to rely on. In such a small central administration, access to expert assistance is understandably limited. In Greenland, this means a significantly different chain of command compared to larger societies, from a political desire/demand for language technology to actual development work.

In practice, the path from wish to execution roughly looks like this: (i) The public/politicians create a sort of "wish list" of desired initiatives and new applications, (ii) Inatsisartut (the Greenlandic Parliament) decides, via the budget allocation to Oqaasileriffik, the financial framework to be prioritised for language technology development in the following fiscal year. Inatsisartut can also decide to initiate specifically selected initiatives and allocate earmarked funding for such initiatives, as happened in 2021 when it was decided to prioritise the development of legal terminology in Greenlandic. (iii) Oqaasileriffik is tasked with prioritising the non-earmarked wishes within the outlined framework and executing earmarked projects.

The path from the political level to the actual developers of Greenlandic language technology is very short. It is therefore reasonably safe to claim that Greenlandic language technology, in practice, is politically anchored, but that this anchoring is far from formalised in the same manner as known in larger societies.

Mutual exchange of information between the small languages.

Greenlandic language technology has from the outset been entirely dependent on collaboration with larger language technology centres outside Greenland, especially with Giellatekno at UiT - the Arctic University of Norway and VISL at SDU. Without support from these two institutions, Greenlandic language technology would undoubtedly have found itself in a much less favourable position than is currently the case.

Conversely, all Greenlandic resources and applications are open source and readily accessisble for inspiration and best practice in LT for other small languages. This was evident when the Nordic Council of Ministers established the initiative Små Språk i Norden, where Greenland's active participation included arranging a best practice seminar for representatives from a number of small Nordic languages.

Recommendations for next steps

After the hesitant first attempt in 2005, Greenlandic language technology has developed rather slowly but steadily over the last decade or so. Maintaining the progress and gathering momentum will under all circumstances be the prime demand for all stakeholders in Greenlandic LT in the years to come.

These are some of the measures needed:

- Greenland needs access to educational programmes to develop new and maintain existing projects that require skills in rule-driven technology, as well as enable students to include AI and other technologies of tomorrow in their professional activities. Such an education is not available anywhere at the moment
- Development of non-existant basic resources, like speech recognition, that are the prerequisite for building essential tools for Greenlandic society (e.g. assistive technology for people with disabilities) must be funded as soon as possible and development work initiated.
- In order to provide native language support for future generations, serious attempts to make Greenlandic a viable alternative to English in man-machine communication, for instance as a query language on the internet, should be undertaken as soon as possible.
- Strong political advocacy to promote accessibility on all major digital platforms, including regional collaboration on a political level.

2.3 Icelandic

Introduction

Spoken by approximately 350,000 people, Icelandic is a North Germanic language closely related to Faroese, Norwegian, Danish and Swedish. Icelandic is the national language of Iceland and while it is used across domains as the primary language in the country, a growing number of immigrants, almost 20% of the population, have other native languages.

Status

For a language with only 350,000 speakers, the current availability of Icelandic language resources and tools is remarkable. Most of these resources and tools are direct or indirect outputs of the Language Technology Programme for Icelandic (LTPI), which was launched by the Icelandic Government in September 2019 and ended in September 2022. Before the programme commenced, a report was written outlining the language technology infrastructure that was needed.[7] This report was essential when the LTPI was implemented. The resources and tools built within the programme are available for free under standard open licence. The ELE report on Iceland briefly describes the most important language resources and tools, projects and initiatives and compares the status of Icelandic LT to other languages.

For Icelandic, a variety of LT tools have been made available, including tools for text analysis, speech processing, machine translation and spell checking.

ABLTagger is the PoS-tagger that achieves the highest accuracy for Icelandic, 96.95% on the MIM-Gold tagset, and further improvement in PoS-tagging may prove difficult. For lemmatizing, Nefnir gives the best results. The most recent version of Nefnir was published in 2019 and while it accurately finds lemmas for most known words, prediction of unknown words can be improved.

In 2012, Google developed speech recognition for Icelandic in cooperation with Icelandic researchers. Around the same time a speech synthesiser for Icelandic was developed by the Polish company Ivona, which later became a subsidiary of Amazon. While these applications have proved useful, their private ownership limits their use. Within the LTPI a number of tools for

[7] See *Language Technology for Icelandic 2018-2022 – Project Plan*
https://rafhladan.is/bitstream/handle/10802/20054/mlt-en.pdf?sequence=1

speech processing were developed, including automatic speech recognition (ASR) models and a text-to-speech (TTS) voice.

The first openly available MT models were published within the LTPI, with both a Moses-based statistical machine translation (SMT) model being made available as well as neural machine translation (NMT) models. The English-Icelandic language pair was part of the shared news translation task at the WMT conference in 2021. A number of systems were submitted and compared on how well they fared in translating in both directions for these two languages. Test and development sets were created in order to evaluate the translation systems. Translation models are available for English->Icelandic and Icelandic-> English, as well as a bidirectional model trained to translate between Icelandic and Polish. While the English–Icelandic models can in some cases generate decent translations, the quality is lacking for many domains.

A neural model for spell checking was published in 2022. While it can accurately correct common errors, it does not explain the errors and why they should be corrected, which is essential for language learners.

Recent developments

Our discussion on the developments since the publication of the ELE report can be divided in three parts. 1) The output of LTPI in 2022 and 2023. 2) The advancement of large language models (LLMs), some of which have capabilities in Icelandic. 3) The resolution of the Icelandic government to start a new language technology programme for Icelandic in 2024, as well as continuing the focus on language and language technology in the strategic research and development programme (*markáætlun*).[8]

The ELE report on Icelandic was published in February 2022. A variety of tools and resources have been published since then, many of which are the output of the last part of the LTPI, but others built in relation to other projects. The most important developments are listed below:

- *Monolingual Corpora:* In the last year, work on the IGC has continued, with the number of tokens in the corpus rising from 1.9 billion at the beginning of 2022 to 2.7 billion by the end of 2023.

[8] https://www.stjornarradid.is/efst-a-baugi/frettir/stok-frett/2023/04/04/Um-tveggja-milljarda-fjarfesting-i-maltaekni-naestu-fjogur-ar-/

- *Monolingual Corpora:* A new corpus of 19th century texts has been published.
- *Parallel Corpora:* A number of other web-based parallel corpora between Icelandic and other languages have been published on opus.nlpl.eu. Most of these are of rather low quality and need extensive filtering, and some of the translations from Icelandic into languages other than English are pivoted. The usefulness of these datasets needs to be assessed.
- *Speech Corpora:* The latest version of Samrómur contains over 143,000 minutes of speech, which is a slight addition to the size of the corpus in February 2022.
- *Speech Corpora:* A number of new speech corpora have been published, including: **Raddrómur,** a 49 hour corpus intended for speech recognition, made up of radio podcasts, mostly from RÚV, the Icelandic National Broadcasting Service. **Spjallrómur** is a conversational speech corpus for speech technology development. It contains 54 conversations, totalling over 21 hours. **Samrómur Children** contains speech from children between 4-17 years old. **Gamli** is an ASR corpus for Icelandic oral histories, derived from the ethnographic collection of the Árni Magnússon Institute for Icelandic Studies. It contains 146 hours of transcribed audio, with the majority being recordings of people over 60 years of age.

Since OpenAI launched ChatGPT in November 2022, LLMs have not only attracted the attention of LT researchers and practitioners, but also of the general public. In early 2023 it was announced that Icelandic was the second language ChatGPT was trained to generate, using reinforcement learning from human feedback. This generated much interest in Iceland for exploring the possibilities to utilise LLMs for various language-related tasks. Creating test suites for evaluating the capabilities of LLMs when using Icelandic is essential for understanding which LLMs can be useful and when.

Research and development of Icelandic language technology has been strengthened considerably by the LTPI. A wide variety of language resources have been introduced, lowering the threshold for application development and enabling a wide area for research in the field. The number of experts in the field has grown considerably, with many working full time on LT both in academia and industry. Icelandic companies are realising the possibilities of employing LT tools in their workflow and the research output of academics working on LT has multiplied in the last few years, with over 30 peer-reviewed LT papers published by Icelandic researchers in 2022 alone, compared to less than 5 papers a year up until five years ago.

LT in the public sector

Language technology is employed by various public bodies. Since 2016, the Icelandic parliament, Alþingi, has used speech recognition to expedite transcription of parliamentary speeches. The City of Reykjavík followed suit in 2022 and uses speech recognition to add captioning in real time to a webcast of city council meetings. The Ministry for Foreign Affairs' translation department has participated in an experiment where MT systems are specially trained to work with regulatory texts to help with the translations of EEA regulations, directives and other documents pertaining to the EEA agreement. Furthermore, the Ministry of Culture and Business is funding a language portal for use in the education sector, targeting primary and secondary school students, as well as second language learners. The portal will contain prescriptive monolingual and bilingual dictionaries, use speech recognition to increase accessibility and MT to translate sentences and phrases.

LT has also been adopted in other sectors,most notably by the Icelandic Association of the Visually Impaired (IAVI), which participated in the development of Icelandic voices for a speech synthesiser in 2010. While these voices are still in use, there is a call for better voices and broader variety from the IAVI, people with dysarthria, and others who rely on speech synthesisers to be able to fully participate in society.

Ongoing or planned LT initiatives

The government plans to launch a new three-year language technology programme in 2024 and invest ISK two billion in language technology. Suggestions from a steering group are currently awaited. Part of the funds to the new programme will be allocated to the Strategic Research and Development Programme for Language Technology, with the aim of strengthening and developing language technology research.

2.4 Karelian

Introduction

Karelian is a northern Baltic Finnic language closely related to Finnish, spoken mainly in Russia but also in Finland. There are approximately 5,000 Karelian speakers in Finland who speak the language as their mother tongue. In addition, another 20,000 people identify as Karelian and can understand and speak the language to some extent. Finland has ratified the European Charter for Regional or Minority Language for Romani as a Part II language.

The language policy programme of the Finnish government includes measures concerning the Karelian language. Under the programme, a government-level Karelian language expert working group was to be set up. Karelian is taught at the University of East Finland, which also has undertaken the implementation of a Karelian language revitalisation programme.

There are two Karelian written standards, North (Viena) Karelian (ISO: krl) and Livvi Karelian, or Olonets (ISO: olo). In addition, Suojärven Pitäjäseura Karelian promotes the development of written Karelian in Finland based on the southern dialects. The YLE Karelian news service is in Livvi Karelian.

Status

The ELE report gives a good picture of the status of Karelian language technology. The basic language technology resource is the grammatical language model. For Olonets, this has a coverage for running text at about 83%, for Karelian proper it is around 62%.

Recent developments

Grammar models, and thereby proofing tools and text analysis tools, exist in the GiellaLT infrastructure both for Olonets Karelian (beta level) and Karelian proper (alpha level). Lexical resources for Karelian are available in the GiellaLT infrastructure and a keyboard layout for Karelian has been produced. Tartu University's machine translation engine now translates 23 Finno-Ugric languages, among them Karelian. Additional attempts have been made to obtain funding for Finnish LT-related Karelian work, but that has not materialised yet.

LT in the public sector

The Ministry of Education and Culture has supported the written use of Karelian by producing learning materials in Karelian and making Karelian-language content openly available online and on social media. The Finnish national broadcaster YLE broadcasts radio programmes and publishes web pages in Olonets Karelian, and a cooperation network has been established for Karelian language instructors in liberal adult education. None of these initiatives make use of LT resources.

Recommendations for next steps

Except for keyboards, the Karelian language community currently has no language technology support. The existing Karelian proper and Olonets

Karelian grammatical language models should be improved to a text coverage of well above 90%, thereby making them useful as spell checkers. Work on speech technology should also be within reach, utilising existing speech corpora, work on Finnish speech technology and existing open infrastructure.

2.5 Kven

Introduction

The Kven language, spoken in Norway, is the result of a northbound expansion of Finnish prior to the 20th century. Originally a part of a dialect continuum, Kven was neither part of the 19th century process of collecting eastern and western Finnish dialects into standard Finnish nor involved in the creation of a modern Finnish vocabulary, and whereas Finnish speakers understand Kven, the changes Finnish underwent during the modernisation process makes Finnish hard to understand for Kven speakers. The orthographic principles behind the Kven orthography are the same as the ones behind Finnish, but there are differences in the vocabulary, in the inflectional morphology and in some central morphophonological processes. One Norwegian municipality, Porsanger, has declared itself trilingual, with Norwegian, North Sámi and Kven. The central institution carrying out Kven language work is Kvensk institutt, working in cooperation with Giellatekno at UiT.

Status

The ELE report gives a good picture of the status of Kven language technology. The basic language technology resource is the grammatical language model. For Kven this has a coverage for running text at about 82%.

Recent developments

Since the publication of the ELE report, the language technology group at Kvensk institutt has mainly been working on improving the Kven - Norwegian - Kven e-dictionary.

LT in the public sector

Kven is used in school textbooks, in press releases from relevant ministries, and in official documents aimed at the general public in districts with Kven speakers. Textbook writers and most Kven translators use Kven proofing tools and e-dictionaries in their work.

Ongoing or planned LT initiatives

The Kvensk institutt has three focus areas in language technology for the foreseeable future: improving the spellchecker, building a larger corpus of Kven text, and improving the quality of the Kven - Norwegian - Kven e-dictionary.

Recommendations for next steps

The grammatical model for Kven is still not good enough to serve as a reliable spell checker. Both the strengthening of Kven in general and Kven language technology work is dependent upon a reliable grammatical model so improving it should be the first priority. What is needed is expanding the group working on language technology at Kvensk institutt.

2.6 Meänkieli

Introduction

The Meänkieli language, formerly known as Tornedalen Finnish, spoken west of the Torne river in northern Sweden, is a result of the northern Finnish language area being divided when Finland was split off from Sweden in 1809. Like Kven, Meänkieli was not part of the 19th century consolidation of Finnish dialects into standard Finnish. The orthographic principles behind the Meänkieli orthography are basically the same as the ones behind Finnish, the main difference being that Meänkieli adheres to the phonological principle ("write as it is pronounced") in a more consistent way than Finnish. There are also differences in the vocabulary, in some central morphophonological processes and to a certain extent also in the inflectional morphology. Meänkieli is a national minority language in Sweden and 9 Swedish municipalities belong to the Meänkieli administration area (initially, there were 5, cf. "Lag (2009:724) om nationella minoriteter och minoritetsspråk"). The central institution for Meänkieli language planning is ISOF.

Status

With one exception (see the next paragraph), the presentation given in the ELE report still gives a good picture of the status for Meänkieli language technology.

Recent developments

The ELE report stated that "... there are plans to start a cooperation project between ISOF and Giellatekno to develop a spelling checker for Meänkieli". This cooperation has indeed started, with work on a grammatical language model for Meänkieli. At present, it has a coverage for running text at about

87%, and is thus coming closer to a level where it can function as a reliable proofing tool.

LT in the public sector

Meänkieli is taught in schools to a limited extent, and there is one textbook in Meänkieli. It is also used to a certain degree in official information from a variety of public institutions, both on a national level but especially in the 5 Meänkieli municipalities. So far, none of these texts have been produced using language technology tools.

Ongoing or planned LT initiatives

In 2024, ISOF and Giellatekno will set up a detailed plan for releasing a Meänkieli spellchecker.

Recommendations for next steps

The grammatical model for Meänkieli is still not good enough to serve as a reliable spell checker. The strengthening of both Meänkieli in general and Meänkieli language technology work in particular is dependent upon a reliable grammatical model. Improving it should be the first priority. What is needed is an expansion of the group working on language technology at ISOF. Also, efforts should be made to enlarge the corpus of Meänkieli text. Possible work on Meänkieli speech technology should be done in cooperation with providers of speech material (SR, Swedish and Finnish archives) in addition to ongoing work on Finnish speech technology and on speech technology infrastructure.

2.7 Romani

Introduction

Romani is one of the largest minority languages in the European Union, with over 3.5 million speakers worldwide.[9] Even though not all Romani varieties are mutually intelligible, Romani is considered one language. Romani is recognised as a national minority language in Norway and Sweden. Finland, Sweden and Norway have ratified the European Charter for Regional or Minority Language for Romani as a Part II language. In Finland, the Institute for the Languages of Finland has been responsible for language planning for Kale Romani since 1997, and in Sweden, the corresponding body is ISOF.

[9] Matras, Yaron 2002. Romani. A Linguistic Introduction. Cambridge: Cambridge University Press.

In Sweden, estimates of the number of Travellers and Roma vary between 35,000 and 100,000. A majority of the Norwegian Roma are Travellers. In addition, a few hundred Vlax Roma have migrated to the country. Finland has approximately 10,000-12,000 Roma, of which the Finnish Kale constitute the majority.

In Sweden and Norway, the Travellers preserve to some extent a Para-Romani variety referred to as Scandoromani[10],[11] (called Romani Rakkripa in Norway and Resanderomska in Sweden). Kale is used in Finland and Sweden. The 'immigrant' dialects spoken by other Romani groups (perhaps most notably Arli, Kalderash, Lovara) in the Nordic countries are usually similar to the dialects of the regions from which the migrants originated. In Sweden, probably more than twenty Romani varieties are used.

Status

The ELE report on Romani gives a good overview of the situation for Romani language technology. Corpora of written and spoken Kale have been compiled since the 1980s, in particular since the turn of the 21st century. In 2004, Finnish Romani lexical data was added to RomLex.[12] In 2014, linguistic terminology in Romani was contributed to the Bank of Finnish Terminology in Arts and Sciences.[13] There are recent dictionaries and a few grammars of Kale.[14] In Sweden, ISOF is collaborating with

[10] Carling, Gerd, Lenny Lindell & Gilbert Ambrazaitis 2014. Scandoromani: Remnants of a mixed language. Leiden: Brill

[11] Wiedner, Jakob 2017. Norwegian Romani: A Linguistic View on a Minority Language in the North of Europe. Oslo: University of Oslo.

[12] http://romani.uni-graz.at/romlex/

[13] *https://tieteentermipankki.fi/wiki/Termipankki:Etusivu*

[14] Granqvist, Kimmo 2007. *Suomen romanin äänne- ja muotorakenne* [Phonology and Morphology of Finnish Romani]. Suomen Itämaisen Seuran Suomenkielisiä julkaisuja 36. Kotimaisten kielten tutkimuskeskuksen julkaisuja 145. Helsinki: Yliopistopaino;

Granqvist, Kimmo 2011. *Lyhyt Suomen romanikielen kielioppi* [Concise grammar of Finnish Romani]. http://scripta.kotus.fi/www/verkkojulkaisut/julk24/ Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja 24. Helsinki. Published in print by the Finnish Romani Association;

Granqvist, Kimmo & Saarni Laitinen, forthc. *Descriptive Grammar of Finnish Romani*. München: Lincom Europa;

Granqvist, Kimmo 2014 (ed.) *Juho Peltosalmi ja Yrjö Temo. Suomi–romani-sanakirja ja Johanneksen evankeliumi*. [Juho Peltosalmi and Yrjö Temo: Finnish-Romani dictionary and Gospel of John]. Helsinki: Suomen romaniyhdistys;

Hedman, Henry 2016. *Suomi-romani-sanakirja : Suomen romanikielen nykyajan sanastoa*. Helsinki: Helsingin yliopisto.

Gialletekno on Romani keyboards and spell checkers. In Norway, Giellatekno has developed a language model for Romani Rakkripa.

No work has been carried out in the Nordic countries on speech synthesis, speech recognition or machine translation for Romani.

Recent developments

In Sweden, the ISOF is responsible for language planning and disseminating knowledge about languages, dialects, folklore, names and intangible cultural heritage in Sweden. ISOF has funded dictionaries for Romani in the Lexin series, as well as terminology lists for several Romani varieties.

In Finland, a Romani language revitalisation programme[15] will run from 2023 through 2030, and contains a number of measures to promote the use of Kale and increase the teaching of it. Developing LT tools for Kale is included among the undertakings.

LT in the public sector

Various Romani varieties are used in schools, and hence in textbooks, for Romani children in Sweden, Finland and Norway. In addition, official information from a variety of public institutions (health authorities, regional authorities etc.) publish information in all national minority languages, also the Romani ones. So far, none of these texts have been produced using language technology tools.

Ongoing or planned LT initiatives

In Sweden, ISOF and Giellatekno are currently collaborating to develop a spell checker for Romani Arli.

In Norway, Giellatekno has made a language model and spell checker for Romani Rakkripa covering the vocabulary of a recent textbook for the language. Gialletekno has also made a language model for Kale, and a pipeline for a spell checker for Kale. Nothing has been done on language technology for the other Romani variety in use in Norway (Lovara Vlax Romani, referred to as “Romanes” in Norwegian documents).

In Finland, a pilot two-level morphological analyser ROMTWOL for Kale was compiled in 2021-2022 using the PC-KIMMO environment version

[15] OPH 2022 = Suomen romanikielen elvytysohjelma toimenpide-esityksineen 2023–2030. Finnish National Agency for Education. https://www.oph.fi/sites/default/files/documents/Suomen_romanikielen_elvytysohjelma_toimenpide-esityksineen.pdf

2.1.8.[16] This model will need rewriting in order to be used as a proofing tool in modern software. Giallatekno and Kimmo Granqvist have plans to unify their LT tools for Kale, using the current Helsinki Finite-State Transducer (HFST).

In Finland, work on an online dictionary of Kale has been commenced by Granqvist and the Finnish Roma Association. Online games and an online course of Kale are being planned by Kimmo Granqvist, Henna Huttu, and Päivi Majaniemi.

Glottolog name	Ortho-graphy	Documented grammar	Monolingual corpus (words)	Lexical resources	Grammatical language model[17]
Kalo Finnish Romani	Yes	small	352000	Kimmo Granqvist/ Finnish Roma Association/ University of Helsinki	Alpha
Tavringer Romani	Yes	small	42000	GiellaLT	Experiment
Romani arli	Yes	small	50000	ISOF, Arli	Alpha
Romani kalderaš	Yes	small	51000	GiellaLT	Experiment
Romani Lovara	Yes	small	49000	GiellaLT	-
Polish Romani	Yes	small	-		-
Traveller Norwegian	Yes	small	-		Alpha

Table 1: Overview of resources for Romani languages.

The written versions in table 1 above are the ones used in official publications. The standardisation process is still open, both with respect to how many varieties should be made into written standards and to what these standards should look like.

[16] Granqvist, Kimmo 2005. ROMTWOL. An implementation of a two-level morphological processor for Finnish Romani. In Schrammel, Barbara, Dieter W. Halwachs & Gerd Ambrosch (eds.) *General and Applied Romani Linguistics. Proceedings of the 6th International Conference of Romani Linguistics*, p. 150–162. München: LINCOM Europa.

[17] Maturity level is defined here: https://giellalt.github.io/MaturityClassification.html

Recommendations for next steps

Orthographic recommendations for the written varieties in use in the Nordic countries should be clarified, preferably as a result of Nordic cooperation. Thereafter, the main priority should be building grammatical models for each written variety, accommodating electronic dictionaries, and spell checkers. The Romani corpora are very small indeed, thus one should give priority to collect whatever published text there is. The work should be conducted in cooperation with normative bodies where they exist, and in any case with the translators producing public texts in Romani.

2.8 Sámi languages

Introduction

There are 7 standardised Sámi languages spoken in the Nordic countries. They may be classified into three groups according to various converging criteria.

First, there is North Sámi, with close to 20,000 speakers, spoken in three countries. It has a written tradition going back almost three centuries, is spoken by the vast majority in two municipalities and plays an increasing role in public administration, including government service functions. It is an official language in four municipalities in Finland and 8-eight in Norway, and taught as first language in some municipalities both in Norway and Finland, where North Sámi can also be used as the main language of instruction in secondary education. The Swedish Law on National Minorities (2009) declared 17 municipalities to be Sámi municipalities (in 2024 the number is 26), without specifying what Sámi language each municipality covers. The University College in Guovdageaidnu is run entirely in North Sámi, and both Oulu University and UiT the Arctic University of Norway conduct study programmes with North Sámi as the language of instruction. North Sámi is also popular as a foreign language subject in both these and other education institutions. Language shift to the majority languages was still going on some decades ago, and whereas the language shift process is now largely halted, its effect can still be felt. The written standard is from 1979 and its position outside formal educational systems is still weak.

The second group consists of South, Lule, Inari and Skolt Sámi. The first two are spoken in Norway and Sweden and the latter two in Finland. Traditionally, Skolt Sámi was spoken more widely in Russia and Norway. Russia still has some speakers and there are plans for revitalisation work in Norway. These four languages all have more than 300 speakers. To a various degree, the languages have a history of religious texts dating back

to the 18th and 19th century. Their current written standards date back to the 1970s and 1980s and they are all taught as a school subject in their respective areas. To a certain extent, all four languages may be studied both in secondary education and at university level, and there are school curricula for the languages at all levels of education. Inari and Skolt Saami are official languages in the municipality of Inari and South and Lule Saami are official languages in several municipalities in Norway.

The third group consists of Ume and Pite Sámi. They are both spoken in Sweden by less than 50 speakers and they were formerly spoken in Norway as well. Ume Sámi was the dominant written Sámi language in Sweden in the 18th century, its written standard ("Sydlapska bokspråket") covering also the South Sámi area south of Ume Sámi. Pite Sámi, on the other hand, is closer to its northern neighbour Lule Sámi. Ume and Pite Sami got an official orthography in 2010 and 2019, respectively. Pite Sámi is taught as a subject in primary school to some degree.

Compared to other indigenous languages of the same size, all the Sámi languages are very well documented. There are 19th and 20th century dictionaries covering the traditional vocabulary of all the languages and there are reference grammars and dictionaries for the contemporary orthographies for all languages except Ume Sámi. The national broadcasters in Finland, Norway and Sweden co-produce a 10 minute television news programme five times a week, in five Sámi languages, as well as web newsflashes in the majority languages, but sometimes also in one of the five Sámi languages. There is a North Sámi newspaper (Ávvir, 5 editions a week) and several magazines in North Sámi. For Inari Sámi, there is a web-based newspaper (https://www.anarasaavis.fi) with (almost) daily updates as well as a regular cultural magazine. Also for Skolt, Lule and South Sámi, there are articles published in various periodicals. Publishing scientific monographs and articles on Sámi topics is gradually becoming more common, and there are two Sámi scientific journals, publishing mainly in North Sámi, to the extent that doing research on Sámi is about to become impossible without reading skills in North Sámi.

One or more of the Sámi languages have official status in a total of 43 municipalities, in Norway (13), Sweden (26) and Finland (4). The implications of this status vary from country to country and from language to language, but at least part of public signage and official documents are available in the relevant Sámi languages, both in the Sámi municipalities, in their respective counties or regions and on the state level. The Sámi parliaments in Norway and Finland have multi-lingual administrations, including three Sámi languages each.

Status

Work on Sámi language technology has, for the last two decades or so, been conducted by two R&D groups at UiT, the Giellatekno and Divvun groups, constituting a research milieu of around 10 persons. The outcome includes proofing tools and morphologically enriched e-dictionaries for all the Sámi languages except for Ume Sámi, as well as work on e-learning and machine translation. Details are found in the ELE report and will not be repeated here.

The coverage on running text for the grammatical language models ranges between 92% and 97.5%, with Skolt Sámi (82%), Pite Sámi (<80%) and Ume Sámi (no grammatical language model) being the exceptions.

Recent developments

Since the ELE report, the Norwegian government has decided to strengthen the UiT R&D groups with what equals approximately four new positions, mainly within infrastructure, speech technology and lexicography. Work on Sámi language technology has continued, e.g. with publishing grammar checkers for three new Sámi languages.

In Helsinki and Tartu, work has been carried out on neural-based machine translation between Uralic languages and between Uralic and Germanic languages. For the Sámi languages, the MT output often looks good (but with pseudo-words), but the meaning of the output often deviates considerably from the original. No systematic evaluation has been published so far and it is thus hard to assess both status and progress. The Norwegian National library is conducting tests on translating North Sámi speech to Norwegian text, but again, it is too early for an evaluation.

Overall, the basic LT tools are still keyboards, proofing tools, MT and e-lexicography. There are still things to be done on these issues, as well as on upcoming issues such as speech technology.

LT in the public sector

Written Sámi has its stronghold in the public sector. The proofing tools were built for exactly that use and are used extensively when translating documents into the Sámi languages. With increasing use of North Sámi in public documents and mass media, there are reports that MT is used to translate Sámi into the majority languages.

Ongoing or planned LT initiatives

The Giellatekno and Divvun R&D groups at UiT the Arctic University of Norway are continuing their work on grammatical language models, proofing tools and MT. Work on speech technology and lexicography will be strengthened from 2024 on. UiT and the Norwegian national library are planning a large-scale corpus collection project, which will benefit all ongoing work on Sámi language technology.

Recommendations for next steps

Work on grammatical language models should continue for all the Sámi languages and both writers and learners need grammar proofing tools as well as spell checkers. Of the languages with school curricula, Skolt Sámi stands out with proofing tools not good enough to be reliable. A priority task should thus be to improve the basic grammatical models for Skolt Sámi.

Sámi speech technology is in an initial phase and should be promoted.

2.9 Summary of language and resource status

In this section, we present a summative overview of available resources and tools for all the languages discussed in this report. The overview provides numbers and facts, which enables direct comparison between the situation in the various languages. The summary consists of three tables.

Table 2 provides an overview of available monolingual text corpora. This type of resource is foundational when working with language and language technology. The availability of corpora is brought forward in this table to illustrate the language resource status in a clear yet significant manner. For each language, an estimate is given of the maximum size of a text corpus in the given language, i.e. how much text is currently produced in the language in total. These estimates should be understood as very rough and given only to provide an overall understanding of the progress each language has made so far, and what could be seen to be the target size (limit). Mentioning an estimated limit also underscores the issue that these languages will by definition experience a maximum in data volumes, as they all have few speakers and thereby relatively low text production, compared to other languages. This will be an ongoing issue, also if the languages actually reach the target corpus size.

Language	**Monolingual corpus (M words)**	**Estimated limit of corpus size (M words)**
Faroese	65	400
Greenlandic	29	100
Icelandic	2 400	12 000
Karelian	0.50	2
Kven	0.69	1.5
Meänkieli	0.74	1.5
Romani	0.54	1
Inari Sámi	3.2	5
Lule Sámi	1.8	3
North Sámi	39	120
Pite Sámi	-	0.1
Skolt Sámi	0.25	0.5
South Sámi	2	5
Ume Sámi	-	0.05

Table 2: Overview of the size of available monolingual corpora together with a rough estimate of corpus size limit. Numbers are given in million word forms/tokens.

Table 3 shows the situation for selected basic language resources and their availability. These resources are prerequisites for the tools outlined in Table 4.

Language	**Standardised orthography**	**Documented grammar**	**Lexical resources**	**Speech corpora**	**Parallel corpora, words L1**
Faroese	yes	yes	good	ASR: 65h+ TTS: yes	en-fao: ~150 k
Greenlandic	yes	yes	medium	ASR: zero	< 0.1
Icelandic			good	ASR: 145+ hours[18] TTS: 85+ hours[19]	en-is: 3.5M + multiple web scraped
Karelian	yes	yes	low	-	-
Kven			low	-	nb-fkv: ~ 200 k
Meänkieli			medium	-	sv-fit: ~ 7 k
Romani	yes	See table 1	See table 1	-	See table 1
Inari Sámi	yes	yes	medium	-	fi-smn: ~370 k se-smn: ~163 k
Lule Sámi	yes	yes	low	ASR: 27 hours TTS: 20 hours	nb-smj: ~160 k se-smj: ~190 k
North Sámi	yes	yes	medium	ASR: 50+ hours TTS: 15 hours	nb-se: ~3.5M
Pite Sámi	yes	(yes)	low	-	-
Skolt Sámi	yes	yes	medium	TTS: 12+ hours	fi-sms: 1 k
South Sámi	yes	yes	low	-	nb-sma: 197 k
Ume Sámi	yes		low	-	-

Table 3: Overview of availability of basic resources for language technology

[18] There are multiple Icelandic corpora available for training ASR systems. Here, we refer to the most recent one, Samrómur.

[19] This number refers to the largest Icelandic speech corpus for training TTS systems, Talrómur.

Table 4 presents an overview of availability and maturity of selected language tools for each of the languages in this report. The distribution of available tools is directly related to the availability of resources, as seen in Table 3.

Language	MT	Grammar model, spell checker[20]	TTS	ASR	Generation	Keyboard
Faroese	multilingual, beta	production	yes	yes	no	yes
Greenlandic	beta	production	yes	no	no	yes
Icelandic	multilingual	production	yes	yes	yes	yes
Karelian	alpha	beta	no	no	no	yes
Kven	alpha	production	no	no	no	yes
Meänkieli	alpha	beta	no	no	no	yes
Romani	-	See section 2.7	no	no	no	no
Inari Sámi	alpha	production	no	no	no	yes
Lule Sámi	alpha	production	alpha	no	no	yes
North Sámi	alpha	production	yes	alpha	no	yes
Pite Sámi	-	beta	no	no	no	yes
Skolt Sámi	alpha	beta	no	no	no	yes
South Sámi	alpha	production	no	no	no	yes
Ume Sámi	-	-	no	no	no	yes

Table 4: Overview of availability and maturity of selected language tools.

These three tables provide an understanding of the overall situation for each language as well as a comparison between the languages. This factual overview highlights the areas on which developmental efforts should be focused.

[20] Maturity level definitions for grammar models / proofing tools are given in https://giellalt.github.io/MaturityClassification.html

3. Methodology

In recent years, the trend in language technology has been to build larger and larger models, using neural network models that need to train on large datasets. In order to work for a given language, the datasets for that language have to contain hundreds of millions of words, preferably billions. These approaches have proved to be powerful for some areas of LT, such as machine translation (MT) and text generation, as well as automatic speech recognition (ASR) and text-to-speech (TTS), given that large enough language resources exist. When working with low-resource languages, such amounts of data simply are not available (the bottleneck being the size of the population) and therefore rule-based approaches can be expected to be competitive for the foreseeable future. In the ELE Report, it was thus recommended that the Nordic minority languages should focus on the rule-based approach.

Rule-based

Rule-based approaches do not rely on corpora to learn from or datasets from which they can draw statistical information. For each problem and language, a set of rules has to be defined. Rule-based approaches can be applied to intervene with a certain set of issues, as in proofing tools, or to systematically analyse and rewrite language, as in the case of machine translation.

Machine learning

Machine learning (ML) approaches in LT use statistical methods, patterns and frequencies to analyse and generate language. They typically require well curated data sets to supervise the ML algorithms.

Neural

Neural approaches rely on large collections of unstructured data, and learn from patterns in the data. They have proved to be able to accurately analyse and generate language in cases where abundant data is available. When given information on multiple languages, they can sometimes transfer their ability to work with related languages.

In light of recent rapid developments in artificial intelligence, the present work group is not as sceptical towards neural approaches to low-resourced languages as the ASTIN workgroup was when it wrote the ELE report. The major bottleneck is still there, but for some languages and tasks it may be possible to compensate for it to a certain extent. Both rule-based and neural models may be facilitating factors for synthetically increasing data volumes (augmentation, back-translation), and can be leveraged for the mechanisms of transfer learning and fine-tuning of larger models (both text and speech).

For Icelandic, experiments with neural methods have produced good results, for example for part-of-speech-tagging, named entity resolution, machine translation and text generation, as well as speech synthesis and speech recognition. The text processing approaches usually require very large datasets for unsupervised training, while the speech processing approaches require specialised speech data that can be collected for any language. Furthermore, experiments have shown that training models on related languages reduces the need for data for the individual languages, as has proved to be the case for Faroese in models trained on large Icelandic datasets, as discussed in Section 2.1. This effect could possibly also be relevant for other languages discussed in this report, for example for the Sámi languages (with models trained on Finnish) and perhaps for Romani (for example with models trained on Farsi or Hindi), but that is a topic of research which needs to be explored.

One should, however, not be blind to the fact that output from data-driven technology will never be better than the input. Applications aiming at improving text production and text quality cannot be based on low-quality training data. The assumption behind using machine learning for normative computer programs is that the corpus represents the norm. When this is not the case, the attempt fails. The necessity of rule-driven technology for a long list of applications will therefore also exist in future LT development alongside machine learning for developments for which such technology makes sense.

The main point to be taken from this section is that new advancements in machine learning and neural techniques can be beneficial if leveraged correctly, and should be explored. However, these new techniques will not solve the fundamental issues with LT in languages with few speakers. The rule-based approach upholds its value and concurrently demands investments. For current large language models to be effective, they are trained on text collections typically containing billions of words. As made apparent in Table 2, not only are datasets of such sizes unavailable for all the languages discussed in this report, perhaps with the exception of Icelandic, but they also will not become available, as not enough text that can potentially be digitised has been written to make large language models viable for these languages.

4. Opportunities for collaboration

This report covers 20 different languages, spoken either in the same countries or in Nordic countries sharing cultural and political values. With relatively few speakers, these languages face difficulties due to small amounts of both data and experts. These languages need to approach language technology in a different manner than languages with many speakers. Collaboration between these languages with similar settings can be a way of compensating for scarcity of resources.

In this section, we first put forward some suggestions for collaboration opportunities in general before providing examples of previous fruitful collaborations.

4.1 Collaboration on work

Education

A major problem for all the language communities covered in this report is the lack of language technology education relevant to the challenges faced when building tools for minority languages. This could be done e.g. in the form of dedicated courses on a Nordic level. What is needed is basic text scripting skills, education for writing and maintaining rule-based systems, and knowledge on how to build and evaluate neural models.

For larger language communities, the university sector typically does research, whereas the large companies make the practical programs. For the small language communities, everything must be done outside the large companies. This means that the smaller language communities have broader educational needs

Infrastructure

Building language technology for small language communities is expensive. An obvious way of making it possible is to cooperate on language-independent infrastructure. The infrastructure work financed for the Sámi languages by the Norwegian Ministry of Local Government and Regional Development has made it possible to produce proofing tools for almost all Nordic minority languages. From a Nordic perspective, this work should be continued and further strengthened. Similar cooperation should also be conducted for speech technology, as speech technology is being developed for two Sámi languages, with the explicit goal of also developing that work

into a general infrastructure for minority languages. This work should be supported, and it could especially benefit the smaller less-resourced languages. A broader collaboration on the infrastructure could facilitate work on speech recognition in other languages.

Exchanging knowledge and methods

The different countries and different languages have different kinds of expertise and knowledge. Collaborating on projects will in itself ensure the exchange of knowledge and methods, and specialists in different areas can combine their efforts. One example could be a collaboration between those working with rule-based approaches and others working with neural or machine learning approaches to build hybrid tools.

Research collaboration

The challenges faced by small language communities are both similar and deviate from the ones faced by larger language communities. A written norm less reliably reflected in texts, small text and speech corpora, limited access to and no support from big companies all contribute to a situation where small language communities have needs deviating from larger language communities. This calls for research to meet these needs. Such research should preferably be done in collaboration between several communities with similar settings.

Network

The exchange of knowledge and research collaboration is also beneficial in a broader sense. Not only do the languages we discuss here face somewhat similar challenges in data scarcity and development, but they also have comparable linguistic and societal conditions as well as similar limitations. These conditions affect the work on language technology, decisions and strategies. Overall, collaborations have the advantage of creating professional networks that can be leveraged for support, discussions and sharing experiences, which in turn can benefit the local environments for language technology and language politics.

4.2 Collaboration on content and data

Grammatical language models must contain the full lexicon of the language in question. For languages with common geographical and cultural domains, there will be large overlap in vocabulary, be it proper nouns (place names and person names) or loanwords from common sources.

The Nordic countries have a strong tradition of lexicography and terminology, disciplines that to an increasing extent are being integrated into language technology. This holds especially true for the Nordic state

languages, but some less-resourced languages (such as Faroese) also have a strong lexicographic tradition. Cooperation within this field, both with respect to methodology and infrastructure but also when it comes to access to lexical and terminological resources, should be given priority.

Leveraging typological similarity when building language models

The languages discussed in this report come from several language families. The languages that are related should explore collaboration on data and methods. The training method of transfer learning is a particularly interesting focus, as this can lower the data volume threshold for quality model and tool development.

4.3 Examples of previous and ongoing collaborations

The Giellatekno team in UiT (The Arctic University of Norway) has created an infrastructure for building text processing tools, containing many resources and tools for the languages in their region. The infrastructure is published under an open licence and hosts resources and tools built in cooperation with relevant R&D groups for almost all languages in this report. For example, the Faroese spell checker, a very fundamental tool, was developed in collaboration between the University of the Faroe Islands and UiT, and is still kindly hosted and maintained by UiT. The same goes for the other languages in this report (save Icelandic). The tools for these languages would not be available if not for this collaboration. This is an example of collaborating on expertise, infrastructure and lexicography in a common project.

Another example is the ongoing project to develop a large Faroese text corpus, the Faroese MegaWord Corpus. This project is a collaboration between The University of the Faroe Islands and The Árni Magnússon Institute for Icelandic Studies. The Faroese corpus is being developed using the same approach and infrastructure as the Icelandic Gigaword Corpus. Leveraging the experience of the Icelandic project makes the process of the Faroese project more efficient.

With the exception of Icelandic and Faroese, there have so far been few and limited attempts at making neural models for the languages covered in this report. What has been tried out is neural machine translation between Uralic languages (in Helsinki and Tartu), and in both cases part of the text resources has come from the open text corpora at UiT.

These examples of collaborative efforts and outcomes clearly demonstrate the possible success of combining forces in language technology work. However, the challenge of collaborative projects like these is the continuing

need for maintenance and ongoing development. Collaborative projects typically have limited funding for a certain time period only, and this can result in otherwise successful projects coming to a halt. Resources and tools need concurrent updates, both technical and linguistic, and such maintenance should be incorporated in funding.

5. Recommendations

We recommend a **continued focus and investment** in language technology for small languages in Nordic language policy

We recommend for each language to **initiate language technology strategies** and implementation

We recommend **collaboration on the political level** to advocate accessibility on major platforms.

We recommend that the **collaboration on research and development** within language technology be continued and intensified

We recommend **collaboration on creating educational programmes/courses** for developing and maintaining local expertise

We recommend **introducing legislation** that facilitates the collecting of corpus material

As part of the task given, the present working group has summed the report up to include the following six recommendations.

a. **We recommend a continued focus and investment in language technology for small languages in Nordic language policy**

It is widely accepted that sufficient language technology for survival in cyberspace is one of the most decisive factors for any language's vitality. In recent years this has been considered within each Nordic country as well as on a supranational Nordic level. We trust that such political support plays an important role in explaining why the small languages in the Nordic countries are comparatively healthy compared to other areas of the world. The Nordic countries are in a leading position internationally.

b. **We recommend for each language to initiate language technology strategies and implementation**

In order to make explicit what is needed for each language (or in some cases for each group of similar languages) and how to achieve it, strategic

plans should be produced for the languages that are currently without one. This has proven fruitful for e.g. Icelandic.

We recommend that the strategic plans be made under the auspices of the respective normative bodies and/or ministries, as a cooperation including also the language communities as well as language technologists working on the various languages. This coordination is essential for each language as the requirements and priorities vary from language to language.

c. We recommend collaboration on the political level to advocate accessibility on major platforms.

The ELE report puts forward some recommendations, which the present working group supports. We specifically want to reiterate the recommendation for working toward accessibility on the major tech platforms: Major technology companies make the possibility of integrating our language technology solutions into our computer, phones and other digital devices increasingly difficult. Getting such access is a political problem and should be dealt with on a Nordic or perhaps even EU level. This obstacle was mentioned in the ELE report, but the situation remains the same.

d. We recommend that the collaboration on research and development within language technology be continued and intensified

In sections 4.1 and 4.2 we list various opportunities for collaboration between the Nordic national and indigenous languages. Carrying out collaboration as described will mutually exploit individual proficiencies and expertise to reach better results. With the approximately 20 languages covered in this report, it is immediately clear that we should cooperate as much as possible. Even disregarding the need to work as cost-efficiently as possible, a major obstacle is still the availability of qualified philologists and language technologists.

e. We recommend collaboration on creating educational programmes/courses for developing and maintaining local expertise

One crucial issue for the Nordic languages with few speakers is educating experts to ensure continuing development in the field. Very few educational programmes exist for language technology, especially programmes that include rule-driven technology, which is still crucial for successful language technology for smaller language communities with limited amounts of data. Having local specialists conduct research is instrumental in maintaining

development. An inter-nordic collaboration on educational programmes or courses targeted at small languages would be an effective support of the small languages in the Nordic region.

f. We recommend introducing legislation that facilitates the collecting of corpus material

Both for rule-based language technology and even more so for neural technologies, linguistic resources in the form of collection of text and speech is crucial, and in many cases also the bottleneck for further development.

This legislation could be in the form of an obligation to make publicly funded publications available for language technology research and development.

6. Conclusions

In our report, we have examined the status of Faroese, Greenlandic, Icelandic, Karelian, Kven, Meänkieli, Romani, and the Sámi languages in the context of language technology.

The languages addressed in the report exhibit variations in corpus size, availability and maturity of LT tools. Robust corpora are readily accessible for Icelandic, whereas there are somewhat robust corpora for Faroese, Greenlandic, and North Sámi, contrasting with the relatively limited corpora for Skolt Sámi, Romani, Karelian, Kven, and Meänkieli. Icelandic boasts an array of lexical resources and tools for text analysis, speech processing, and machine translation. For Faroese, the dictionary situation is good, and there are tools for text analysis, proofing and speech processing. In the case of Greenlandic, endeavours have been undertaken to enhance the fundamental resources, fortifying the linguistic groundwork for the language. For the Sámi languages (except Ume Sámi), and Kven, grammatical models have been made. Most of these are of high quality, but all need to be updated continuously. For Karelian and Meänkieli, the grammatical language models still require substantial work. Among Romani varieties, early stages of developmental efforts for grammatical language models and spell checkers are concentrated on Kale (in Finland), Arli (in Sweden/Norway), and Romani Rakkripa (Norway), while other varieties such as Kalderaš, Lovara, and Polish Romani are yet to receive attention.

The availability of LT tools is closely tied to resourcing and institutional support. In Iceland, the government-funded Language Technology Programme for Icelandic (2019-2022) has played a pivotal role in the success of existing language resources and tools. Faroese has seen major developments through the establishment of a funded institution for LT research and development. Sámi language technology has been advanced by researchers at UiT, Giellatekno, and the Divvun group, forming a research milieu of approximately ten individuals. These groups have also contributed to the development of grammatical models for Kven, Greenlandic, and Romani varieties. In Finland, dedicated full-time resources for LT of the languages in this report are lacking.

There is a need for adopting LT in the public sector as well as in industry and for the general public, and this need is as urgent for languages with few speakers as for major languages. The public sector in the Faroe Islands incorporates LT to some extent, while Greenlandic LT garners attention across various public sector domains. Icelandic companies have embraced the use of LT tools in their workflows, and plans are underway in Finland for online games and an online textbook of Kale Romani. Notably, proofing tools for Sámi are actively employed by translators.

Key concerns for LT in the languages discussed include broad applicability, functionality, and dissemination of solutions, along with challenges related to collaboration with major IT corporations such as Microsoft, Apple, and Google. Emphasis is placed on the imperative need to enhance grammatical language models, particularly for Karelian, Kven, Meänkieli, and Romani. Considerable efforts will be required, especially in the development of speech technologies for these languages.

In conclusion, despite the challenges and needs, language technology has experienced substantial growth in recent years, firmly establishing itself as an integral component of Nordic language work. Effective collaboration stands as a cornerstone for fostering development.