Siberian Ingrian Finnish: FST and IGTs

Ivan Ubaleht
Omsk State Technical University
Omsk, Russia
last@gmail.com

Abstract

This paper presents the current version of the finite-state transducer for the Siberian Ingrian Finnish. Our finite-state transducer uses two-level morphology. We use LexC and TwolC languages together with HFST tools to develop lexicons and phonological rules, as well as to compile the transducer. The paper also provides a description of the morphological system of Siberian Ingrian Finnish. In addition, we present a collection of interlinear glossed texts in Siberian Ingrian Finnish, provided in a machine-readable format.

1 Introduction

The solution of computational morphology tasks is an important stage of language processing. Chapter 3 proposes a finite-state transducer based on the two-level morphology for addressing computational morphology tasks for Siberian Ingrian Finnish. Labeled data are also required for the successful solution of computational morphology tasks. In Chapter 4, we present interlinear glossed texts in Siberian Ingrian Finnish that have been published in a machine-readable format. Siberian Ingrian Finnish is a language with a rich morphological system; a brief overview of the language and its available resources is provided in Chapter 2.

2 Background

2.1 An overview of Siberian Ingrian Finnish

The Siberian Ingrian Finnish Language is an Ingrian Finnish – Ingrian (Izhorian) mixed language. The ancestors of the speakers of Siberian Ingrian Finnish spoke Lower Luga Ingrian Finnish and Lower Luga Ingrian varieties and lived in the lower reaches of the Luga River (Yamburgsky Uyezd). They were exiled to Western Siberia in 1803–1804 for their participation in a peasant uprising against Baron von Ungern-Sternberg (Kuznetsova, 2016, p. 14; Sidorkevich, 2014, pp. 23–24).

This language has been investigated by a number of linguists. D. V. Sidorkevich conducted research on this language between 2008 and 2014 (Sidorkevich, 2014, 2011). She introduced the term "Siberian Ingrian Finnish" (Russian "Сибирский ингерманландский идиом"). Siberian Ingrian Finnish was also studied by R. E. Nirvi (Nirvi, 1972), V. Zlobina (Zlobina, 1971, 1972), N.V. Kuznetsova (Kuznetsova, 2016; Kuznetsova and Verkhodanova, 2019), M. Z. Muslimov and F. I. Rozhansky. V. Zlobina introduced the term "Korlaks" (Russian "Корлаки", Finnish "Korlakat") to refer to the group speaking this language.

In 2025, there is still a group of elderly people who use Siberian Ingrian Finnish in domestic communication in the Ryzhkovo settlement (Krutinsky District of Omsk Oblast). Small groups and isolated speakers of Siberian Ingrian Finnish also live in other settlements of Omsk Oblast and in Estonia. A pessimistic estimate of the number of Siberian Ingrian Finnish speakers is about 30. This estimate is based on the fact that the author of this paper personally knows or is aware of 21 speakers of the language. An optimistic estimate, including semi-speakers, is about 100–150 people.

2.2 The language resources of Siberian Ingrian Finnish

The resources of the Siberian Ingrian Finnish language are summarized in Table 1. As can be seen in Table 1, a certain number of texts are currently available for Siberian Ingrian Finnish. Additional texts are planned to be collected through audio transcription. These texts require morphological glossing. Finite-state transducers provide significant assistance in the glossing process.

We have quite a large amount of audio data for Siberian Ingrian Finnish, see Table 1. Therefore, we previously created annotations for these au-

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 123–126

Resource type	Resource size
Audio data (2008-2025)	120 hours
Audio data published under a Creative Commons 4.0 license	5 hours
Video data	2 hours
Texts (mostly manual transcriptions of audio data)	42,000 tokens
IGT collection	150 sentences
Number of speakers recorded	31

Table 1: Language resources of Siberian Ingrian Finnish.

dio data and developed software for working with the annotations (Ubaleht and Raudalainen, 2022). However, the process of annotating audio data also requires automation, which became another reason for developing the Siberian Ingrian Finnish finite-state transducer.

3 Development of the Siberian Ingrian Finnish Finite-State Transducer

Currently, many computational morphology tasks, including those for low-resource languages, have been effectively addressed using models based on neural networks (Goldman et al., 2023; Wu et al., 2020; Liu, 2021). Nevertheless, approaches grounded in linguistic knowledge and employing finite-state transducers continue to provide benefits under certain conditions (Morozov et al., 2024; Merzhevich et al., 2022; Beemer et al., 2020). This is particularly pronounced in scenarios where processing morphologically rich languages is required and where training data are limited.

Siberian Ingrian Finnish lacks available training data. Currently, interlinear glossed texts for this language (which could serve as training data in the future) are still being created. Therefore, for solving morphological analysis and synthesis tasks for Siberian Ingrian Finnish, we are developing a solution based on finite-state transducers. We use the two-level morphology approach for developing finite-state transducers for Siberian Finnish, using LexC, TwolC, and the HFST toolkit (Lindén et al., 2011). The source code of the LexC and TwolC files for the current version of the finite-state transducer is accessible to the public on GitHub[1]. Currently, the transducer includes approximately 100 stems.

3.1 Morphological processing of Siberian Ingrian Finnish nouns

The morphological paradigm of nouns in Siberian Ingrian Finnish includes the declension of nouns by case and number, see Table 2. In its present state, Siberian Ingrian Finnish has eleven cases and two numbers. In practice, the adessive case and allative case have merged into a single syncretic adessive–allative case (Sidorkevich, 2011). However, in our finite-state transducer, we treat these cases separately. Siberian Ingrian Finnish nouns have five stems:

- NOM.SG: the stem used for the nominative singular form.
- OBL.SG: the stem used for all singular oblique cases except the illative and the partitive.
- PART.SG: the stem used for the partitive singular form.
- ILL.SG: the stem used for the illative singular form.
- OBL.PL: the stem used for all plural oblique cases.

Siberian Ingrian Ingrian Finnish nouns do not have a regular plural suffix like *-de* in Estonian or *-loi* in Izhorian. Plural forms are formed through stem alternations. These alternations are expressed by the OBL.PL stem. For example, the words *koir* (dog) and *sisar* (sister) belong to the same morphophonological type CS2 (Sidorkevich, 2014, p. 172), but *koira-n* (dog.SG-GEN), *koiri-n* (dog.PL-GEN) vs. *sisara-n* (sister.SG-GEN), *sisaro-n* (sister.PL-GEN).

Sixteen morphophonological types have been identified for Siberian Finnish nouns. D. V. Sidorkevich labels them as follows: CS1–CS8 for words with a consonant stem (Sidorkevich, 2014, pp. 164-165), and VS1–VS8 for words with a vowel stem (Sidorkevich, 2014, pp. 165-166).

In Siberian Ingrian Finnish, there are also morphophonological types that have not yet been documented. In Siberian Ingrian Finnish, a set of rules reflecting consonant alternations (CA1–CA5) and vowel alternations (VA1–VA5) is defined (Sidorkevich, 2014, pp. 161-162). Using alternation rules, it is not always possible to reliably derive the

[1] https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/src/morphological-analyzer/fst

Case	Singular			Plural		
	Stem	Affix	Example	Stem	Affix	Example
Nominative	NOM.SG	ø	*käsi*	OBL.SG	*-t*	*käe-t*
Genitive	OBL.SG	*-n*	*käe-n*	OBL.PL	*-n*	*kässi-n*
Partitive	PRT.SG	ø	*kätt*	OBL.PL	*-j*	*kässi-j*
Illative	ILL.SG	ø	*kätte*	OBL.PL	*-s*	*kässi-s*
Inessive	OBL.SG	*-s*	*käe-s*	OBL.PL	*-s*	*kässi-s*
Elative	OBL.SG	*-st*	*käe-st*	OBL.PL	*-st*	*kässi-st*
Allative	OBL.SG	*-l*	*käe-l*	OBL.PL	*-l*	*kässi-l*
Adessive	OBL.SG	*-l*	*käe-l*	OBL.PL	*-l*	*kässi-l*
Ablative	OBL.SG	*-lt*	*käe-lt*	OBL.PL	*-lt*	*kässi-lt*
Translative	OBL.SG	*-ks*	*käe-ks*	OBL.PL	*-ks*	*kässi-ks*
Comitative	OBL.SG	*-nkA*	*käe-nkä*	OBL.PL	*-nkA*	*kässi-nkä*

Table 2: Declension paradigm of nouns in Siberian Ingrian Finnish, with the example for morphophonological type VS3.

other noun stems from the main stem NOM.SG for all morphophonological types. When it is difficult to derive the other stems from the main stem NOM.SG, we record all five stems in the transducer lexicon, see *käsi* (1). As an example, (2) shows a lexicon that can be used to generate word forms from stem OBL.SG.

In some cases, it is possible to derive all stems from the NOM.SG stem using phonological rules from TwolC, so the word is represented in the lexicon by a single stem, see *koir*, *sisar* (1). We assume that in the future, for most morphophonological types, it will be possible to find phonological rules for a convenient representation of words in the lexicons.

(1) Lexicon containing noun stems

```
LEXICON NounStems

käsi:käsi   NomSgStem ;
käsi:käe    OblSgStem ;
käsi:kätt   PrtSgStem ;
käsi:kätte  IllSgStem ;
käsi:kässi  OblPlStem ;
koir:koir CS2-I ;
sisar:sisar CS2-O ;
```

3.2 Morphological processing of Siberian Ingrian Finnish verbs

The paradigm of verb inflection in Siberian Ingrian Finnish is not described in detail in this paper. The morphophonological types of Siberian Ingrian Finnish verbs remain poorly studied. D. V. Sidorkevich identifies 13 stems in the verbs of the Siberian Finnish language (Sidorkevich, 2014, p. 210).

Verbs such as *korja* (to pick up) and *harja* (to comb) can be represented in the lexicon by their infinitive stem, and all other forms are derived simply by adding suffixes. For all other morphophonological verb types, the lexicon must include between 3 and 13 verb stems. We assume that, by applying phonological rules, the number of verb stems in the lexicon can be reduced.

(2) Morphotactics for generating word forms from stem OBL.SG

```
LEXICON OblSgStem

+N+Gen+Sg:n # ;
+N+Ine+Sg:s # ;
+N+Ela+Sg:st # ;
+N+All+Sg:l # ;
+N+Ade+Sg:l # ;
+N+Abl+Sg:lt # ;
+N+Tra+Sg:ks # ;
+N+Com+Sg:nk%{A%} # ;
+N+Nom+Pl: t # ;
```

4 The Interlinear Glossed Texts in Siberian Ingrian Finnish

There are often no written texts available for many low-resource languages. Therefore, collections of interlinear glossed texts (IGTs) are important for providing these languages with linguistic resources.

D. V. Sidorkevich collected and glossed texts in Siberian Ingrian Finnish, but these texts were in a format not suitable for computational processing (Sidorkevich, 2014). We converted this collection of IGTs into a machine-readable format (3) and

made it openly available[2]; for example, a similar IGT format was used in the SIGMORPHON 2023 Shared Task on Interlinear Glossing[3].

(3) An example from our IGT collection

```
\t Miltajst sié kahest podarkast tahot?

\m miltajs-t sié kahe-st podarka-st taho-t

\g which-PRT 2SG two-ELA gift-ELA want-2SG

\l Which of the two gifts do you want?
```

5 Conclusion

In this paper, we have presented a finite-state transducer for Siberian Ingrian Finnish and a collection of interlinear glossed texts in this language. Future work includes: (i) expanding the transducer's lexicon to cover a larger vocabulary (we plan to add approximately 400-500 new stems to the lexicon by February 2026); (ii) developing phonological alternation rules to improve verb processing; (iii) glossing new Siberian Ingrian Finnish texts (including audio data annotations) using the FST; (iv) applying the FST and the IGTs in the language revitalization practices of Siberian Ingrian Finnish.

References

Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, and 1 others. 2020. Linguist vs. machine: Rapid development of finite-state morphological grammars. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 162–170.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. Sigmorphon–unimorph 2023 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125.

Natalia Kuznetsova. 2016. Evolution of the non-initial vocalic length contrast across the finnic varieties of ingria and adjacent areas. *Linguistica Uralica*, 52(1):1–25.

Natalia Kuznetsova and Vasilisa Verkhodanova. 2019. Phonetic realisation and phonemic categorisation of the final reduced corner vowels in the finnic languages of ingria. *Phonetica*, 76(2-3):201–233.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A Pirinen, and Miikka Silfverberg. 2011. Hfst—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Ling Liu. 2021. Computational morphology with neural network approaches. *arXiv preprint arXiv:2105.09404*.

Tatiana Merzhevich, Nkonye Gbadegoye, Leander Girrbach, Jingwen Li, and Ryan Soh-Eun Shim. 2022. Sigmorphon 2022 task 0 submission description: Modelling morphological inflection with data-driven and rule-based approaches. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–211.

Dmitry Morozov, Timur Garipov, Olga Lyashevskaya, Svetlana Savchuk, Boris Iomdin, and Anna Glazkova. 2024. Automatic morpheme segmentation for russian: Can an algorithm replace experts? *Journal of Language and Education*, 10(4 (40)):71–84.

Ruben Erik Nirvi. 1972. Siperian inkeriläisten murteesta ja alkuperästä. *Kotiseutu*, 2(3):92–95.

Daria V Sidorkevich. 2011. On domains of adessive-allative in siberian ingrian finnish. *Acta Linguistica Petropolitana.* Труды института лингвистических исследований, 7(3):575–607.

Daria V Sidorkevich. 2014. *The Language of Settlers from Ingria in Siberia: Structure, Dialect Features, Contact Phenomena. (*Язык ингерманландских переселенцев в Сибири: структура, диалектные особенности, контактные явления*)*. Ph.D. thesis, Institute of Linguistics of the Russian Academy of Sciences.

Ivan Ubaleht and Taisto-Kalevi Raudalainen. 2022. Development of the siberian ingrian finnish speech corpus. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–4.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv preprint arXiv:2005.10213*.

Vieno Zlobina. 1971. Who are the korlaks? (Кто такие корлаки?). Советское финно-угроведение, 7(2):87–91.

Vieno Zlobina. 1972. Mitä alkujuurta siperian suomalaiset ja korlakat ovat. *Kotiseutu*, 2(3):86–92.

[2] https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/IGT

[3] https://github.com/sigmorphon/2023glossingST/