EUROPEAN LANGUAGE EQUALITY

D1.38

Report on the Nordic Minority Languages

Authors	Sjur Nørstebø Moshagen, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, Per Langgård

Dissemination level	Public
Date	09-05-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.38
Deliverable title	Report on the Nordic Minority Languages
Type	Report
Number of pages	51
Status and version	Final (*Note: this document is not a contractual ELE deliverable.*)
Dissemination level	Public
Date of delivery	Plan: 28-02-2022 – Actual: 09-05-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Sjur Nørstebø Moshagen, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, Per Langgård
Reviewers	Jane Dunne, Maria Giagkou
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ASR	Automatic speech recognition
CLARIN	Common Language Resources and Technology Infrastructure
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
EU	European Union
HPC	High-Performance Computing
ISOF	Institutet för språk och folkminnen (Institute of language and folklore)
LT	Language Technology/Technologies
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
SOV	(the word order) Subject Object Verb
SVO	(the word order) Subject Verb Object
DVT	Swedish Television
TTS	Text-to-speech
UiT	The Arctic University of Norway

Abstract

The present report provides an overview of regional and minority languages with very few speakers in the Nordic countries and addresses a number of challenges specific to LT development for very small languages. The elementary facts that a language is a whole language notwithstanding the number of its speakers and that the costs of developing LT for very small language communities do *not* come with a discount are omnipresent when developing LT for small languages. They need to be considered in every step of the development process.

These considerations include the critical demand that projects must be anchored locally and be accessible locally to secure sustainability of the programs and successful dissemination of them to the end-users they are intended for. Considerations also include the need of broad versatility in everything developed for very small language communities. With the scarcity of economic and human resources as well as data resources that exists in all the languages dealt with here, launching a large number of isolated projects simply is unviable. It is therefore a condition that development should focus on providing basic resources that with limited efforts can be expanded into several different applications.

The report also deals with the exclusion of small languages in much of the digital sphere, where large companies dominate and whose platforms do not allow for inclusion of third party LT solutions. Inclusion of smaller languages with LT of low market value in the digital sphere should be secured by large international bodies such as the Nordic Council and the EU.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and it seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – to identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners, experts in more than 30 European languages have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed in the frame of the European Language Equality (ELE) project.[2] With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

The authors of this report would like to emphasise that small language communities and minority languages enter the digital age with different resources and prerequisites. The technological optimism that surrounds new language technology and the need for smaller data sets for machine learning, can not be transferred to the languages in this report. Given the typology of most of the languages in combination with the absolute scarcity of resources, at present, it is quite simply not realistic to produce functioning language technology for these

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

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

languages by the same methods used for larger language communities. It is, however, realistic to develop a functional set of tools with the resources that either are available already or which can be produced by the means outlined here, and slow down, or even reverse the current trend towards digital inequality for small language communities.

2 The Nordic Minority languages in the Digital Age

The Nordic countries are home to about twenty minority languages. The languages are briefly described in the following sections, one section per country, to give an overview of their status in the digital age.

Nordic sign languages are not included in this report. The authors do not have the expertise required, and the technology is, as far as the authors are aware, not mature enough. Even though some resources do exist, such as sign language corpora and lexicons for some of the sign languages, a proper consideration of language technology for these languages has to be deferred to another report.

Three of the official minority languages are at the same time majority languages in another country: Finnish (minority in Sweden, majority in Finland), Swedish (minority in Finland, majority in Sweden) and German (minority in Denmark, majority in Germany and other countries). These languages are covered thoroughly by the respective country reports, and will not be further dealt with here other than in passing. It should be mentioned, though, that for language technology tools to work properly for each of these minorities, the tools must be adapted to the specifics of each minority society. Adaptations could be country specific terminology, speech technology adaption and more.

The declaration on a Nordic language policy,[3] states that:

> "Multilingualism provides the basis for skills, creativity, perspective, and international contacts to an extent that is impossible in monolingual societies. Developing it requires a unified, long-range, and effective language-policy effort.
>
> Nordic language policy is based on all Nordic residents having the right:
>
> - to acquire both spoken and written skills in a language essential to society, so that they can participate in the workings of society
> - to acquire an understanding of and skills in a Scandinavian language and an understanding of the other Scandinavian languages so that they can take part in the Nordic language community
> - to acquire a language of international importance so that they can take part in the development of world society
> - to preserve and develop their mother tongue and their national minority language."

The declaration is from 2006, when the language technology situation was very different from today's where LT is an essential prerequisite for a language.[4] However, the intentions behind the declaration is clear:

> "Nordic language policy has a responsibility to world society to see that in particular the languages that are not national languages anywhere continue to live and develop, and that all minority languages can continue to exist. It is important that sign language also be granted a strong position."

[3] https://www.sprakradet.no/globalassets/spraka-vare/deklaration-om-nordisk-sprakpolitik-2006.pdf
[4] It should be noted that a revision of the declaration is in preparation

In order to live up to the responsibility from the declaration of 2006, and given the important position of language technology in today's digitalised society, extra support is needed for minority languages to prevent their digital extinction. In today's situation, individuals as well as the language societies for those who speak a number of the minority languages dealt with in this report, are kept out of the digital sphere. For some languages, this is due to the lack of technological readiness. For other languages, such as some Sami languages and Greenlandic, their exclusion is not due to the lack of language technology – because the technology exists – these languages are being kept out of the large international platforms, even when the technology is in place. An example of this is access to existing proofing tools within web-based office solutions, where third-party proofing tools are not allowed in any form resembling that what users expect. While it is fully understandable that the major players do not want to allow unknown third-party software on their servers, one can easily imagine alternative solutions that would give users the expected user experience. This has not happened, for reasons unknown to the authors.

Both issues above must be dealt with if we are to ensure that minority languages "continue to live and develop" and "can continue to exist" (cf footnote 3).

The status and digital readiness of the minority languages in this report vary. They span from languages that do not have the linguistic resources needed to build language technology via languages that have some resources, but no technology built, languages which have the technology, but whose usage is prevented by monopolising language solutions, to a language such as Faroese, which is the official language of a specific geographic area, and has a very different status and digital readiness.

For those languages that lack fundamental linguistic resources such as a standardised orthography, dictionaries and written text, these resources need to be in place before language technology is possible. Languages that do have such basic linguistic resources can build language technology if the language community is interested in a digital existence. The lack of resources as well as missing technology can (but need not) be dealt with at a local level, if the financing is in place and the expertise needed for the technological development exists.

However, all small language communities, even those that do have functioning language technology, have problems in accessing the big platforms. By access, we mean for instance the possibility of adding new language applications, such as a Sami spell checkers, to already existing programs, such as the Microsoft Office web apps. By preventing access for smaller languages to these platforms, we find ourselves in a situation where international companies are in fact monopolising access to language. Because these big platforms are bought by the public sector and used in public administration, schools and cultural enterprises, the very same public institutions which are supposed to protect and vitalise the minority languages, end up contributing to the monopolisation of access and to the exclusion of these languages from the digital sphere.

Access to the big platforms is not something that can be solved locally, or even at a national level. The low number of speakers make these languages economically less attractive, and they do not have the political muscle to demand such access. This is probably true even if national policies support such access. Potential political bodies that could help solve these issues could be the Nordic Council or the European Union.

2.1 Faroese islands

There are two languages spoken on the Faroese islands: Faroese and Danish. Danish is covered in a separate report, and will not be further mentioned here.

The Faroese language is a separate branch of the North Germanic languages. It is spoken by about 50,000 people, mainly on the islands, some in Mainland Denmark.

The normative body for Faroese is Málráðið[5], and the language codes is `fo/fao`. It is written using the Latin alphabet, and is used in all of the society and the whole educational system.

Faroese is well established in the digital sphere for websites under the `.fo` top domain. There is basic support for writing the language (i.e. a keyboard) in all operating systems. There is also a spell checker available developed in cooperation between Faroese language authorities and UiT The Arctic University of Norway, in Tromsø. Although no exact figures have been accessible, it is assumed that Internet access is widespread, as in the rest of the Nordic countries.

2.2 Finland

The official minority languages of Finland are listed in Table 1.

Language	Language code	Appr. no of speakers	Official status	Normativity body
Karelian	`krl`	5,000	National minority	Univ. of Joensuu
Romani Kale	`rmf`	4,000	National minority	KOTUS
Sámi, Inari	`smn`	300	Indigenous lang.	Giellagáldu
Sámi, North	`se/sme`	25,000[6]	Indigenous lang.	Giellagáldu
Sámi, Skolt	`sms`	300	Indigenous lang.	Giellagáldu

Table 1: Official Minority Languages of Finland

The Sami languages are protected by law, as indigenous languages. But Finland has no legal definition of national minority, as far as the authors have been able to determine. Karelian and Romani are included in this report based on the following criteria: a normative body exists in Finland; and the languages are mentioned in the political programme for the current government. The idea is that these two facts (mentioned in the governmental programme, and having a normative body) is a strong endorsement and expression of the will to support these languages, and constitutes a de facto definition of a national language minority.

All languages in Finland use the Latin alphabet.

2.3 Greenland

According to the Self-government Act of 2009 there is only one official language in Greenland, Greenlandic (iso kal). Danish language, though, has a special status in Greenland after more than 300 years of co-existence. The normative body of Greenlandic is fixed by law as Oqaasiliortut/ Language Council.

No reliable, official census of the number of speakers exist but there can be no doubt that Greenlandic is prolific with a majority of children born and raised as monolinguals in Greenlandic before being introduced to Danish as first L2 in compulsory school. For a rough estimate there are about 40,000 L1 speakers of Greenlandic in Greenland.

Greenlandic is not endangered according to UNESCO's taxonomy. It is the first language in all public affairs including law and administration. It is compulsory in childrens' school and widely used in secondary and tertiary education.

[5] http://malrad.fo
[6] Total in all countries

Greenlandic has had a national ortographical standard since 1851. It is a polysynthetic Inuit language, and member of what used to be termed the Eskimo/Aleut family of languages. There are three main dialects in Greenland with West Greenlandic being the by far biggest and the base for the standard orthography. West Greenlandic is 'Vulnerable' according to UNESCO. East Greenlandic is spoken on the East coast by (estimated) 3,000 persons. In the high Arctic around Qaanaaq the archaic dialect of Inuktun is spoken by (estimated) less than 500 persons. Both East Greenlandic and Inuktun are 'Definitely endangered' according to UNESCO.

2.4 Norway

The national minority languages of Norway are Kven, Romanes and Romani. In Norway, *Romani*, sometimes referred to as *Romani rakkripa* or *Scandoromani* is spoken by Norwegian Travellers (Norwegian: *Romanifolket, tatere*) and is similar to Romani Tavringer, spoken in Sweden. According to the new Language act of 2022, section 6, "as expressions of language and culture, Kven, Romani and Romanes are equal in value to Norwegian." Under the same act, section 5, Sami languages are recognised as indigenous languages and have standing equal to Norwegian under Chapter 3 of the Sami Act.[7] Norway has ratified the European Charter for Regional or Minority Languages and all the above mentioned languages are protected under the Charter.

Language	Language code	Appr. no of speakers	Official status	Normativity body
Kven	fkv	2,000-8,000[8]	Minority language	Kvensk Språkting
Romanes	rmy	N/A	Minority language	None
Romani rakkripa	rmg	N/A	Minority language	None
Sámi, Lule	smj	1,000[6]	Indigenous language	Giellagáldu
Sámi, North	se/sme	25,000[6]	Indigenous language	Giellagáldu
Sámi, South	sma	600[6]	Indigenous language	Giellagáldu

Table 2: Official Minority Languages of Norway

All minority languages in Norway use the Latin alphabet. The Sámi languages and the Kven language all belong to the Uralic language family, and are characterised by complex inflectional and derivational morphology, and also highly complex morphophonology. Syntactically the Sámi languages have historically been SOV, but to various degrees they are nowadays following a more Germanic SVO word order, with South Sámi being the most conservative, and still very much SOV.

The two Romani languages of Norway are quite different from each other. What in Norway is called Romanes is actually a Kalderash dialect of Vlax Romani. Romani Rakkripa, on the other hand, has been spoken in Norway for centuries, and has swapped most of its morphology with the Norwegian counterpart.

Kven and the Sámi languages are all used digitally. There exists keyboards and proofing tools for all of them, and North Sámi even has a grammar checker and machine translation engine for translation to both Norwegian Bokmål and other Sámi languages.

To our knowledge, the Romani languages of Norway are hardly used digitally. There is no official body responsible for any them, although the Norwegian Language Council follows their development, and they are used to a certain extent in primary education.

[7] https://lovdata.no/dokument/NLE/lov/2021-05-21-42
[8] https://www.kvenskinstitutt.no/kvener/

2.5 Sweden

Since year 2000, five languages are officially recognised as national minority languages in Sweden: Finnish, Yiddish, Meänkieli, Romani and Sami. Sweden has a Language Act[9] and an Act on National Minorities and National Minority Languages[10] stating language status and rights for the languages in Sweden. Note that two of these languages (Sami and Romani) are macrolanguages, both of them covering 5 different written languages or varieties. For a short overview of the status in legislation and in practice for the national minority languages in Sweden, see (Ekberg, 2010).

Language	Language code	Appr. no of speakers	Official status	Normativity body
Meänkieli	fit	N/A	National minority	ISOF
Romani chib[11]	rom	N/A	National minority	ISOF
Sámi, Lule	smj	1,000[6]	Indigenous lang.	Giellagáldu
Sámi, North	se/sme	25,000[6]	Indigenous lang.	Giellagáldu
Sámi, Pite	sje	50	Indigenous lang.	Giellagáldu
Sámi, South	sma	600[6]	Indigenous lang.	Giellagáldu
Sámi, Ume	sju	10	Indigenous lang.	Giellagáldu
Yiddisch	yid	N/A	National minority	ISOF

Table 3: Official Minority Languages of Sweden

The Sámi languages are historically spoken in an area from Härjedalen in Middle Sweden to the Northern Swedish border. Meänkieli is historically spoken in the Torne River Valley and the mountain areas around and between Gällivare and Kiruna. Finnish is historically spoken in Stockholm and Värmland, in the district of Mälardalen and in the big cities in general. Speakers of Romani live in the three largest cities which is also the case for Yiddish-speakers.

Of the five minority languages, South Sami and Meänkieli are most severely threatened. There are few young speakers and the bilingual education must be strengthened in order to protect the languages. With regards to Romani (and also Meänkieli) there is a great demand for documentation and standardisation.

It is difficult to estimate the number of speakers of the different minority languages since Sweden does not collect official statistics about this. The figures below are coarse estimations by (Parkvall, 2015). Finnish is the second largest language in Sweden with approximately 175,000 speakers. The estimates give about 20,000 – 25,000 speakers of Meänkieli, 10,000 – 20,000 speakers of Romani, and about 7,000 – 10,000 speakers of Sami; three quarters of these speak North Sami, around 15% speak Lule Sami and only 10% speak South Sami. Yiddish is spoken only by around 4,000 people of which about 1,000 have Yiddish as their mother tongue.

According to the Language Act, speakers of the national minority languages have special rights to their language. Speakers of Finnish, Meänkieli and the Sami languages have extended rights in the administrative districts where these languages have had a long history. Yiddish and Romani are territorially independent minority languages with more generally-worded protection provisions. Swedish sign language is not defined as a national minority

[9] SFS 2009:600
[10] SFS 2009:724
[11] ISOF presents 7 Roman varieties in Sweden, 5 of them with written standards: Arlikane, Kelderašicka, Lovaricka, Polsko romanes, Tavringer, (Gurbetikane, Kalikane).

language by the EU's minority languages convention, but it bears the same status in the Language Act as the national minority languages and should therefore be granted equal protection.

The responsibility for monitoring the language situation and the implementation of the Language act lies on the Institute of language and folklore (ISOF) – the official language planning and policy organisation in Sweden. As a guidance for implementing the legislation in practice, the LC has produced a guidebook containing information on language rights and practical advice for complying with the new requirements by making information and services accessible for all citizens, including national minorities. It has started to survey the minority language situation and the need for developing language technology resources for national minority languages (Domeij et al., 2019). Available resources are made accessible through the National Language Bank in cooperation with The Royal Institute of Technology and the University of Gothenburg.

3 What is Language Technology?

Natural language[12] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialised field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, *Language Technology* (LT). While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably Artificial Intelligence (AI)), mathematics and psychology among others. In practice, these communities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

Language Technology is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in Machine Learning (ML), rule-based systems have been displaced by data-based ones, i. e., systems that learn implicitly from examples. In the recent decade of 2010s, we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

In recent years, the LT community has been witnessing the emergence of powerful new deep learning techniques and tools that are revolutionising the way in which LT tasks are

[12] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of Artificial Intelligence (AI) has been possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power, and 4) application of simple but effective self-learning approaches.

LT is trying to provide solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e., the generation of speech given a piece of text, Automatic Speech Recognition, i. e., the conversion of speech signal into text, and Speaker Recognition.
- **Machine Translation**, i. e., the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.
- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e., the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction** which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). A very popular application within this area are conversational agents (better known as chatbots).

LT is already fused in our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just very few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance, for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[13]

[13] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is

4 Language Technology for the Nordic Minority Languages

While the description of language technology and artificial intelligence in our everyday lives in Chapter 3 above holds true for users of many majority languages, it does not hold for speakers of the Nordic minority languages and their communities. A native speaker of a Sami language may use a common tool like a spell checker without even thinking about it, but only in Norwegian, Swedish or another majority language. Shifting to his or her native language, the spell checker might not be available. And while language technology in general has a fast growing economic impact, localised LT for lesser used languages does not have a high market value, and is usually created by public funding. The languages in this report have not taken part in the rapid evolution of LT, machine learning and AI for a number of other reasons than purely economical ones.

As pointed out by Wiechetek et al. (in press), for the time being rule-based technology is the dominant paradigm for the Nordic minority languages, and will continue to be the foundation for language technology solutions for them, although combinations of rule-based and machine learning technologies in a hybrid setup could be a useful complement in the future or for specific domains.

Currently, rule-based language technology is playing a major role in developing working LT solutions for these languages in initiatives such as the Apertium[14] and the GiellaLT[15] projects. The rich morphology of the languages leads to a high type-token ratio in corpora, where even relatively common inflectional forms for not so frequent lemmas can be completely absent in available text collections. And free compounding adds another layer of morphological complexity. In combination with minimal, and in some cases non-existing, text resources, each word form will appear less frequently (if at all) compared to a language such as Swedish, which has little morphology and plenty of text. There will be little material to learn from, and very little information on each word form, making machine learning very difficult. Additionally, since few writing support tools exist, the error rate is high in existing text. This is perhaps also due to the fact that the minority language communities are less exposed to written text in their own language, and because there are fewer arenas where the written language is used. Until there are more resources available, AI or similar technologies are not useful for producing basic language technology tools such as proofing tools and other writing aids.

Even though there have been attempts at machine learning for low resource languages through transfer learning with little training data, there is no sign that rule-based technologies will be replaced any time soon for morphology-rich languages. Rather, machine learning can be used in limited circumstances where rule-based solutions are either non-existing or clearly inferior to the machine learned ones, such as speech synthesis, and then preferably as a complement or in combination with rule-based solutions. For example, in a Lule Sami TTS project (described below), rule-based technologies are used for text processing (disambiguation and normalisation), and the processed text (possibly converted to IPA) is sent to the synthesis engine, built using machine learning. When the machine learning does not have to deal with text normalisation and disambiguation, a much smaller voice corpus is needed, and it becomes possible for much smaller language communities to take on such a project. It still presupposes that all language independent infrastructure exists and is maintained outside such a project, otherwise the costs will be insurmountable for any minority language community.

anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://tinyurl.com/2p9ed6tp). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25,7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

[14] https://apertium.org

[15] https://giellalt.github.io

That does not mean that AI and machine learning techniques won't have a place in the future development of LT for these languages. But as already referenced, for the time being rule-based technology is the dominant paradigm, although combinations in a hybrid setup could be a useful complement, as exemplified above in the case of TTS.

Another reason for keeping most machine learning at arms length from minority languages is that the needs of the language communities are crucially different from those of the majority language communities on some core issues. As has been mentioned earlier in this report, the literary tradition will often be weak, and text production and exposure low. For many language communities their language has been prohibited from education for a long period. All of these factors make many of the language community members unsure of how to write their own language, and they need to rely much more on writing aids than a typical majority language writer. But building writing aids using machine learning techniques or list based methods on a corpus of texts is not going to fly. It was tried for Greenlandic, and failed completely. After this fail, the Greenlanders turned to rule-based, and have never looked back.

One further point strengthening the case for rule-based technologies vs machine learning for minority languages is that of resource reuse. With limited resources, both financial, computational and human, one has to plan the LT work so as to minimise duplicate and repeated tasks. Building a multitask lexicon and morphological transducer is fairly trivial, there are established routines and best practices for that in both the Apertium and GiellaLT infrastructures: you build your lexicon once, and use it for everything, removing automatically some features or content that does not match a certain use case. It is very unclear how this would even work using machine learning methods to create first a tagger, then a speller, then a machine translation system, and so on. One would need to build training resources differently every time, from the same set of limited text collections, essentially a rinse and repeat process that would not build sustainable LT resources for the language community.

The LT overview presented above in Chapter 3 defines the major areas for LT work as outlined in the other ELE reports. As outlined in this section, LT for (the Nordic) minority languages have a very different starting point, and this report thus adds the following application areas to the tables and overviews on the specific languages:

- **Language identification** is the process of identifying the language of the text or speech given to the system. In the following tables, this will be restricted to whether operating systems have any knowledge of the language in question at all, as in central registers for locales or languages available to applications and system services. In many cases, a lack of such knowledge on a system level prohibits proper integration of most LT tools, or even makes them impossible to use in some cases. In the resource tables in this report, the availability of a keyboard is taken as a proxy for a language being known to the system when no other information is available. None of the languages covered in this report are supported on ChromeOS, thus the tables will state that the support is *Partial* for a language also when all other operating systems support that language.
- **Text Input** is the means by which text is entered into a computing device. Text input can be given both spoken and typed, but in this report it will be restricted to typed input only, ie to keyboards. Most of the Nordic minority languages have no built-in keyboard support in any of the existing operating systems.
- **Proofing tools** are LT applications to help users write their language according to established norms and guidelines. Such tools, especially spell checkers, are both taken for granted by majority language users and often seen as unnecessary or distracting, especially by younger majority speaker generations. For minority language speakers on the other hand, the situation is the opposite: proofing tools are rarely available, but very much desired by all writers of the language.

It bears emphasising that in the tables of tools and resources throughout this report, **Models and grammars** does *not* equal a machine learned model. On the contrary, except for the case of Yiddish (for which we have no accurate information), all models and grammars are rule-based models, built using an explicit lexicon, rules for inflections and sound change processes, and disambiguation and parsing rules for the syntax. The result is still very much a concise and complete model and grammar of the language, just not in the sense of the machine learning paradigm. As argued earlier in this chapter, using rule-based methods and technologies is presently the only viable way of developing LT resources and tools for the languages covered in this report as well as for many, if not most other minority languages. More than twenty years of experience and development of working tools for language communities has proved our point.

4.1 Access to LT for Nordic minority languages

Beyond the situation for language technology specifically for the Nordic minority languages described in the previous section, the Nordic minority languages – and indeed almost any minority language – face another battle unknown to the majority languages: the battle for access to the language technology services that exist. The situation for the Sámi languages is described in (Moshagen, in press), a generalised summary of the main points are given below.

4.1.1 Keyboards

On many platforms, installing software keyboards is relatively straightforward, but many of the platforms have quirks, issues and downright prohibitions in certain cases. The following is a list of concrete issues, all existing in the latest OS versions as of April 1, 2022:

- **ChromeOS:** All keyboards not preinstalled as part of the OS must be listed as a variant of one of the preinstalled keyboards, that is, the minority language becomes a “variant” of the majority language. That does not give a signal of language equality, and seems unnecessary.
- **Windows:** installing new keyboard layouts is straightforward, but to be able to use the keyboard with your language tools like spellers, the language must be known to Windows. For most minority languages of the world this is not the case, and registering a new language is an error-prone, undocumented but in principle supported exercise. The result is large extra costs for minority language communities, costs that do not exist for the majority languages.
- **iOS/iPadOS and Android:** it is not possible to map an onscreen keyboard to a physical keyboard, so minority language pupils in schools using iPads or Android tablets have to switch back and forth between their physical (majority language) keyboard and the on-screen software keyboard for their native language. This is both destructive for the writing process as well as causing tension. Why should access to the physical keyboard be restricted to majority language speakers?

4.1.2 Proofing tools

The most basic proofing tool of all, the spell checker, has been around for more than forty years. It is also one of the first applications of language technology. Still it is incredibly hard to get a tool like that to work for minority language users in their environments. Some examples are:

- **Google Docs:** there is no way to install spellers such that they appear as any other speller that Google provides, with a red squiggly underline and a context menu with suggestions. There is also no way of specifying the language of the text, to enforce a certain set of tools for that language. instead developers are given an API for extensions, where none of the core functionalities exist for providing a speller in the way users expect them to behave.
- **Microsoft Office, Web:** exactly the same as for Google Docs, with the single improvement that it is possible to specify a language for the selected text. The only problem is: most minority languages are not on the list provided by Microsoft, and there is no way to add to that list.
- **Microsoft Office, Windows:** it is in principle possible to install proofing tools for any language, and have it recognised and used by the system (see the section about keyboards above). But for most minority languages it will not be identifiable by its language name, rather through a cryptic label as seen in Figure 1. And getting there is an enormous amount of work, a work that majority language speakers do not have to do.
- **Microsoft Office, Mac:** Minority language users use macs too, and they often use Microsoft Office. However, Microsoft has decided that a number of the languages recognised and listed on Windows, are *not* going to be accessible in the macOS version of Office. This was the case with the top five Sámi languages (see Section 4.8) in 2005, and it is still the case for these languages in 2022. There is no supported way of adding your own language.

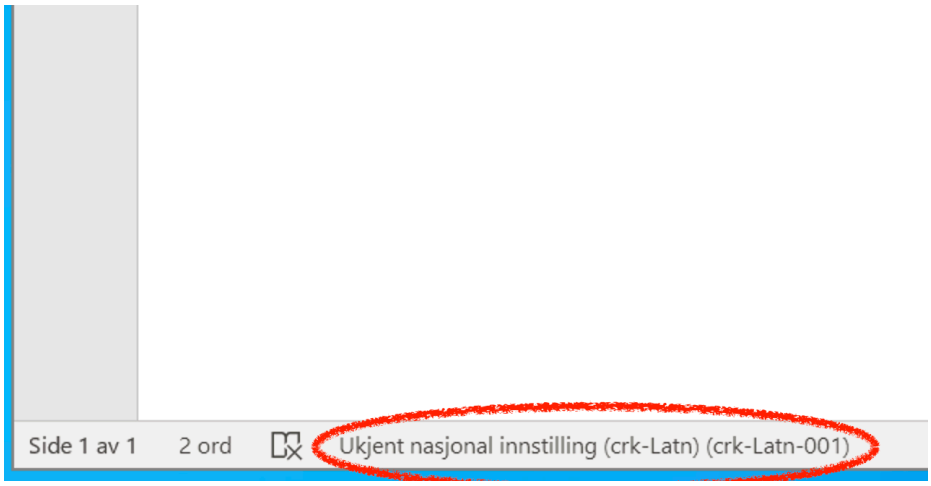

Figure 1: What Microsoft Word on Windows tells you if you try to use a minority language

Other proofing tools, like automatic hyphenation and grammar checking, face the same problems.

4.1.3 Localisation

An important part of (re)vitalising a language community is to make the language visible and accessible, in all environments. In education, an important part of this is the digital work space. Menus, window titles, button texts, etc. This everyday language teaches the users about the name of functions and parts of their digital devices, and shows them that their

language is valued also in this space. Except that for most languages on most platforms such localisation is not an option.

Many language communities will not have the resources to localise (adapt to the needs of a specific geographic area) a full office package, not to say an operating system. But if it is done (and it was done for North Sámi on Linux, which shows that it is possible and doable), the effort would be questionable, because of the difficulty of applying the localisation to the systems. Again, the language community does not have of control over their language.

4.1.4 Speech services

The future is speech we are told. Cars, computers, refrigerators, intelligent loudspeakers and phones – they will all speak to us, and some already do. But only in a majority language. The kitchen, the place that used to be the safe harbour for minority language communities, will be invaded by "intelligent" appliances that only understand a handful of languages.

A number of researchers are working on speech technology for minority languages, and at some point such services will be a reality for at least some of these languages. But unless the system providers open up their platforms to third party language service providers, these academic works will stay academic, instead of supporting the language communities.

4.2 The Faroese language

Authors: Sjur Nørstebø Moshagen, Peter Juel Henrichsen

Faroese is a separate branch of the North Germanic languages, descendant from old Norse. It has undergone a number of sound changes over the centuries, and the morphology is a bit simplified compared to Icelandic. Modern Faroese was codified in 1846.

The faroese islands were part of the Norwegian kingdom, later Dano-Norwegian kingdom, until Norway became independent from Denmark in 1814, at which point the Faroese islands continued to be part of Denmark. They still are, but with a lot of self governance. Local administration, education from daycare to university and all aspects of society are by all measures monolingual Faroese. Danish is taught in school, and holds a strong position as a second language for most Faroese speakers.

Faroese is the official language of the Faroese islands. It belongs to the North Germanic branch of the Indo-European language family, and is written with the Latin alphabet. Of the languages covered in this report, it belongs to the group with the least complex morphology and morphophonology.

The population of about 50,000 is highly connected to the net, internet coverage is close to 100%.[16] There are about 6,100 domains below the `.fo` top level domain.[17]

4.2.1 Availability of Language Data and Tools

All basic tools are available for Faroese: keyboards, morphological analysers and syntactic parsers, a speller, and even a first version of a grammar checker. There is also a Faroese text-to-speech system available from a private entity, unfortunately the resources for it are also under control of the same private entity.

There exists a number of corpus resources in various places, some of them registered in CLARIN. And there is an ongoing project to build a Faroese BLARK, documented in Simonsen et al. (in press).

[16] 97.6% in January 2021 according to https://datareportal.com/reports/digital-2021-faroe-islands
[17] cf https://research.domaintools.com/statistics/tld-counts/

4.2.2 Projects, Initiatives, Stakeholders

The Faroese islands is lacking a general language technology strategy, but has funded a couple of projects lately. There was funding for developing an open-source spelling checker, which at the same time built a morphological analyser. There is presently a large, on-going project to develop a Faroese BLARK a.o. targeted at speech technology.

Resources:	**Size or availability**	**Access**	**Source / Description**
Language identification	Partial	`fo/fao`	Android, iOS, Linux, macOS, Windows
Mobile keyboard definitions	Yes	Closed	OS vendors
Desktop keyboard definitions	Yes	Open	CLDR
Monolingual text corpora	10.6M words	Open	Part of SIKOR
Multi-lingual text corpora	Yes	Unknown	Legal & governmental texts
Multimodal corpora	Yes	Open/Closed	Broadcast archives, CLARIN resources, FADAC, Nordic word order database, new BLARK (see below)
Lexical resources	67k lemmas	Unknown	sprotin.fo
Models and grammars	Yes	GPLv3	github.com/giellalt/lang-fao
Tools:			
Mobile keyboards	Yes	Part of OS	Supplied by OS vendors
Desktop keyboards	Yes	Part of OS	Supplied by OS vendors
Proofing tools	Speller	Free	divvun.org, based on model above
Text analysis tools	Yes	Free	Based on model above
Speech synthesis	Yes	Closed	Acapela group
Speech recognition	Dev	Open	A BLARK for a.o. ASR purposes is under construction
Machine translation	Yes	Open	Apertium: fao-dan, fao-nor, fao-eng
Information extraction & IR	No	–	–
Language generation & sum.	No	–	–
Human computer interaction	No	–	–

Table 4: Resource overview for Faroese

4.3 The Greenlandic language

Author: Per Langgård

Greenland was colonised by Denmark/ Norway when the first missionary, Hans Egede, arrived in Greenland in 1721 hoping to find descendants of the Norse settlers who lived in Greenland almost 500 years, but disappeared or died out about 300 years before Hans Egede's arrival. Instead Hans Egede established the Danish mission for the Inuit he met where he first landed close to the present capitol of Nuuk and over the next generation or so for the inhabitants in a number of settlements to the North and South of Nuuk.

Activities were financed by a trade company, Bergenskompagniet, against a monopoly on the purchase of whale and seal blubber from the Greenlandic hunters for export to European lamps and industries.

From the very beginning of the Greenlandic mission it was evident to both the Egede family and the agents of the trade company that there was no way around Greenlandic speaking brokers in spiritual as well as in commercial affairs for daily life to function in the colonies. As a result, attempts to train a class of professionals and provide information about the absolutely unintelligible language that in no respects showed familiarity with the Nordic languages were given high priority.

The first dictionary, Paul Egede's *Dictionarium Grönlandico-Danico-Latinum, complectens primitiva cum fuis derivatis, qvibus interjectœ funt voces primariæ è Kirendo Angekkutorum* was printed in 1750 and the first translation of the Bible was produced in 1766.

Also local education was highly successful so that all cathecists in schools and churches were Greenlanders in early 1800 and illiteracy was allegedly eridcated by mid 1800-s. From 1848 Greenlanders were formally trained in the two cathecist-training colleges in Nuuk and Ilulissat for service in Greenlandic schools and churches.

The same picture prevails until after world war II when Greenland's colonial status formally was given up and Greenland turned an integrated part of Denmark concomitant with a very high priority given to Danish language and Danish culture. The years between 1950 and 1975 are often termed *the Danification period*. During this period Greenlandic was clearly stressed but it soon gained strength when the movement towards political autonomy gained power leading to the Home Rule Act of 1979 and the further empowerment in the Self Government Act of 2009.

4.3.1 Status of Greenlandic

Apart from the short period of danification, Greenlandic always was the unquestioned language of Greenland including education, mediae and public administration. This status was formalised in the Home Rule Act where Greenlandic is termed *the primary language of Greenland* and reinforced in the Self Government Act where §20 simply states that *Greenlandic is the official language of Greenland.*

It should be clear that Greenland does not have and actually never had problems with linguistic rights. Still, Greenlandic is according to the UNESCO taxonomy vulnerable as are all languages with a very limited amount of speakers. Unfortunately, no reliable count of the number of speakers of Greenlandic exist but for a rough estimate about half of Greenland's 56,562 inhabitants (by January first 2022) are Greenlandic monolinguals with limited L2 competence in Danish, about a quarter of the population have Greenlandic L1 plus a strong command of Danish L2 and a quarter has Danish L1 with no or limited command of Greenlandic L2. Accordingly Greenlandic is a strong L1 for not less than 40,000 persons.

4.3.2 Typology, orthography, status and language planning

Greenlandic is the biggest dialect of the Esk-aleut family of languages that nowadays often is termed the Inuit Languages. Greenlandic is like all other Inuit Languages polysynthetic with an extremely rich morphology and thus with a type/token ratio as low as it gets.

Greenlandic has had a very consistent, shallow standard orthography since 1851. It was replaced in 1973 by a modern (morphophonemic) orthography that is still in use and the principle of national standard irrespective of dialectal varieties is still adhered to.

There is one normative body for Greenlandic in general namely the national language committee, *Oqaasiliortut*, and another official body for place names in Greenland namely the Place Name Committee, *Nunat Aqqinik Aalajangiisartut*. Both committees are manned by members appointed by the government. Since 1999 The Language Secretariat, *Oqaasileriffik*, has provided academic service to the committees mentioned.

Before 2009 all access to the digital world went via satellite at rates that made the Internet inaccessible to major parts of the population. But in 2009 the sea cable from Europe to Canada landed in Greenland and brought with it dramatically lowered prices and much more stable connections including public access to free terminals at libraries, in schools etc. Today almost all Greenlanders all over Greenland have access to Cyberspace .. and exploit the possibilities extensively.

4.3.3 Availability of Language Data and Tools

Greenland always had a considerable amount of written text in spite of its total population count being lower than 50,000 persons. There are two main reasons for this. (i) The fact that reading materials in Greenlandic have been easily available to all Greenlanders since 1861 when the Greenlandic newspaper, *Atuagagdliutit* (literally "reading matters given away for free"), was published for the first time. From 1861 till 1952 *Atuagagdliutit* was monolingual in Greenlandic but since 1952 it has been published in both Danish and Greenlandic. Besides *Atuagagdliutit* several local newspapers have been published on and off since 1913. (ii) From the early days of christian mission, reading materials in Greenlandic for school use were produced. A small catechism came first but soon locally produced supplements were also added. In 1880 the first professionally printed mother tongue primer, *Atuainiutit*, was published.

Already before 1950 basically everything regarding colonial administration and school and church matters were translated into Greenlandic. With the abolition of colonial status and the introduction of the many needs for communication in many new contexts, the amount of information in Danish that needed to be translated into Greenlandic grew dramatically everywhere in society.

Finally it should be noted that literary production in Greenlandic has a long history and still is an active part of Greenlandic culture with thousands of original and translated works available for instance the National Library.

There is thus access to a considerable amount of Greenlandic data but compilation of text and literary resources in formats suited for language technology projects started only recently. Earlier attempts to create corpora soon proved to be too time-consuming and costly and had to be given up. But with the introduction of efficient household-utensils like orthographic converters, a speller and a reliable POS-tagger, Greenlandic corpora have reached considerable sizes.

It should, though, be noted that aligned corpora still are next to non-existing and that it is unlikely that the human and economic resources needed to start the process will be available over the coming years.

An overview of basic Greenlandic resources is found below in Table 5.

4.3.4 Projects, Initiatives, Stakeholders

There is no national programme guiding the development of Greenlandic language technology and there never was. Instead the political demand for LT-development was passed on

Resources:	Size or availability	Access	Source / Description
Language identification	Partial	`kl/kal`	Android, iOS, Linux, macOS, Windows
Mobile keyboard definitions	No	–	The Danish keyboard is used
Desktop keyboard definitions	No	–	The Danish keyboard is used
Monolingual text corpora	20.4M words	Free for reading, closed for download	Oqaasileriffik
Multi-lingual text corpora	Unknown	Unknown	Legal, governmental & news texts
Multimodal corpora	Yes	Unknown	Old dictionary project recordings
Lexical resources	60k lemmas	Unknown	oqaasileriffik.gl
Models and grammars	Yes	GPLv3	github.com/giellalt/lang-kal
Tools:			
Mobile keyboards	No	–	The Danish keyboard is used
Desktop keyboards	No	–	The Danish keyboard is used
Proofing tools	Speller	Free	divvun.org & oqaasileriffik.gl, based on model above
Text analysis tools	Yes	Free	Based on model above
Speech synthesis	Yes	Free	oqaasileriffik.gl
Speech recognition	No	–	–
Machine translation	No	–	–
Information extraction & IR	No	–	–
Language generation & sum.	No	–	–
Human computer interaction	No	–	–

Table 5: Resource overview for Greenlandic

to *Oqaasileriffik/The Language Secretariat* once it was established in 1999 basically without detailed constraints or comprehensive guidelines ... and with very limited funding.

In practice Oqaasileriffik had no alternative but to start with the beginning compiling lexical resources and educating Greenlandic students as future staff. After a few years *Oqaasileriffik* was ready for the next step and started to develop the fst-automaton that has been the kernel in all Greenlandic LT since. An alpha-version including a speller and a hyphenation tool was launched in 2006. After that Greenlandic LT almost ran out of funding for several years and very little happened but from 2011 private funding bought *Oqaasileriffik's* senior researcher free from other duties three years and funded two apprentice positions specifically for LT development.

Thanks to the external funding *Oqaasileriffik* succeeded in developing a comparatively stable syntactic and semantic parser that was published together with a number of utensils. This success paved the way for public funding of a few positions as LT developers. Lately, a few extra positions were added to the staff so that the LT-group at *Oqaasileriffik* at the moment amounts to 5 positions including 1 position ear-marked for terminology.

No Greenlandic LT development whatsoever takes place outside *Oqaasileriffik* and when the department of Greenlandic at Greenland's University from time to time offers introductory courses in LT, the teaching is handled by staff from *Oqaasileriffik*.

Greenlandic is a polysynthetic language and has as such an extremely complex morphology and ample very deep syntactic embedding plus extensive derivation. Add hereto the fact that Greenlandic with the increased opening towards the surrounding world via the Internet constantly imports new concepts and loanwords and/or creates scores of neologisms. For the parser to deal with this requires daily attention and the need for frequent updates is obvious. Accordingly keeping the parser up to date is still is one of our primary duties.

Apart from this, a number of other projects are currently being developed at *Oqaasileriffik*:

- **MT kal2dan and dan2kal:** In a 5-year period from 2017 to 2021 *Oqaasileriffik* in collaboration with the Danish software house *GrammarSoft* has developed rule-driven MT to and from Danish. The applications were launched as freely accessible alpha-versions on January first 2022. Upgrading and debugging these programs to a future beta-version form major parts of *Oqaasileriffik's* obligations these years.
- **Continued corpus compiling:** Refining and expanding Greenlandic corpora is ongoing. The monolingual corpus recently grew past 20 million items and a number of new genres have been included. It should be mentioned that all texts automatically are tagged by the parser. Aligned corpora still are scarce but they too will be addressed as soon as resources are at hand.
- **Terminology:** The need for consistent Greenlandic terminology is obviously urgent in the modern Greenlandic society having Greenlandic as the only official language. Terminology including authorising neologism is therefore an ear-marked part of *Oqaasileriffik's* duties these years.
- **Automatic dependency grammar:** During development of the alpha-version of kal2dan it became clear that a new and more precise dependency grammar must be developed not least in order to deal properly with inderivation. This attempt obviously puts new pressure on the parser that needs to be refined in a number of respects. It is challenging but the work is in progress and it is expected that a new and advanced dependency tagger will be launched during 2022.
- **Basic resources for English:** There are next to no basic resources to build Greenlandic-English interaction on. This creates enormous problems not only for daily communication but also for English L2 in compulsory school where all teaching materials are

based on Danish language and Danish children's needs. Simultaneous with this inadequate teaching, English – not seldomly pidgin – like English – acquired via *YouTube* and similar programs more and more becomes part of Greenlandic children's reality. In an attempt to counterbalance this situation *Oqaasileriffik* has developed programs and algorithms that are expected to be able to build a Greenlandic-English lexical database very fast once funding is secured.

4.3.5 Summary

Basically Greenlandic is vital and strong in comparison with about all other small languages in the world and the language has been standardised and attended to according to a deliberate language policy for generations.

Greenland is a modern society with a comparatively well-educated population. It is therefore questionable whether Greenlandic should be termed a lesser-resourced language along with the rest of the wold's languages of similar or smaller size.

Still, Greenlandic is like all other small languages vulnerable not least in the shadow of the tech-giants' English as the most recent developments in Greenlandic children's language seem to suggest with unstructured English seeping into their daily Greenlandic language at a pace never experienced before.

The situation has given rise to a lot of lay and political anxiety and calls for some kind of action. The only action *Oqaasileriffik* can point at is increased focus on better language education in school in L1 as well as L2 and a dramatically intensified rate of Greenlandic LT development to strengthen Greenlandic in general. This is expected to pave the way for future applications among which Greenlandic-English MT as an important part of a tool that hopefully will enable Greenlandic to surf the Internet in line with English and other big languages is considered an urgent need.

There can be no doubt that the endeavours needed to make this action a success are challenging but bearing in mind how far Greenland actually made it during a dozen or so of years the aspiration might not be impossible if properly structured which would include the revised and improved education of future staff. Greenland has enough linguistic know-how and human resources needed for the job, albeit not sufficient funding.

However, with the rate of speed the English influence seems to work it is definitely not the time to wait too long for future funding!

4.4 The Karelian language

Authors: Sjur Nørstebø Moshagen, Trond Trosterud, Flammie Pirinen

The lion's share of Karelian speakers live in the Karelian republic in Russia. Here, it had official status until 1940, when this status was granted to Finnish. Contemporary Karelian is written with the Latin alphabet, and thus cannot have official status in Russia, making the Karelian republic the only republic in Russia where the titular language is not an official language. Karelian is protected by a specific law within the Republic of Karelia, but in practice this law has little significance[18] In Russia, there are 25,605 self-reported speakers of Karelian[19]. There are diverging reports on the number of Karelian speakers in Finland. Finland's fifth report to the Council of Europe on minority language rights reports 5,000 speakers[20]

[18] Cf. p. 269 in Sarhimaa, Anneli 2022: Karelian. In: The Oxford guide to the Uralic languages. Edited by Bakró-Nagy et al., Oxford University Press.

[19] Sarhimaa, op.cit. p. 269.

[20] The Fifth Periodic Report By The Government Of Finland on The Implementation of the European Charter for Regional or Minority Languages. Finnish Ministry for Foreign Affairs, 2017.

(152 of which have reported Karelian as their mother tongue in official registries). Sarhimaa[21] reports 11,000 fluent speakers. Ethnologue reports a total number of 36,000 Karelian speakers[22]. The language is endangered both in Russia and in Finland. In Finland, the language has been recognised as a national minority language since 2009[23], and it is written using the Latin alphabet. Karelian is also among the languages Finland reports on to the European Charter for regional and minority languages.

Karelian is a Northern Baltic Finnic language, thus closely related to Finnish. Mutual understanding between the two languages is possible to a certain extent.

The position of Karelian on the Internet is more marginal than one should expect, given its size and status as a national minority language in the Nordic countries. The Finnish national broadcaster publishes both written and spoken news broadcasts in Karelian[24].

Practically all Karelian speakers in Finland have access to Internet. In Russia, it is estimated that in 2020 71.6 % of the rural population (to which most Karelian speakers belong) had access to the Internet, as compared to 82.8 % of the urban population[25]. Given that Karelia is situated in the western part of Russia, the number may be somewhat higher than the country average.

One of the problems in the computational versions of the Karelian language processing is that the languages are covered by two separate ISO-639 codes (krl and olo), a division which is not always accurately represented either in historical use (c.f. e.g. *Karjalan Kielen Sanakirja*[26]) or contemporarily. There have been attempts to clarify the situation within ISO standard in the change request 2019-037 [27], however, there was no agreement on the change request, and it was ultimately rejected.

The main freely available and open text corpus source for Karelian languages is the VepKar corpus, a collection of texts that includes both Karelian variants arranged by the ISO codes and Veps. The collection contains folklore stories as well as modern news texts, and is at the moment over 3,000 articles in size and whilst being curated continuously.

The freely available lexicons are developed within Apertium's and Giellalt's infras by interested language activists and linguists.

As far as we know, there is no large freely available parallel corpora, or spoken corpora.

4.4.1 Projects, Initiatives, Stakeholders

- Work on Karelian dictionaries for Apertium has been done by Jack Rueter and his working group with funding from Google's Summer of Code[30] and Kone foundation[31]
- The number of linguistic projects aimed at researching Karelian language and translations at uef.fi also produce computational resources; https://kianna-hanke.blogspot.com/ mainly funded by Kone foundation
- VepKar Open corpus by Krizhnazovsky et al. in Petrozavodsk includes not only curated corpora, but also lexical resources and web platforms

[21] Sarhimaa op.cit p. 269, for a discussion, see Sarhimaa 2017: *Vaietut ja vaiennetut – Karjalankieliset karjalaiset Suomessa*. SKS, pp. 111–115.
[22] https://www.ethnologue.com/language/krl
[23] Cf. the Finnish ministry of Justice, https://oikeusministerio.fi/muut-kielet
[24] https://yle.fi/uutiset/18-44136
[25] https://www.statista.com/statistics/1004225/household-internet-usage-by-area-russia/
[26] https://kaino.kotus.fi/cgi-bin/kks/kks_etusivu.cgi
[27] https://iso639-3.sil.org/request/2019-037
[28] universaldependencies.org
[29] dictorpus.krc.karelia.ru
[30] https://summerofcode.withgoogle.com/
[31] https://koneensaatio.fi/apurahat/ specifically: researchportal.helsinki.fi/...

Resources:	Size or availability	Access	Source / Description
Language identification	Yes	krl	Android, ChromeOS(?), iOS, Linux, macOS, Windows
Mobile keyboard definition	No	–	
Desktop keyboard definition	No	–	
Monolingual text corpora	Yes (3k texts)	Open (CC BY)	UD,[28] VepKar[29]
Multi-lingual text corpora	Unknown	–	
Multimodal corpora	Yes	Closed	YLE have archives of radio and television broadcasts, no agreement for use exists
Lexical resources	≈60k	Free & open	GiellaLT infra
Models and grammars	Yes	Free & open	github.com/giellalt/lang-krl
Tools:			
Mobile keyboards	No	–	
Desktop keyboards	No	–	
Proofing tools	No	Free & open	A very experimental alpha version exists in the GiellaLT infrastructure
Text analysis tools	Yes	Free & open	Via the GiellaLT infrastructure
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	Yes	Free & open	Apertium: fin-krl-olo
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 6: Resource overview for Karelian

- annotated corpora such as UD are only a startup project with several hundreds of sentences annotated by hand by language activists and linguists.

4.4.2 Summary

Karelian is a threatened language, with very limited LT support. It is crucial that more work and resources be targeted at the language if it is going to survive the next century.

4.5 The Kven language

Authors: Sjur Nørstebø Moshagen, Trond Trosterud

Kven is spoken in Northern Norway. It may be divided in 3 main dialects, the Varanger dialect, the Porsanger dialect and the dialect of the large river valleys (mainly in Western Finnmark and Northern Troms). Rasmussen (2005) finds 10,840 people "speaking or understanding Kven or Finnish" in Northern Norway, but 36 % of them also speak Sami. Many of these live in Inner Finnmark and are probably trilingual Samis. The national institution *Kvensk institutt* lists 2,000 – 8,000 speakers[32]. Ethnologue lists the number of speakers between 5,000 – 8,000.

The Kven language society faced language shift in different phases (Trosterud, 2008), especially during the years following WWII. The stronghold for Kven today is the municipality Porsanger, which declares itself trilingual. Kven is one of the national minority languages of Norway. It is used in language nests in Porsanger and northern parts of Troms, as well as a school subject. Kven is a Northern Baltic Finnic language resembling both Meänkieli and northern Finnish dialects, but without the language planning work that has gone into Finnish during the last century or so. Kven shares its orthographic principles with Meänkieli and Finnish.

4.5.1 Availability of Language Data and Tools

Language technology tools available so far are a spellchecker and an e-dictionary between Kven and Norwegian. In addition to that, there is a corpus of approximately half a million of words, mainly translations from Norwegian.

Finnish dialect archives contain both text and recorded speech of the traditional language.

4.5.2 Projects, Initiatives, Stakeholders

- Language technology for Kven was initiated at Giellatekno at UiT
- The central institution behind developmental work today is *Kvensk institutt* in Børselv in Porsanger, the national institution for Kven language and culture.
- The tools available for Kven are in use by the institutions offering Kven education: Schools, language centres and UiT.

4.5.3 Summary

The Kven language is a threatened language, with limited LT support. It is crucial that more work and resources be targeted at the language to ensure its continued use and survival.

[32] https://www.kvenskinstitutt.no/kvener/

Resources:	Size or availability	Access	Source / Description
Language identification	No	fkv	–
Mobile keyboard definition	No	–	
Desktop keyboard definition	No	–	
Monolingual text corpora	500k	Open/Closed	SIKOR
Multi-lingual text corpora	200k	Open/Closed	SIKOR
Multimodal corpora	Yes	Closed	NRK: Broadcast archives
Lexical resources	12k	Open	GiellaLT infra
Models and grammars	Yes	Open	github.com/giellalt/lang-fkv
Tools:			
Mobile keyboards	No	–	
Desktop keyboards	No	–	
Proofing tools	Yes	Free & Open	GiellaLT via divvun.no
Text analysis tools	Yes	Free & Open	GiellaLT via divvun.no
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	No	–	
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 7: Resource overview for the Kven language

4.6 Meänkieli and Sweden Finnish

Authors: Rickard Domeij

Meänkieli, formerly often called Tornedalian Finnish, is a Finno-Ugric language closely related to Finnish and Kven. Both Meänkieli and Finnish have the status of national minority languages in Sweden since 2000. Sweden does not collect sensus data on the number of speakers of languages spoken in the country, but there are rough estimations for example by (Parkvall, 2015) that Finnish has about 175,000 speakers and Meänkieli about 20,000 – 25,000 speakers.

Conditions bode well for the Finnish spoken in Sweden with regard to internet use, since it belongs to the same language community as the majority language in Finland; here we find a wide range of content, online language tools, services and internet forums. The problem is that access to content and services in Finnish in Sweden is heavily restricted. Some information in Finnish is offered on the government agencies' websites, particularly in the administrative districts, but far too little in relation to the demand. To the extent the information exists, the translation quality is often inferior, a major problem along with the lack of content.

Meänkieli, linguistically very close to Finnish, has been acknowledged as a language in its own right since 2000. Online teaching and internet forums are regarded as important ways to get young people to learn and use the language. An explicit need – as well as a statutory right – exists for using Meänkieli in communications with government agencies, particularly with the municipal administration in the administrative districts. This parallels the Finnish situation: information is available on government websites but mostly as downloadable forms, rarely as main web page content. The content mainly covers laws and similar types of basic information. The information is said to be flawed and varies significantly. Web services in Meänkieli are non-existent.

Language technology for Finnish is well developed since Finnish is the majority language in Finland, as can be seen in the language report for Finnish. Therefore, we will here focus on resources for Meänkieli and apart from that only mention some resources specifically related to Sweden Finnish.

4.6.1 Availability of Language Data and Tools

There are no special mobile and desktop keyboard definitions for Meänkieli. Instead, Finnish or Swedish keybords are used. This becomes a problem when needing language support, such as spelling correction, for writing Meänkieli on the mobile phone or the computer, since it is not possible without a keyboard definition.

There are some corpora for Meänkieli, some systematically collected, others in the wild.

Giellatekno has published monolingual corpora of meänkieli texts in different genres comprising nearly 40,000 sentences.[33]

The Finnish Language Bank has published a collection of web corpora of small Uralic Languages of which Meänkieli is one (Jauhiainen et al., 2019).

Swedish Television (SVT) and Radio (UR) have archives containing some material in Meänkieli and Sweden Finnish. Many central public agencies regularly publish written multilingual information containing Romani as one of many minority languages, some of which has been collected and made available by Språkbanken Sam, the National Language Bank department at ISOF. In the archives at ISOF there is also some material in Meänkieli and Sweden Finnish, both written and oral. At the National Library of Sweden almost everything published in Sweden is being made available for restricted use.

[33] https://gtweb.uit.no/f_korp/?mode=fit#?lang=en

Resources:	Size or availability	Access	Source / Description
Language identification	No	`fit`	–
Mobile keyboard definition	No	No	
Desktop keyboard definition	No	No	
Monolingual text corpora	450k	Yes	Giellatekno, ISOF
Multi-lingual text corpora	Yes	Open/Closed	ISOF, public agencies
Multimodal corpora	Yes	Closed	SVT/SR: broadcasts archives
Lexical resources	Yes	Closed	STR-T, Meänsuomen föreeninki
Models and grammars	No	–	Under development by Giellatekno & ISOF
Tools:			
Mobile keyboards	No	–	
Desktop keyboards	No	–	
Proofing tools	No	–	Planned by Giellteknо and ISOF
Text analysis tools	Yes	Open, Alpha	Prototype by Giellatekno
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	No	–	
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 8: Resource overview for Meänkieli

A Wikipedia for Meänkieli is in the making containing more than 1,800 articles[34].

The Lexin series[35] contains a comprehensive Finnish-Swedish dictionary. There are two larger dictionaries for Meänkieli:

Dictionary Meänkieli-Swedish. Bilingual electronic dictionary edited by Academia Tornedaliensis. Ca 33,000 uppslagsord[36].

Meänkielen sanakirja. Bilingual web dictionary for Meänkieli-Swedish ordbok under continuous development. Presently about 5,000 words and 70,000 word forms[37].

The Language Council publishes a host of glossaries in areas like education and medicine, primarily for Finnish (Sweden-Finnish terminology). For Meänkieli some thematic glossaries are available, such as a hunter's dictionary and a water and weather glossary (str-t.com)

A project has been started to make a thorough grammatical description of Mäenkieli[38].

Giellatekno has a text analyser in Meänkieli in the making and an inflecional paradigm generator[39].

4.6.2 Projects, Initiatives, Stakeholders

The National Association of Swedish Tornedalians is an interest organisation for Meänkieli. It includes Academia Tornedaliensis which has developed Dictionary Menkieli-Swedish.

Meänsuomen föreeninki – Byfinska kulturföreningen has several projects regarding Meänkieli but also Finnish, among other things the online dictionary Meänkieli sanakirja and a Wikipedia for Meänkieli.

The ISOF is responsible for language planning and disseminating knowledge about languages, dialects, folklore, names and intangible cultural heritage in Sweden. ISOF is funding minority language projects such as the electronic dictionaries Meänkieli sanakirja and Dictionary Meänkieli-Swedish which it maintains and develops.

Presently, there are plans to start a cooperation project between ISOF and Giellatekno to develop a spelling checker for Meänkieli.

4.6.3 Summary

Language support for Meänkieli is practically non-existent, apart from a few electronic dictionaries. More are needed. Spell-checking, interactive language learning, translation support and more do not exist and represent a major failing. The proximity to Finnish and Kven would allow spell checking to be done by adapting Finnish or Kven technology. The Finnish market offers several proofing and spelling checkers, for example Lingsoft's spelling and grammar checkers (lingsoft.fi). Language learning aids and interactive training materials are requested the most by the community. Above all, fundamental and broad revitalisation efforts are needed both on and off the Internet to help ensure the language's future survival. (Domeij et al., 2019) For Sweden Finnish, language resources and tools for Finnish can be used. These resources need only be complemented with resources such as bilingual word collections and corpora from communication with public agencies in Sweden.

4.7 The Romani languages

Authors: Rickard Domeij

[34] https://incubator.wikimedia.org/wiki/Wp/fit/Alkusivu
[35] https://lexin.nada.kth.se
[36] https://www.isof.se/stod-och-sprakrad/spraktjanster/ordbok-meankieli-svenska
[37] https://meankielensanakirja.com
[38] https://www.isof.se/lar-dig-mer/forskning/projekt/projektet-en-grammatisk-beskrivning-av-meankieli
[39] https://giellatekno.uit.no/cgi/index.fit.eng.html

Romani is one of Sweden's five national minority languages. There are 10,000 – 20,000 speakers of Romani according to rough estimations by (Parkvall, 2015). The actual numbers of speakers today may be considerably higher. Many different dialectal varieties are spoken, such as Kale, Lovari, Gurbeti, Tavringer romani, Kalderash, Arli, Polish romani and several others. The speakers are spread over the country but most of the speakers live in the three largest cities. In Norway, Romanes and Romani Rakkripa are recognised as two separate minority languages (see Section 2.4).

For Romani, the internet has opened up a whole new opportunity to connect with other language users in the globally dispersed language community. Strengthening cultural ties within the community is a primary reason for the need to communicate. People who have arrived in Sweden in recent years may also need to communicate with government agencies in their minority language. The problems highlighted are an acute lack of standardisation and technical prerequisites.

Classifying Romani varieties for language technology purposes is complicated. There are more Romani varieties than there are ISO codes and some ISO codes thus cover several varieties.[40] One and the same language may have several diverging normative bodies.

The Swedish Language Council has initiated a project aimed at revising the orthographies of Romani varieties in Sweden. At present (spring 2022), all varieties in Sweden have their own distinct orthographies, but one possible outcome of the Swedish project is thus that several of them may be unified.

4.7.1 Availability of Language Data and Tools

Apart from a few electronic dictionaries, Romani language support on the internet is lacking. There is no keyboard layout for Romani, in order to obtain the necessary characters (č, ř, š, ž) one may use keyboard layouts for e.g., Serbian or Bosnian. The best ergonomic keyboard layout for Romani is the Northern Sami layout, which illustrates how efforts in one minority language can support the other. People who speak Romani can use the Northern Sami layout directly (it's available on all computers) or make minimal changes (ŧ > y and đ > é), improving it even more.

As for corpora, there are a few written and/or oral collections at universities (Borin, 2000) and in folklore archives (Hyltén-Cavallius and Fernstål, 2020). Swedish Television (SVT) and Radio (UR) have archives containing some material in Romani. Many central public agencies regularly publish written multi-lingual information containing Romani as one of many minority languages, some of which has been collected and made available by Språkbanken Sam, the National Language Bank department at ISOF. At the National Library of Sweden almost everything published in Sweden is being made available for restricted use.

The Lexin series[41] contains image-based dictionaries for Romani Arli, Kalderas and Lovara, as well as a comprehensive dictionary for Romani Arli[42].

The Language Council at ISOF also publishes terminological glossaries in areas like education, health care and social service for Romani Arli, Kalderas and Lovara, as well as single ones for other varieties like Kale, Tavringer Romani and Polish Romani.[43]

Giellatekno has developed experiment language models for the varietes Arli, Kalderas and Tavringer and an alpha language model for Kale. Alpha indicates working language models with some content. Experiment indicates a working setup with no linguistic content.[44]

[40] see https://giellalt.github.io/lang-rmy/romani-languages.html for more details.
[41] https://lexin.nada.kth.se/lexin/
[42] https://www.isof.se/lar-dig-mer/publikationer/publikationer/2007-01-01-lexin-svensk-romskt-arli-lexikon
[43] https://www.isof.se/lar-dig-mer/publikationer?sv.target=12.5f8cc396177db5159bd1a78&sv.12.5f8cc396177db5159bd1a78.route=/&category=Romska
[44] https://giellalt.github.io/lang-rmy/romani-languages.html

Resources:	Size or availability	Access	Source / Description
Language identification	No	`rmf/rmg/rml/ rmn/rmu/rmy`	–
Mobile keyboard definition	No	–	
Desktop keyboard definition	No	–	
Monolingual text corpora	Yes	Unknown	
Multi-lingual text corpora	Yes	Closed	ISOF, public agencies. Few texts
Multimodal corpora	yes	Closed	SVT/SR: broadcasts archives
Lexical resources	Yes	Closed	ISOF, only for Arli
Models and grammars	Yes	Printed	E.g. Granqvist for Kale
Tools:			
Mobile keyboards	No	–	
Desktop keyboards	No	–	
Proofing tools	No	–	Plans by Giellatekno and ISOF
Text analysis tools	No	–	Plans by Giellatekno and ISOF
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	No	–	
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 9: Resource overview for the Romani languages

The situation for Romani and Romanes in Norway is similar to that in Sweden. A few dictionaries exist for Romani Rakkripa and an ABC for Romanes (Theil, 2022) has recently been published. Along with a few texts, these dictionaries are the total of available resources for these languages in Norway.

4.7.2 Projects, Initiatives, Stakeholders

The ISOF is responsible for language planning and disseminating knowledge about languages, dialects, folklore, names and intangible cultural heritage in Sweden. ISOF has been funding dictionaries for Romani in the Lexin series.

Presently, there are plans to start a cooperation project between ISOF and Giellatekno to develop a spelling checker for Romani Arli.

4.7.3 Summary

For the Romani languages, the basic foundations must be built: standardisation of the written language, font management, electronic dictionaries, spelling checkers, translation functions, and search features. Except for some initial language modeling work of Giellatekno and ISOF, there is no development of language technology for Romani to use or build on elsewhere, the work needs to be initiated where there is a political will to protect and promote the use of romani in the digital age, as in Sweden, Finland and Norway.

4.8 The Sámi languages

Authors: Sjur Nørstebø Moshagen

The Sámi people is the only indigenous people in Europe. The traditional area of the Sámi people stretches from southern Norway and Sweden all the way to the Kola peninsula in Russia, and at least nine living Sámi languages are recognised today (ordered roughly from southwest to northeast):

- South Sámi (Norway, Sweden)
- Ume Sámi (Sweden)
- Pite Sámi (Sweden)
- Lule Sámi (Norway, Sweden)
- North Sámi (Finland, Norway, Sweden)
- Inari Sámi (Finland)
- Skolt Sámi (Finland)
- Ter Sámi (Russia)
- Kildin Sámi (Russia)

The number of speakers varies from around 10 – 15 to more than 20,000 (North Sámi). All Sámi languages are endangered or heavily endangered, some on the brink of extinction. All of them are written using the Latin alphabet except the two in Russia, which are written using the Cyrillic alphabet. The Sámi languages in Russia will not be further described in this report.

South, Lule, North, Inari and Skolt Sámi are official languages in their country/countries in the area where they are traditionally spoken, and are used in education and administration. These will be designed *the top five Sámi languages* throughout the rest of this chapter. Pite and Ume Sámi have only recently been standardised.

The Sámi languages belong to the Uralic language family, together with Finnish, Estonian, Hungarian and a large number of other languages. They are characterised by an extensive suffixational morphology, and complex – indeed very complex – morphophonology.

The top five Sámi languages have seen an increased use on the Internet over the last ten years, mainly due to access to keyboards and proofing tools. It is hard to estimate the number of sites using these languages, as none of them are identified by leading search engines. But some indications can be had by Googling for language specific words. As an example, the South Sámi conjuction *jïh* (= *"and"*) returns 26,700 hits for Norway, and 17,500 hits in Sweden, using Google. That is, about 34,000 hits together. Hits does not equal sites, and it is a frequent word, but at least it proves that South Sámi is present on the net beyond just a few stray sites.

The Sámi population, as the rest of the population in the Nordic countries, is highly digital, and has high expectations based on what is available in the majority languages.

Language technology for the Sámi languages have been developed by researchers since the 1990's. Serious development with the aim of delivering end user tools started in 2004, and the first set of proofing tools was delivered in 2007: spelling checkers and automatic hyphenation for North and Lule Sámi. The work was done by two groups: a research group at UiT The Arctic University of Norway, and a development group named Divvun at the Norwegian Sámi Parliament. The tools and models were developed using Finite State Transducers (Beesley and Karttunen), and later development has also included machine translation technology from Apertium and VislCG3 (Constraint Grammar) from the University of Southern Denmark.

All technologies used are rule based – it is the only working way for languages with complex morphology and morphophonology and few pre-existing digital resources. But as can be seen in the following tables and sections, a broad set of tools have been developed and are in daily use by the language communities.

4.8.1 Availability of Language Data and Tools

The tools and resources for the top five Sámi languages are listed in Tables 10 through 14, while the status for Pite and Ume Sámi is listed in Tables 15 and 16.

4.8.2 Projects, Initiatives, Stakeholders

There is presently one main initiative concerning Sámi language technology: the GiellaLT infrastructure at UiT. The main stakeholders are the Divvun and Giellatekno groups at that university. These groups are also the main developers of that infrastructure.

The work is followed by several other stakeholders, like the Norwegian Sámi Parliament, the Sámi University, the Ministry of Local Government and Regional Development, as well as several Sámi municipalities.

The infrastructure and its support tools allow for rapid prototyping and guided development of language technology for about any language, with a focus on indigenous and minority languages with complex morphology or morphophonology (or both). Almost all technologies and development patterns use a rule-based approach, allowing even languages with no digital resources to get the tools and support it deserves. Tools like spell checkers and keyboards are easily installed and automatically updated.

Resources:	Size or availability	Access	Source / Description
Language identification	Partial	`se/sme`	Android, iOS, Linux, macOS, Windows
Mobile keyboard definition	Yes	Open	GiellaLT infra
Desktop keyboard definition	Yes	Open	GiellaLT infra, CLDR
Monolingual text corpora	39M	Open / Closed	gtweb.uit.no/korp
Multi-lingual text corpora	3.5M	Open / Closed	gtweb.uit.no/korp
Multimodal corpora	Yes	Closed	NRK/SVT/SR/YLE
Lexical resources	132k	Open	GiellaLT infra
Models and grammars	Yes	Open	github.com/giellalt/lang-sme
Tools:			
Mobile keyboards	Yes	App Store	GiellaLT infra
Desktop keyboards	Yes	Part of OS	OS vendors, GiellaLT
Proofing tools	Yes	Free & Open	GiellaLT via divvun.no
Text analysis tools	Yes	Free & Open	GiellaLT infra
Speech synthesis	Yes	Closed	Acapela (discontinued)
Speech recognition	No	–	
Machine translation	Yes	Free & Open	GiellaLT + Apertium
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 10: Resource overview for the **North Sámi** language

Resources:	Size or availability	Access	Source / Description
Language identification	Partial	`smj`	Windows
Mobile keyboard definition	Yes	Open	GiellaLT infra
Desktop keyboard definition	Yes	Open	GiellaLT infra
Monolingual text corpora	1.8M	Open / Closed	gtweb.uit.no/korp
Multi-lingual text corpora	231k	Open / Closed	gtweb.uit.no/korp
Multimodal corpora	Yes	Closed	NRK/SVT/SR: broadcast archives
Lexical resources	76k	Open	GiellaLT infra
Models and grammars	Yes	Open	github.com/giellalt/lang-smj
Tools:			
Mobile keyboards	Yes	App Store	GiellaLT infra
Desktop keyboards	Yes	Free & Open	GiellaLT via divvun.no
Proofing tools	Yes	Free & Open	GiellaLT via divvun.no
Text analysis tools	Yes	Free & Open	GiellaLT infra
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	Beta	Free & Open	GiellaLT + Apertium
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 11: Resource overview for the **Lule Sámi** language

Resources:	Size or availability	Access	Source / Description
Language identification	Partial	`sma`	Windows
Mobile keyboard definition	Yes	Open	GiellaLT infra
Desktop keyboard definition	Yes	Open	GiellaLT infra
Monolingual text corpora	2M	Open / Closed	gtweb.uit.no/korp
Multi-lingual text corpora	198k	Open / Closed	gtweb.uit.no/korp
Multimodal corpora	Yes	Closed	NRK/SVT/SR: broadcast archives
Lexical resources	58k	Open	GiellaLT infra
Models and grammars	Yes	Open	github.com/giellalt/lang-sma
Tools:			
Mobile keyboards	Yes	App Store	GiellaLT infra
Desktop keyboards	Yes	Free & Open	GiellaLT via divvun.no
Proofing tools	Yes	Free & Open	GiellaLT via divvun.no
Text analysis tools	Yes	Free & Open	GiellaLT infra
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	Alpha	Free & Open	GiellaLT + Apertium
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 12: Resource overview for the **South Sámi** language

Resources:	Size or availability	Access	Source / Description
Language identification	Partial	smn	Windows
Mobile keyboard definition	Yes	Open	GiellaLT infra
Desktop keyboard definition	Yes	Free & Open	GiellaLT via divvun.no
Monolingual text corpora	3.1M	Open / Closed	gtweb.uit.no/korp
Multi-lingual text corpora	85k	Open / Closed	gtweb.uit.no/korp
Multimodal corpora	Yes	Closed	YLE: broadcast archives
Lexical resources	46k	Open	GiellaLT infra
Models and grammars	Yes	Open	github.com/giellalt/lang-smn
Tools:			
Mobile keyboards	Yes	App Store	GiellaLT infra
Desktop keyboards	Yes	Free & Open	GiellaLT via divvun.no
Proofing tools	Yes	Free & Open	GiellaLT via divvun.no
Text analysis tools	Yes	Free & Open	GiellaLT infra
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	Alpha	Free & Open	GiellaLT + Apertium
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 13: Resource overview for the **Inari Sámi** language

Resources:	Size or availability	Access	Source / Description
Language identification	Partial	`sms`	Windows
Mobile keyboard definition	Yes	Open	GiellaLT infra
Desktop keyboard definition	Yes	Free & Open	GiellaLT via divvun.no
Monolingual text corpora	250k	Open / Closed	gtweb.uit.no/korp
Multi-lingual text corpora	No	–	
Multimodal corpora	Yes	Closed	YLE: broadcast archives
Lexical resources	46k	Open	GiellaLT infra
Models and grammars	Yes	Open	github.com/giellalt/lang-sms
Tools:			
Mobile keyboards	Yes	App Store	GiellaLT infra
Desktop keyboards	Yes	Free & Open	GiellaLT via divvun.no
Proofing tools	Yes	Free & Open	GiellaLT via divvun.no
Text analysis tools	Yes	Free & Open	GiellaLT infra
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	No	–	
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 14: Resource overview for the **Skolt Sámi** language

Resources:	Size or availability	Access	Source / Description
Language identification	No	`sje`	–
Mobile keyboard definition	No	–	
Desktop keyboard definition	No	–	
Monolingual text corpora	No	–	
Multi-lingual text corpora	No	–	
Multimodal corpora	Yes	Closed	SVT/SR: broadcast archives
Lexical resources	7k	Open	GiellaLT infra
Models and grammars	Alpha	Open	github.com/giellalt/lang-sje
Tools:			
Mobile keyboards	No	–	
Desktop keyboards	No	–	
Proofing tools	No	–	
Text analysis tools	Alpha	Free & Open	GiellaLT infra
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	No	–	
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 15: Resource overview for the **Pite Sámi** language

Resources:	Size or availability	Access	Source / Description
Language identification	No	sju	–
Mobile keyboard definition	Alpha	Free & Open	github.com/giellalt/keyboard-sju
Desktop keyboard definition	Alpha	Free & Open	github.com/giellalt/keyboard-sju
Monolingual text corpora	No	–	
Multi-lingual text corpora	No	–	
Multimodal corpora	Unknown	Closed	SVT/SR: broadcast archives?
Lexical resources	No	–	
Models and grammars	No	–	
Tools:			
Mobile keyboards	No	–	
Desktop keyboards	No	–	
Proofing tools	No	–	
Text analysis tools	No	–	
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	No	–	
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 16: Resource overview for the **Ume Sámi** language

4.8.3 Summary

The resource situation for the Sámi languages varies quite a lot, from acceptable for North Sámi to non-existing for Ume and Pite Sámi.

4.9 The Yiddish language

Authors: Rickard Domeij

Yiddish has been one of Sweden's official minority languages since 2000. Of the 15,000–20,000 Jews living in Sweden, mainly in the big cities, about 3,000 are speaking Yiddish. For many Yiddish speakers in Sweden, it is not their first language, but strongly associated with family life and Jewish culture. It is used among relatives and friends and in connection with association and cultural activities. (Minoritet.se) For Yiddish, like Romani, the internet has opened up a whole new opportunity for speakers to connect with other Yiddish speaking people in the globally dispersed language community. Strengthening cultural ties within the community is a primary reason for the need to communicate.

4.9.1 Availability of Language Data and Tools

Compared to Romani, Yiddish has come further, with keyboards for Hebrew, corpora and basic language technology like spelling checking provided by the international community.[45] There also seems to be enough data for Yiddish for Google to develop a translation engine. In this report, the focus is on language resources and tools for Yiddish in Sweden.

The Language Council at ISOF has published a bilingual dictionary for Swedish – Yiddish containing about 8,200 entries. The dictionary will be published on the web, free to use.

The Language Council has also published a glossary in Yiddish containing terms related to the corona pandemic.

Swedish Television (SVT) and Radio (UR) have archives containing some material in Yiddish. Many central public agencies regularly publish written multilingual information containing Yiddish as one of many minority languages, some of which has been collected and made available by Språkbanken Sam, the National Language Bank department at ISOF. In the archives at ISOF some material in Yiddish may also exist. At the National Library of Sweden almost everything published in Sweden is being made available for restricted use.

4.9.2 Projects, Initiatives, Stakeholders

The ISOF is responsible for language planning and disseminating knowledge about languages, dialects, folklore, names and intangible cultural heritage in Sweden. The ISOF has been funding the Swedish-Yiddish dictionary.

4.9.3 Summary

In Sweden, language resources such as bilingual word collections and corpora need to be constructed with respect to the communicative needs of using Yiddish in the Swedish society. Otherwise, the development of language technology for Yiddish is a task that is carried out in international communities where it already has started to take important steps.

[45] See for example https://yiddishinstitute.org/yiddish-resources/ and https://www.cs.uky.edu/~raphael/yiddish.html.

Resources:	Size or availability	Access	Source / Description
Language identification	Partial	`yi/yid`	Android, iOS, Linux, macOS, Windows
Mobile keyboard definition	Yes	Unknown	
Desktop keyboard definition	Yes	Unknown	
Monolingual text corpora	Yes	Unknown	
Multi-lingual text corpora	Yes	Unknown	Swedish-Yiddish by ISOF. Open license under discussion.
Multimodal corpora	Yes	Unknown	SVT and SR have archives of radio and television broadcasts. no agreement for use exists.
Lexical resources	Yes	Unknown	
Models and grammars	Yes	Unknown	
Tools:			
Mobile keyboards	Yes	Unknown	
Desktop keyboards	Yes	Unknown	
Proofing tools	Yes	Unknown	yiddish-sources.com/yiddish-spell-checker
Text analysis tools	No	–	
Speech synthesis	No	–	
Speech recognition	No	–	
Machine translation	Yes	Closed	Google Translate
Information extraction & IR	No	–	
Language generation & sum.	No	–	
Human computer interaction	No	–	

Table 17: Resource overview for Yiddish

5 Conclusions

As outlined within this report there are many factors which create barriers to language use in digital arenas, thus effectively blocking all attempts at language equality. What is happening is that the major players in the digital sphere are controlling access to language; they are becoming language gatekeepers. Both they as companies, and we as a society, want to develop the digital world and the services it can provide, but as described, the situation for minority languages is leading to *less* language use, not more, and adds to the pressure causing language extinction.

It is imperative that ownership of languages, including control of access and use, is transferred back to the language communities, and away from the major players. This should also be in the interest of these players: no single company can alone serve all languages, and noone expects them to either. On the contrary, working solutions for most of the languages in the world will have to be developed by various third party groups: academics, open-source groups, voluntary organisations, language activists, small and medium sized business, etc. The only thing required is that the major actors should include the languages, and tools and services for them, in their platforms.

5.1 Nordic minority language resource status

To sum up, the Nordic minority languages can be grouped according to language technology maturity and resource availability as presented in Table 18.

As a general note, it should be added that even when there are resources and tools available, various platforms and computing environments prohibit the use of these tools, usually because they are not delivered by the system provider. Examples are voice assistants and various types of voice technology services on several operating systems, and proofing tools in web-based office packages. It is frustrating for both language communities and developers to know that the tools exist but are still beyond reach. This situation is not specific to the minority languages in the Nordic countries, but applies equally to all minority and indigenous languages throughout Europe and the rest of the world.

5.2 Recommended actions to improve the resource situation

The previous chapters outline the basic missing tools and resources for each language or language group. Filling the gaps is a first step towards digital language equality. The cost of LT development is the same for all languages, regardless of the number of speakers. This must be taken into consideration when funding projects for smaller languages. It should be noted, though, that LT frameworks and projects like Apertium and GiellaLT provide an infrastructure that has basic support in place, including integration with host applications and operating systems, so that the upfront costs are substantially reduced compared to building everything from scratch. This shows that the costs for supporting a minority language is not prohibitive, even though the language community might be very small.

Given the difference in support levels, the languages in this report have different immediate needs even though the end goal of digital equality should be the same for all. For the languages in group 4, a basic requirement such as defining a norm for the written language is the most immediate need. Without it, even basic spell checker programs are unachievable. Languages in group 2, on the other hand, have reached a stable level of digital readiness and are impeded by the lack of access to digital platforms.

For languages with little or no text resources, normalisation and language support tools become all the more important, since there is not enough text from which a ”majority con-

Group	Languages	Description
1	German in Denmark, Finnish in Sweden, Swedish in Finland	Have access to all the tools that are accessible to the same language as a majority language in another country, but lack resources for adaptation to country specific terminology, dialects or other linguistic parameters to make the tools really useful in their local context. The language is always recognised by operating systems, but usually not the particular country variant.
2	Faroese, Greenlandic, North and Lule Sámi	Have access to all basic tools on all platforms, as well as the resources to build and maintain them. This includes keyboards and spelling checkers. They also have access to one or more advanced tools, like grammar checkers, speech synthesis or machine translation. They still lack consistent access to all advanced tools, and do not (yet) have resources to tackle demanding tasks like automatic speech recognition and dialog systems. These languages are usually recognised by many operating systems and computing environments, but that is often not enough to make language tools available to users.
3	South, Inari, Skolt Sámi, Kven Finnish, Meänkieli	Have access to some or most basic tools on most platforms. Work may have started on more advanced tools, but nothing has been released. Solid or good lexical coverage and grammatical models, but very limited corpus resources. These languages are usually *not* recognised by operating systems or computing environments.
4	Romani languages, Karelian, Pite and Ume Sámi	Have (close to) no basic tools on any platforms. No (or close to no) resources to develop the tools from, except a printed grammar or similar. These languages are presently *not* recognised by *any* computing environment or operating system.

Table 18: Groups of Nordic minority languages according to language technology maturity and resource availability

sensus" can be derived. To create such tools, lexical resources, parsers and grammars should be prioritised since they are versatile and can be re-used to create other tools and resources.

When prioritising solutions and resources for languages in group 2 and 3, we should keep in mind that most minority language users are bi- or multilingual and understand the majority language. Users of these languages may find it more useful to be able to write their own language on a digital platform and have it translated to a majority language rather than the other way around, in which case a translation from a minority to a majority language should have the higher priority.

The need for proofing tools to support the text creation process is much more imperative for minority and indigenous languages due to the socio-linguistic situation of these languages. The speakers are much less exposed to their own written language than the majority language speakers are. There are fewer arenas that allow for the use of the written language, and there is a constant pressure from the majority language. All of this is making the writing process much more demanding. Since minority speakers do not have the tools to generate digital texts in their language, and, as they are often bilingual, they shift to the majority language, simply because the tools to type in the majority language are available.

It is crucial that the language community is encouraged to use their language also in writing, firstly to build a stronger language community as part of a vitalisation process, but secondarily also to generate the digital language data that can fuel future development of more advanced tools. Last but not least: by using the languages in writing, lacunae in terminology and the general lexicon related to the general development of the society can be detected.

As the resources for these languages must be developed using rule-based technology (cf above), to improve the situation the following need to be addressed:

- Creation of missing basic resources for all languages in the report. This is a responsibility for all the Nordic countries, and Nordic cooperation is particularly important for languages that cross borders.
- Responsible bodies for ensuring such resources must be appointed and financed.
- Routines for language documentation must be established at national level, and, in languages that cross borders, on a Nordic and even European level.
- Early involvement of the language communities themselves.
- Raise awareness in the language communities of the importance of language data and the availability of such data for language technology.
- Make language data and tools available through repositories such as ELG.

A summary of the resource status for the languages in groups 2-4 can be seen in Table 19.

5.3 Recommended actions to improve access to language technology services

Language technologies for Nordic minority languages and other small Nordic languages are currently excluded from participation on large, digital platforms. This is a serious threat to the future of small languages, since individuals as well as whole language societies are kept out of the digital sphere. While language rights are often seen as the right an individual has to learn and use his or her own language, it is also true that it is not possible without a linguistic community.

The ownership the linguistic community has to a tool seems to be a key to its success. Most of the successful LT tools and projects have been developed in cooperation with the local

Resource	Status	Recommended actions
Language identification	Missing for many / most languages	Urge OS providers to add support for all languages
Mobile keyboard definition	Missing for many languages	Develop working layouts for all languages
Desktop keyboard definition	Missing for several languages	Develop working layouts for all languages
Monolingual text corpora	Missing for many languages	Systematically collect the corpora that exist
Multi-lingual text corpora	Missing for most languages	Systematically collect the corpora that exist
Multimodal corpora	Missing or unaccessable for most languages	Systematically collect the corpora that exist
Lexical resources	Missing for most group 3 and 4 languages	Develop based on printed and electronic material
Models and grammars	Missing for most group 4 languages	Develop based on printed grammars and available resources

Table 19: Resource status and development actions for Nordic minority languages

linguistic communities, and answer to their particular needs. It should also be noted that not all minority language communities wish to share their language resources or participate in the solutions that are used by the majority society.

When it comes to access to the big platforms, we can think of two possible ways of making them available to smaller languages. One way would be for big companies with a monopoly-like market share to recognise the responsibility they have not to exclude some members or some groups of society. They could open their systems to include language technology developed by third parties for smaller languages.

The same openness can be achieved if large organisations, such as the Nordic Council of Ministers and the EU, use their economic and political muscle to put pressure on companies to open up their platforms to all minority and small languages. It could be in the form of a digital language technology act modelled on the anti-gatekeeping policies of the Digital Markets Act, and ensure that individuals or groups are not kept outside of the society at large, thereby avoiding the most serious threat to the future of minority and indigenous languages.

The following actions should be taken in order to make language technology accessible to minority languages:

- Involve minority language users and communities in all processes and projects on digitalisation and language technology.
- Institutions for higher education should offer language technology courses and research on rule based technology that suits low-resource languages.
- A legal framework to achieve digital inclusion should be created. Such a legal framework can be modelled on the Digital Market Act, and regulate access to language technology for all parts of a digital platform: from keyboards to digital assistants to localisation.

- Large organisations, such as the Nordic Council and the European Union must take action to ensure access for small languages on large platforms.

There is reason to believe that the situation described in this report for the Nordic countries is similar to that of small and minority language communities in other European regions and countries and that the proposed actions will benefit also these language communities across Europe.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

Lars Borin. A corpus of written finnish romani texts. In *LREC 2000. Second International Conference on Language Resources and Evaluation. Workshop proceedings. Developing Language Resources for Minority Languages: Reusability and Strategic Priorities*, pages 75–82. Athens: ELRA, 2000. URL http://www.ling.uu.se/lars/pblctns/lrec2-romani.pdf.

Noam Chomsky. *Syntactic structures*. The Hague: Mouton, 1957.

Rickard Domeij, Ola Karlsson, Trond Trosterud, and Sjur Moshagen. Enhancing information accessibility and digital literacy for minorities using language technology : The example of sami and other national minority languages in sweden. In *Indigenous Writing and Education :*, number 37 in Studies in Writing. Brill, 2019. ISBN 978-90-04-29850-7. doi: 10.1163/9789004298507. URL https://brill.com/abstract/book/edcoll/9789004298507/BP000008.xml.

Lena Ekberg. The national minority languages in sweden – their status in legislation and in practice. In Gerhard Stickel, editor, *National, regional and minority languages in Europe: contributions to the annual conference 2009 of EFNIL in Dublin*, Duisburg papers on research in language and culture. European Federation of National Institutions for Language, 2010. ISBN 9783631603659. URL https://books.google.se/books?id=fFYa2ooeVXgC.

Charlotte Hyltén-Cavallius and Charlotte Fernstål. *Ett lapptäcke av källor : Kunskapsproduktion om romer och resande vid arkiv och museer*. 1 edition, 2020. ISBN 978-91-88909-57-2.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Linden. Wanca in korp: Text corpora for underresourced uralic languages. In Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen, and Katja Västi, editors, *Proceedings of the Research data and humanities (RDHUM) 2019 conference*, number 17 in Studia Humaniora Ouluensia, pages 21–40, Finland, 2019. University of Oulu. ISBN 978-952-62-2320-9.

Sjur Nørstebø Moshagen. Samisk språkteknologi i 2021. *Språk i Norden / Sprog i Norden*, in press.

Mikael Parkvall. *Sveriges språk i siffror : Vilka språk talas och av hur många?* Number 20 in Språkrådets skrifter. Institutet för språk och folkminnen, Språkrådet / Morfem, 2015. ISBN 978-91-980922-7-1.

Annika Simonsen, Iben Nyholm Debess, Sandra Saxov Lamhauge, and Peter Juel Henrichsen. Creating a basic language resource kit for faroese. In *LREC 2022. 13th International Conference on Language Resources and Evaluation*, in press.

Rolf Theil. *ABC Romani Shib*. Oslo: Almater Forlag, 2022.

Trond Trosterud. Language assimilation during the modernisation process: Experiences from norway and north-west russia. *Acta Borealia*, 25:93–112, 12 2008. doi: 10.1080/08003830802496653.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur N. Moshagen, Flammie A. Pirinen, Trond Trosterud, and Børre Gaup. Unmasking the myth of effortless big data — making an open source multilingual infrastructure and building language resources from scratch. In *LREC 2022. 13th International Conference on Language Resources and Evaluation*, in press.