Language technology for the minority Finnic languages

Flammie A Pirinen
Divvun — UiT Norgga árktalaš universitehta
Tromsø, Norway
first.last@uit.no

Trond Trosterud
Giellatekno — UiT Norgga árktalaš universitehta
Tromsø, Norway
first.last@uit.no

Jack Rueter
Helsingin Yliopisto
Helsinki, Finland
Affiliation / Address line 3
first.last@helsinki.fi

Abstract

This article gives an overview of the state of the art in language technology tools for Balto-Finnic minority languages, i.e., Balto-Finnic languages other than Estonian and Finnish. For simplicity, we will use the term *Finnic* in this article when referring to all members of this language branch *except* the Estonian and Finnish literary languages. All in all, there are nine standardised languages represented in existing language technology infrastructures with keyboards, grammatical language models, proofing tools, annotated corpora and (for one of the langauges) extensive ICALL programs. This article presents these tools and resources, discusses the relation between language models and proofing tool quality, as well as the (potential) impact of these tools on the respective language communities. The article rounds off with a discussion on prospects for future development.

1 Introduction

In contemporary Uralic language technology, the majority languages of the countries such as Finnish, Estonian and Hungarian are well researched and documented, whereas minority languages lack some of the resources. For example, in terms of mapping the status of language technology of European languages, there exist two series of white-papers from the central European research infrastructures, one by Springer (Koskenniemi et al., 2012; Liin et al., 2012; Simon et al., 2012) and another by ELE (Muischnek, 2022; Linden and Dyster, 2022; Jelencsik-Mátyus et al., 2022). For minority languages in the Nordic countries, there are also two such reports (Moshagen et al., 2022 and Steingrímsson et al., 2024). Two Finnic languages were covered by the two last reports (Kven and Meänkieli), but, to our knowledge, no such overviews exist for the Finnic minority languages as a whole. One of our aims is to fill that gap.

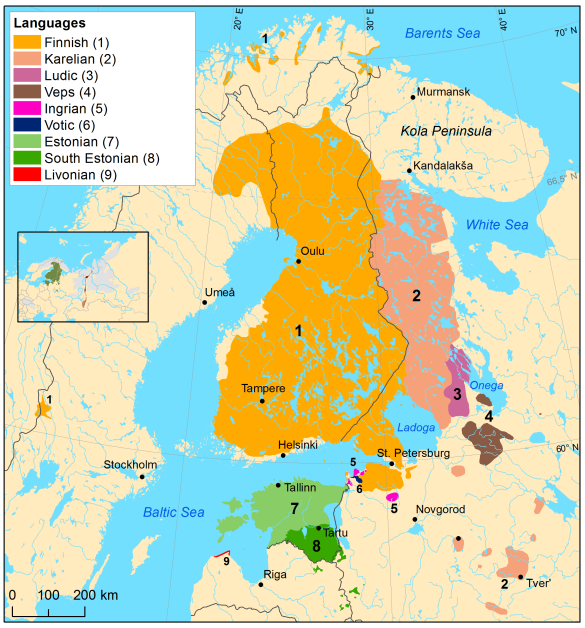

Figure 1: The Finnic Languages (Rantanen et al., 2022)

Much of the Finnic language technology has been done within the GiellaLT infrastructure[1], where the present authors all have been active, but both the Apertium[2] and Neurotõlge[3] machine translation systems have been applied to Finnic languages as well. In this paper, we give an overview of current and ongoing work in the field of Finnic language technology.

2 Background

This section gives a brief presentation of the languages and thereafter the technological foundation for the language technology used with them.

2.1 Languages

The Finnic language area is shown on the map in Figure 1. The map is ordered according to linguis-

[1] https://giellalt.github.io, see also Pirinen et al. (2023); Moshagen et al. (2023)
[2] https://apertium.org, Khanna et al. (2021)
[3] https://neurotolge.ee, Yankovskaya et al. (2023)

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 39–48
December 10-12, 2025 ©2025 Association for Computational Linguistics

tic criteria and does not quite correspond to the written Finnic languages. Subsumed under (1) in the map are also Meänkieli and Kven (marked as "Finnish" on the Swedish and Norwegian side of the border in Northern Fennoscandinavia, respectively. Within the South Estonian area (8) there is only one written standard, whereas the Karelian area (2) covers North Karelian Proper (krl) and Livvi (see 3.1 below for a discussion). Outside the present presentation fall the majority languages (Estonian and Finnish). This leaves us with a linguistic map quite close to the 11 Finnish language codes, shown in Table 1.

Language	**ISO**	*Glottolog*	*Finnish*
Meänkieli	fit	torn1244	meänkieli
Kven	fkv	kven1236	kveeni
Karelian	krl	kare1335	karjala
Livvi	olo	livv1243	livvi
Ludic	lud	ludi1246	lyydi
Veps	vep	veps1250	vepsä
Ingrian	izh	ingr1248	inkeroinen
Votic	vot	voti1245	vatja
Võro	vro	sout2679	võro
Livonian	liv	livv1244	liivi

Table 1: Names and codes for the Finnic minority languages

All the Finnic minority languages are written in the Latin script, using orthographic principles much in line with the ones used for Finnish. Typologically, the language branch is quite homogenous, the languages are mainly agglutinative with rich case systems for the nominals and tense-mode systems for the verbs. The size of the case systems ranges from 8 (Livonian, Viitso and Ernštreits (2012), Laakso (2022)) to 18 (Veps, Grünthal (2022)), and most of the languages use possessive suffixes for all nouns. Most of the languages have consonant gradation and vowel harmony, whereas Livonian and Veps have neither.

All the Finnic languages are presented in two recent handbooks on Uralic languages, Bakró-Nagy et al. (2022)[4] and (Abondolo and Valijärvi, 2023)[5]. Kven is presented in Söderholm (2017) and Meänkieli in Pohjanen (2022).

[4]See especially the chapters on Ingrian (Markus and Rozhanskiy, 2022a), Karelian (Sarhimaa, 2022), Livonian (Laakso, 2022), Seto (Pajusalu, 2022), Veps (Grünthal, 2022), Votic (Markus and Rozhanskiy, 2022b).

[5]Relevant chapters are Grünthal (2023) on Finnic and Plado et al. (2023) on Võro.

2.2 Technologies

The main technologies used for language modelling in the GiellaLT infrastructure are *Finite State Morphology* (Beesley and Karttunen, 2003, FSM), *Constraint Grammar* (Karlsson, 1990, CG), and *Two-Level Morphology* (Koskenniemi, 1983b, TWOL). This means that morphology and syntax is implemented based on (hand-written) dictionaries of lemma-stem pairs and on rules governing morphology, morphophonology and syntax. These dictionaries and rules are then compiled into finite-state automata for efficient processing. Contextually determined disambiguation and higher level syntax rules are written in constraint grammar and processed programmatically. The grammatical models are compiled with Helsinki Finite-State Technology (HFST) (Lindén et al., 2009) and the constraint grammars with VISL CG 3 (Bick and Didriksen, 2015), both free and open source products. HFST is based on weighted finite-state automata and can contain statistical information about words and word-forms. Throughout this article, we use the term *language model* broadly for any system that can analyse or validate word-forms and may or may not have statistical information. The *grammatical model* is used to point to the rule-based model consisting of the traditional FSM, CG and TWOL.

The source code for the grammatical models is stored on Github as open source [6]. The applications that can be developed with the language models include spell-checking and correction, grammatical error correction, computer-assisted language learning and speech technology applications.

The GiellaLT infrastructure also holds corpora. They are used both for development and testing of the language models and are presented as annotated corpora, accessible via dictionaries or for corpus linguistics[7]. The tools are also used in collaborative infrastructures, such as the Language Bank of Finland Korp server Rueter (2024). For minority Uralic languages, the availability of texts in general is limited, and certain genres might be totally absent. The variance in "quality" in relation to standards is more extensive than what is available for majority languages that have long established writing systems.

The universal dependencies project (Zeman et al., 2025) contains several Finnic language datasets:

[6]`https://github.com/giellalt/`, see `https://github.com/divvungiellatekno` for a full overview

[7]`https://gtweb.uit.no/korp`

Karelian and Livvi have been built based on GiellaLT analysers and manual annotation (Pirinen, 2019).

The grammatical models generate paradigms and the corpora present usage expamples for digital dictionaries for most of the Finnic languages[8]. The dictionaries are very useful for language communities and language learners[9].

The underlying technology for rule-based machine translation of the minority Uralic languages is traditionally based on the Apertium tools (Khanna et al., 2021). What this means in practice is that we can make use of the above-mentioned Finite State Morphology for language modelling, and add to that bilingual (translation) dictionaries, and grammatical rules concerning about structural re-ordering of words and phrases to implement the machine translation.

In recent years we have also started to develop speech technologies, while this is not yet production quality for the languages mentioned in this article, we are hopeful that the successes shown, for example, for Saami languages by Hiovain-Asikainen and De la Rosa (2023) will be transferable to Finnic minority languages as well.

In recent years within natural language processing, the use of large language models and neural networks has become more popular and widely replaced rule-based technologies. While this works for larger languages with plenty of available language data covering all textual genres and containing largely grammatically correct and correctly spelled language, this is more challenging and produces still less optimal results for minority Uralic languages. For this reason, the first step for us is usually to get rule-based tools that promote language revitalisation and writing normative language, that is, creating more language data that these large language models need as a prerequisite.

There exists some work done in the Uralic neural network model space, especially within machine translation, Yankovskaya et al. (2023) have released systems for minority Uralic languages, see Table 5.3 below for a discussion.

[8]The dictionaries are available at `https://sanat.oahpa.no` (Kven, Livvi, Meänkieli, Veps) and `https://sonad.oahpa.no` (Ingrian, Liv, Võro and Votic), respectively.

[9]See e.g. Räisänen et al. (2024) for an analysis of the role of the Kven dictionary in revitalisation.

3 Grammar models and standardisation

When making grammatical language models, one always has to make choices: Some grammatical forms are included in the model, others are not. When the models are turned into proofing tools and similar programs, the normative aspects become central linguistic questions. On the other hand, when models are used in search engines or speech technology, a completely different set of questions over inclusion of words and word-forms arises.

3.1 How many standard languages?

The international standard ISO 639-3, *Codes for the representation of names of languages*, lists 9 Finnic languages (c.f Table 1), in addition to standard Finnish and Estonian. This has profound consequences in a language technology setting, as the ISO codes are used by the operating systems as identification of languages for proofing tools, for example, in text editors, localisation of user interfaces, speech technology, etc. A language without an ISO 693-9 code is thus invisible to the computer. Any language community in search of literacy thus needs an ISO language code.

According to (Laakso and Skribnik, 2022, 93f), there are literary languages for Veps, Livonian, Meänkieli and Kven as well as a common literary language for Võro and Seto. Laakso and Skribnik do not mention written languages for Ingrian, Ludic or Votic but for Karelian they report that there exist "at least three different written forms for the diverse dialects of Karelian".

As seen in Table 1, there is no separate tag for Seto, and **vro** is assigned to *Võro*. Glottolog the ISO standard, aligns the ISO code **vro** with Glottolog code **sout2679** for South Estonian, this node then contains 13 subnodes, two of them are **seto1244** for Seto (itself with 3 subnodes) and **voro1243** for Võro. If Laakso and Skribnik are correct, the ISO code **vro** may be used for identifying the Seto-Võro written language.

The most problematic part is Karelian. ISO offers the code quadruplet **krl, olo, lud, vep**, for Karelian, Livvi, Ludic and Veps, respectively. The traditional distribution is shown in Figure 2.

According to the corpus data presented in Boyko et al. (2022), Chapter 2.1, the ISO codes are actually quite appropriate for the situation at hand. They present 4 corpora, for the languages "Veps, Livvi, Ludian and Karelian proper", i.e., an exact match with the existing language codes. As long as no

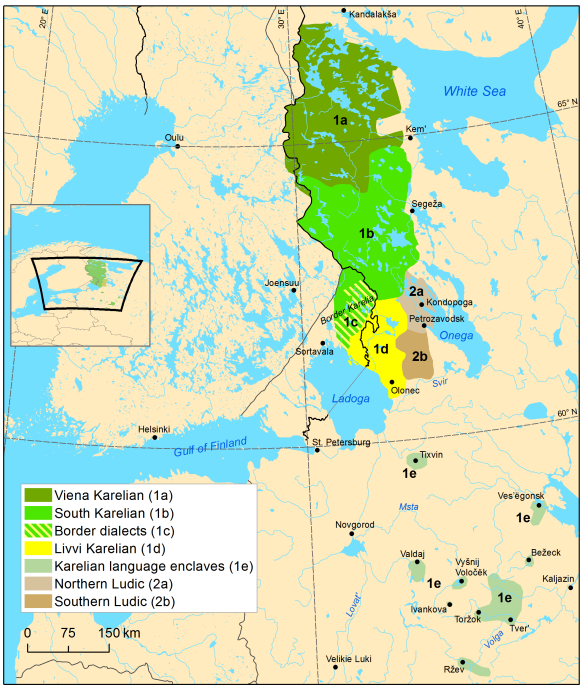

Figure 2: Karelian and Ludic around 1900 (Rantanen et al., 2022)

standard is claimed for South Karelian ((1b) in Figure 2, the ISO code inventory provides a good tool for making proofing tools for the Finnic languages of Russia.

3.2 Meänkieli and Kven: Many norms in one

Kven and Meänkieli pose a different type of challenge. Here, the ISO codes, are unambiguous, the problem is rather that some speakers would like to distinguish between three standardised varieties for both Kven (c.f. Söderholm (2017)). and perhaps also for Meänkieli (incidentally, Glottolog offers 3 codes for Meänkieli dialects but none for Kven). Obtaining different ISO language codes for these would probably be problematic, but so is the situation of missing support for the (emerging) varieties. So far, the problem has been solved in different ways for these two langauges. For Meänkieli, the analyser includes all variant forms on an equal footing, thus allowing for (even inconsistent) variation in writing. For Kven, there is one grammatical model for all three dialects. We here show a snippet of code for nouns with short vowel stems for two of the Kven dialects, Porsanki and Varanki. Both share the same genitive suffix, but the partitive suffix is set to the archiphoneme `^A` for the Varanki dialect and `^V` for the Porsanki dialect. Then TWOL rules[10] will spell out the actual forms of `^A` (as *a* when the stem contains *aou* or *ä* elsewhere) and `^V` (as a copy of the preceding vowel). During compilation, we build one transducer for each dialect, by removing the strings containing the other dialect tags for each dialect, and thereafter the dialect tag of the dialect desired (but not the string containing it). The genitive case is common to both dialects (as is most of the morphology), it receives no dialect tags and is kept throughout compilation.

```
LEXICON n_11 ! päivä, syksy, kuva, ...
...
+N+Sg+Gen:^WG%>n # ;
+N+Sg+Par+Dial/Var:%>^A # ;
+N+Sg+Par+Dial/Por:%>^V # ;
```

So far, only the Porsanki dialect has been distributed to language users. Having all three coexisting in the same computer would not be possible, as they must be referred to by the same ISO code, so if the need should arise we would have to ask the users to install only one of them.

3.3 Data-driven and/or rule-based language technology

A hot topic in NLP of 2020's is, what all can be done with large language models and chatbots. Our approach to NLP is based on traditional rule-based systems, with expert curated dictionaries and handwritten rules. For languages we talk about in this article, it can be easy to point out that for data-driven approaches we simply do not have enough data (c.f. Sable 4.2 for some statistics), while the methods of using little data improve, the amounts of data available for Baltic Finnic languages is insufficient for large language modelling. Another aspect that one has to keep in mind is the quality of the data: for machine learning to work, the data needs to be representative: follow the standards that the chatbot-based AI is supposed to use and contain ample examples of correct usage in various genres. With limited data and plenty of non-standard usage, the large language models will not be usable for spell and grammar checking and correction, while rule-based approaches can be steered to prefer and suggest current norms if available.

4 Resources and evaluations

In this section we list grammatical models in the GiellaLT infrastructure as well as corpus resources in GiellaLT and elsewhere. The statistics shown in this chapter are valid for the time of writing, since the language models are developed constantly, the

[10] for detailed technical description on TWOL refer to Koskenniemi (1983a)

figures will be outdated by the time of publication already. For this reason, automated generation of resources and evaluations are evaluated in the *continuous integration / continuous deployment* (CI/CD) systems and presented as up-to-date online statistics [11]. The relevant scripts are available in the github repositories[12].

4.1 Grammatical models

Within GiellaLT, there are grammatical models for 9 of the Finnic minority languages, cf. Table 2, which gives an overview of the lexical and morphosyntactic descriptions of the language models in our infrastructure.. Only two of them are described in publications (Meänkieli Trosterud (2020), Kven Trosterud et al. (2017)).

The size of morphosyntactic models can be measured in terms of how many lexemes they contain and the complexity of the morphophonological system can be approximated by combining the number of affixes used with the number of morphophonological alteration rules, covering suprasegmental and non-concatenative morphology as well as sandhi phenomena).

Language	ISO	Stems	Affixes	Rules
Ingrian	izh	2,163	2,361	45
Karelian	krl	66,096	555	1
Kven	fkv	46,354	5,096	56
Liv	liv	15,276	6,247	68
Livvi	olo	60,008	5,456	84
Ludic	lud	-	-	-
Meänkieli	fit	65,872	3,436	63
Veps	vep	6,280	2,011	10
Võro	vro	36,591	8,672	156
Votic	vot	1,030	190	10

Table 2: Grammatical models in the GiellaLT infrastructure (`https://giellalt.github.io/LanguageModels.html#uralic`)

4.2 Corpora

We have also curated corpora for some of these languages. The corpora are used for the development of the language technology tools: we collect spelling and grammar errors to test and develop writers tools, we collect the words and word forms to test the morphological implementations and use the sentences to test the automatic machine translation, to name a few. The GiellaLT corpora are summarised in Table 3.

There are also corpora for minority Finnic languages outside the Giellalt infrastructure. MetaShare contains a parallel corpus Võro - Estonian containing 171,252 Võro words as well as a monolingual Võro corpus of 350000 words (*https://metashare.ut.ee*). There are Bible texts available for Viena Karelian, Livvi and Veps (*https://www.finugorbib.com*), a parallel Bible corpus (Helsingin yliopisto, FIN-CLARIN et al., 2022) and an open corpus containing (in total) 2,66 million words for the same languages (cf. Boyko et al. (2022) for a presentation).

4.3 Evaluation

Using the corpora, it is possible to measure a naïve *coverage* gives an impression of how much of real world texts can be successfully processed with the resulting analyser; a näive coverage is measured as a proportion of surface tokens that gets *any* analysis at all without considering correctness, this gives a rough estimate of how well the analyser models the language in the form that is used in real world texts. It may be noteworthy to remember that, in the case of minority languages, real world texts can show a variance of non-standard forms and orthographies wider than established and standardised majority languages. In order to perform more thorough evaluation, we would need to co-operate with a language expert and develop hand-annotated gold standard corpora, for this article, that is left for future work. To get a qualitative insight on the quality of the analysers (or the data), for example the commonest words that are not analysed for each languages are: [13]:

- Meänkieli: *oova, och, nytten*
- Kven: *kirj., muist, đ*
- Livvi: *grigorianskoin, kargavusvuon, kalenduaruan*
- Veps: *km, Vellest, nell*
- Võro: *q, NOTOC, de*

5 Practical tools

Several language technology tools and softwares are implemented based on the morphological ana-

[11] `https://giellalt.github.io/CorpusResources.html`

[12] `https://github.com/giellalt/giella-core` and `https://github.com/divvun/actions`

[13] Both the source code for analysers and the corpora can be found at `https://github.com/giellalt`, in the repositories *lang-xxx* and *corpus-xxx*, respectively, where *xxx* is the relevant ISO code. Compilation is docuented at `https://giellalt.github.io`. Analysis was run at Oct 18th 2025.

Language	ISO	ktkn	MiB	Cov
Meänkieli	fit	528	12	90 %
Kven	fkv	1,115	21	92 %
Livvi	olo	242	4	87 %
Veps	vep	859	9	88 %
Võro	vro	265	4	90 %
Finnish	fin	16,694	382	—

Table 3: Corpora in the GiellaLT infrastructure. Finnish is listed for its relevance to machine translation. **ktkn** = thousand tokens, **MiB** = million bytes, **Cov** = coverage, or percentage re cognised by the analyser.

lysers and text collection. These tools are developed to support the language community, language revitalisation, standardisation, etc. We provide here experimental results of using these analysers in the context of these applications and corpora.

5.1 Keyboards and proofing tools

Keyboard drivers and tools for checking written language and correcting mistakes are crucial for literacy development in the digital era. Each literary language needs its own keyboard layout, for several reasons. The Finnic languages have different sets of letters in addition to the basic a-z set, typically around 6 additional ones, but ranging from 3 (Meänkieli) to 21 (Livonian). The optimal keyboard should be a compromise between keyboard tradition and placement of letters according to their frequency in running text. Then the keyboard users will expect non-letter symbols to be in the same positions as they are on the majority language keyboard. Kven and Meänkieli share the same alphabet (except for the Kven đ), but in addition, symbols such as `@, ', §, $, €` are placed (and engraved!) on different positions on Norwegian and Swedish keyboards, and the users of each minority language will expect these symbols to be in the same positions as they hold on the majority language keyboard. Finally, in Windows, the language of third-party proofing tools are identified by sharing ISO code with a keyboard driver. The same goes for mobile phones, where language support is always linked to the keyboard language.

The GiellaLT infrastructure contains a pipeline for easily setting up keyboard layouts for all computer and mobile phone operative systems, as well as keyboards for 8 of the Finnic minority languages [14].

Proofing tools include spell-checking and correction as well as grammatical error correction. The GiellaLT infrastructure is set up so that even a grammatical model can be turned into a spellchecker. The availability of proofing tools is thus obviously dependent upon the quality of the language model. The language models (see Table 2) are classified according to a 4-grade evaluation scale[15]. In addition, the spellchecker is dependent upon a suggestion mechanism as well as a text corpus in order to give precedence to more common words when correcting. A minimal suggestion mechanism contains approximately 50 rules (one for each letter or symbol to be suggested). Even a well-developed spellchecker in the GiellaLT does not contain more than appr. 300 suggestion rules. Table 4 gives an overview of status for the Finnic minority languages.

Language	ISO	Keyb	Spell	Sugg	W
Ingrian	izh	yes	Beta	56	-
Karelian	krl	yes	Alpha	89	-
Kven	fkv	yes	Prod.	301	yes
Liv	liv	yes	Alpha	109	-
Livvi	olo	yes	Beta	88	-
Ludic	lud	-	-	-	-
Meänkieli	fit	yes	Beta	220	yes
Veps	vep	-	Alpha	68	-
Võro	vro	yes	Beta	62	-
Votic	vot	yes	-	-	-

Table 4: Proofing tools in the GiellaLT infrastructure. **Spell** = quality level, **Sugg** = number of suggestion rules, **W** = corpus for weighting of suggestions

5.2 Rule-based machine translation

There are 6 Finnic language pairs within the Apertium (Khanna et al., 2021) rule-based machine translation system, cf. Table 5. Each language pair contains bilingual dictionaries, grammatical language models for analysis of L1 and generation of L2 as well as grammars for lexical selection and grammatical differences. As can be seen from the number of lexical entries, the language pairs range from usable machine translators to early stage projects.

5.3 Neural machine translation

The neural machine translation project *Neurotõlge* (*neurotolge.ee*, see Yankovskaya et al. (2023)) offers

[14] For an overview and links to the keyboards, see *https://giellalt.github.io/KeyboardLayouts.html#uralic-languages*

[15] For a definition of the various grades, see *https://giellalt.github.io/MaturityClassification.html*

Pair	**Entries**
Finnish—Livvi	30,212
Karelian—Livvi	6,419
Finnish—Kven	4,624
Karelian—Finnish	2,297
Vorõ—Estonian	161
Livonian—Finnish	37

Table 5: Machine translation models

Language	**Paradigm info**
Kven	10,557
Livonian	5,693
Livvi	3,538
Meänkieli	1,526
Veps	392
Võro	4,023

Table 6: Paradigm info

machine translation between (among other Uralic languages) the Finnic minority languages Livvi Karelian, Viena Karelian, Lude, Veps, Livonian and Võro and the majority languages Finnish, Swedish, Norwegian Bokmål and Russian. The monolingual corpora presented in Yankovskaya et al. (2023, 765) range from 5,000 (Ludic) to 115,300 and 162,000 (Veps and Võro) sentences. The amount of parallel sentences for the languages in Russia with Russian are 10,000 – 27,000, with the Bible dominating for all languages except Ludic.

Compared to their result for Finnish to Inari Saami and Norwegian to South Saami (which boast the quite good BLEU scores of 67.34 and 60.79, respectively), their results for the Finnic languages (op.cit. p. 768) are far worse (BLEU 24.17 for Estonian to Livonian and 30.63 for Estonian to Võro, the latter even worse than their previous result of 34.11). As shown by Yankovskaya et al. (2023), the main reason for this is the paucity of text, and the lack of balance for the parallel text, for the Finnic languages.

There are some existing critical evaluations of Neurotõlge for Sámi languages, c.f. Wiechetek et al. (2024, 2023), but these evaluations concentrate upon key semantic and grammatical elements of the translated texts rather than the overall closeness between translation and reference, as Yankovskaya et al. (2023) do.

6 Possibilities and perspectives

There are grammatical models for most Finnic minority languages, they show a coverage for running text on around or slightly 90 % (cf. Table 3). This is typical result achieved by rewriting formal grammars as grammar models. Grammars are seldom comprehensive, they typically sketch main patterns and obvious exceptions. In order to go the time-consuming work of getting a coverage of, say, 98 %, one has to include native speakers with knowledge of the norm in the team, so that they can add the description not included in the grammars. It is thus important that language researchers, teachers and learners are included in the process.

One way that the teachers and learners might help, is to simply provide paradigmatic information on word inflection. Providing simple information on a single word *häkki+N+Sg+Ade: häkil*, for example, provides the coder with information on gradation, and an adjacent plural form *häkki+N+Pl+Ade: häkkilöil*. These bits of information can be generated in a class environment where each student is given nouns, verbs or adjectives to describe in paradigms. The teacher checks to see that the forms are correct and the paradigmatic information is added to the infrastructure testing.

The GiellaLT infrastructure provides two different kinds of testing: One is impressionistic testing: Tools that generate parts of the model for the developer to inspect (e.g. generating all forms of a certain case). Another type is regression testing. Here, the linguist has set up for example model paradigms for parts of the morphology, and the model is tested continuously in order to ensure that it does not get worse.

There are test paradigms for the grammatical models of the Finnich minority languages to a various degree. Table 6 gives an overview of paradigm cells in the testing setup for the different languages. The figures might provide us with a picture of the time allocated to developing the different models. One could, of course, also add language-form information to the paradigmatic information, which could help solve problems in Veps, for example, where the Veps magazine *Kodima*[16] and the Veps edition of *Wikipedia*[17] are written in two different orthographies.

There is always a continuum of dialects and languages and standards within these minority lan-

[16] https://omamedia.ru/fi/publication/kodima
[17] https://vep.wikipedia.org

guages, one benefit of rule-based approaches is that they offer good control over the variation: It is possible to implement morphophonological rules and lexical analyses that concern specific variants. When this language technology is combined with a tool like spell-checking and correction, it is a powerful tool for language normativisation and support of writing culture. Experience with Kven has shown that the same lexica and morphological tagging structures can be used for describing language variants by river valley. Applied to Karelian languages, this might allow us to share mutual word stems, on the one hand, but distinguish morphological branches on the other. When it comes to sharing mutual lexica, it should be noted that the shared lexica are set off as their own groups. In work with Saami languages, proper noun lexica are shared. Even here, however, not all proper nouns can be shared. In work with the Permyak-Komi and Zyrian-Komi, additional sharing of lexica has been included for 100% matches in Russian loan words. For the Karelian languages using shared lexica is dependent on the use of parallel phonematic writing practices.

For future work, there is a lot that can be done in curating more lexical data and corpora for these languages. There is also a potential of developing speech technology applications based on the example of existing systems in Sámi languages. All of this requires collaboration, of course, between language communities and computational linguists. An important and ever more relevant issue in collaboration of language communities and computational linguists is ethical issues related to ownership of the language data and language itself, there has been a lot of research on this topic by us and others and we want to point towards (Wiechetek et al., 2024, 2022) for further references.

7 Conclusion

In this article, we have summarised the state of the art in minority Finnic language technology. We have shown that there exist some resources and have compared them to related languages to highlight the potential future possibilities these languages already have available.

The main part of the language technology work on Finnic so far has been concentrated on language models and proofing tools. For 5 of the 9 languages, we have developed grammatical models showing a coverage on running text extending 85 % (for three of them, 90 %).

The situation for available corpora is rather limited. Only for Kven and Meänkieli are there text collections available other than text from (Incubator) Wikipedias. To what extent the content of the corpora follow established standards is unclear. The corpora referred to here do not include all published text, but it is clear that the basis for data-driven language technology is shaky. In this perspective, we note on the positive side that despite this, there is neural-based MT for 5 of the languages presented here.

References

Daniel Abondolo and Riitta-Liisa Valijärvi, editors. 2023. *The Uralic Languages*.

Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors. 2022. *The Oxford Guide to the Uralic Languages*, 1 edition. Oxford Guides to the World's Languages. Oxford University Press, Incorporated, Oxford.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Eckhard Bick and Tino Didriksen. 2015. Cg-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 31–39.

Tatyana Boyko, Nina Zaitseva, Natalia Krizhanovskaya, Andrew Krizhanovsky, Irina Novak, Nataliya Pellinen, and Aleksandra Rodionova. 2022. The open corpus of the Veps and Karelian languages: Overview and applications. *KnE Social Sciences*.

Riho Grünthal. 2022. Veps. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 291–307. Oxford.

Riho Grünthal. 2023. The Finnic languages. In *The Uralic Languages*.

Helsingin yliopisto, FIN-CLARIN, Jack Rueter, and Erik Axelson. 2022. Raamatun jakeita uralilaisille kielille, rinnakkaiskorpus, Korp.

Katri Hiovain-Asikainen and Javier De la Rosa. 2023. Developing TTS and ASR for Lule and North Sámi languages. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 48–52.

Kinga Jelencsik-Mátyus, Enikő Héja, Zsófia Varga, and Tamás Váradi. 2022. *Report on the Hungarian Language*. European Language Equality (ELE).

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING '90 Proceedings of the 13th conference on Computational linguistics*, volume 3, pages 168–173, Helsinki.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hector Alos i Font. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

Kimmo Koskenniemi. 1983a. Two-level morphology: A general computational model for word-form recognition and production.

Kimmo Koskenniemi. 1983b. *Two-level Morphology. A General Computational Model for Word-forms Production and Generation*, volume 11 of *Publications of the Department of General Linguistics*. University of Helsinki.

Kimmo Koskenniemi, Krister Lindén, Lauri Carlsson, Martti Vainio, Antti Arppe, Mietta Lennes, Hanna Westerlund, Mirja Hyvärinen, Imre Bartis, Pirkko Nuolijärvi, and Aino Piehl. 2012. *The Finnish Language in the Digital Age*. Springer.

Johanna Laakso. 2022. Livonian. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 380–391. Oxford.

Johanna Laakso and Elena Skribnik. 2022. Graphization and orthographies of Uralic minority languages. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 91–100. Oxford.

Krista Liin, Kadri Muischnek, Kaili Müürisep, and Kadri Vider. 2012. *The Estonian Language in the Digital Age*. Springer.

Krister Linden and Wilhelmina Dyster. 2022. *Report on the Finnish Language*. European Language Equality (ELE).

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 28–47. Springer.

Elena Markus and Fedor Rozhanskiy. 2022a. Ingrian. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Elena Markus and Fedor Rozhanskiy. 2022b. Votic. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Sjur Nørstebø Moshagen, Flammie Pirinen, Lene Antonsen, Børre Gaup, Inga Mikkelsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain-Asikainen. 2023. *The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP*, volume 2 of *NEALT Monograph Series*, pages 70–94. NEALT.

Sjur Nørstebø Moshagen, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, and Per Langgård. 2022. *Report on the Nordic Minority Languages*. D1.38. European Language Equality (ELE).

Kadri Muischnek. 2022. *Report on the Estonian Language*, volume D1.12. European Language Equality (ELE).

Karl Pajusalu. 2022. Seto South Estonian. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Tommi A Pirinen. 2019. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136.

Helen Plado, Liina Lindström, and Sulev Iva. 2023. Võro South Estonian. In *The Uralic Languages*.

Bengt Pohjanen. 2022. *Meänkieli – Grammatik, lärobok, historik, texter*. Barents Publisher, Överkalix.

T. Rantanen, H. Tolvanen, M. Roose, J. Ylikoski, and O. Vesakoski. 2022. Best practices for spatial language data harmonization, sharing and map creation – a case study of uralic. *PLoS ONE*, 17(6).

Jack Rueter. 2024. Testing and enhancement of language models (transducers) from GiellaLT (scientific blog). *HAL*. 23 pages.

Anna-Kaisa Räisänen, Aili Eriksen, Thomas Brevik Kjærstad, and Trond Trosterud. 2024. Kvensk revitalisering, normering og leksikografi. *LexicoNordica*, 1(31).

Anneli Sarhimaa. 2022. Karelian. In *The Oxford Guide to the Uralic Languages*. Oxford University Press.

Eszter Simon, Piroska Lendvai, Géza Németh, Gábor Olaszy, and Klára Vicsi. 2012. *The Hungarian Language in the Digital Age*. Springer.

Steinþór Steingrímsson, Iben Nyholm Debess, Kimmo Granqvist, Per Langgård, and Trond Trosterud. 2024. *Language Technology for Less-Resourced Languages in the Nordics. Current Development and Collaborative Opportunities*. Stjórnaráð Íslands.

Eira Söderholm. 2017. *Kvensk grammatikk*. Cappelen Damm.

Sindre Reino Trosterud, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto, and Kaisa Maliniemi. 2017. A morphological analyser for Kven.

Trond Trosterud. 2020. Språkteknologi for meänkieli.

Tiit-Rein Viitso and Valts Ernštreits. 2012. *Līvõkīel-ēstikīel-leţkīel sõnārõntõz: = Liivi-eesti-läti sõnaraamat = Lībiešu-igauņu-latviešu vārdnīca*. Tartu Ülikool, and Latviešu valodas aģentūra.

Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur Moshagen, Flammie Pirinen, Trond Trosterud, and Børre Gaup. 2022. Unmasking the myth of effortless big data-making an open source multi-lingual infrastructure and building language resources from scratch. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1167–1177.

Linda Wiechetek, Flammie Pirinen, and Per Kummervold. 2023. A manual evaluation method of neural MT for indigenous languages. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 1–10.

Linda Wiechetek, Flammie A. Pirinen, Børre Gaup, Trond Trosterud, Maja Lisa Kappfjell, and Sjur Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephtey Adolphe, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Antonios Anastasopoulos, and 741 others. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).