Fedor Rozhanskiy (University of Tartu)

The corpus of Soikkola Ingrian: key issues and solutions

This paper discusses the experience of compiling a corpus for the Ingrian language – one of the minor moribund Finnic languages (Markus & Rozhanskiy 2022). The number of published texts in the Soikkola dialect of Ingrian is critically small. There are two fairy tales in Porkka (1885: 130–134, 143–145), three fairy tales in Sovijärvi (1944: 222–228), small collections of texts in Ariste (1960: 7–87) and Laanest (1966: 111–154) with translations into Estonian, several short texts in (Virtaranta 1967), and one short text in (Virtaranta & Suhonen 1978). The only published Soikkola text with morphological glossing is a fairy tale from (Porkka 1885: 130–134) annotated and commented by Rožanskij & Markus (2012: 448–503). The work on the corpus of the Soikkola dialect of the Ingrian language began in 2011, when a collection of Soikkola texts was recorded and processed as part of the project "Documentation of Ingrian: collecting and analysing fieldwork data and digitising legacy materials" (funded by the Endangered Languages Documentation Programme). The corpus work continues now in course of the project "Creating a multimodal corpus of Soikkola Ingrian" (funded by the Kindred Peoples' Programme). The aim of this talk is to discuss the difficulties and conceptual challenges encountered during this process and the solutions chosen to address them. The discussion is structured according to the various stages of the work.

Stage 1 – Data collection

At this stage, there were three main challenges. First, many speakers had not regularly practised communicating in the Ingrian language for a long time. Second, they were unaccustomed to speaking Ingrian in the presence of people they did not consider to be Ingrian. Third, the houses of many Soikkola residents were located close to a motorway, which was a source of considerable noise, and it affected the quality of the audio recordings. In order to overcome these challenges, spontaneous speech was recorded after some preliminary "warming up" elicitation work had been done, all researchers were instructed on techniques for encouraging speakers to switch to their native language, and the team was trained in audio recording methods under various conditions.

Stage 2 – Transcription of the texts

Transcribing Ingrian texts turned out to be an extremely time-consuming and even exhausting process, as the average age of the speakers was over 80. For people of this age, accurately repeating a fragment of text they had heard is a difficult task, due to both physiological factors (many elderly people have hearing issues), and psychological factors (speakers tend to

"correct" the grammar, semantics, and pragmatics of fragments they do not consider ideal). Also, it was not always possible to transcribe the text with the same speaker from whom it had been recorded. As a result, transcribing one minute of text often took 2–3 hours of work. To speed up the process, the following techniques were used: (1) the text was preliminarily transcribed by the researcher, and during the session with the speaker, this transcription was checked, and problematic fragments were clarified, (2) the entire audio file was divided into short, convenient fragments for listening, and (3) the entire transcription process was recorded on a voice recorder, so that it could be available anytime later.

Stage 3 – Compiling a primary collection of texts in ELAN

The main challenge at this stage, both technically and conceptually, was developing a transcription method. On a technical level, it was decided to minimise the number of characters with double diacritics in order to avoid future problems with searching for forms that could have been inputted in various ways. Conceptually, it was decided to create separate tiers for phonetic transcription and for standardised phonological transcription. For the latter, decisions were made regarding which types of variability to neutralise (e.g., pronunciation features of a particular idiolect) and which to retain (e.g., variation in grammatical forms). A separate tier for comments was also introduced to explain all differences (except for regular ones) between the phonetic and standardised transcriptions.

Stage 4 – Morphological analysis

Texts were converted from ELAN to FLEX, where morphological annotation is performed, after which the texts were converted back into ELAN. The main issues at this stage relate to the limitations of FLEX. Specifically, there are difficulties in handling compounds or compound-like forms (FLEX does not allow the use of a hyphen as a separator for parts of a compound and cannot process forms where the inflected part is not the last, e.g., *ke-l–ikkee* who-ADE-DER$_{INDEF}$ 'by someone'). Additionally, FLEX is designed for inflectional morphology and cannot analyse syncretic forms without markers as forms of the same word, e.g., *ava-* as the stem of the verb 'open' and *ava* as the 2Sg imperative form 'open!'. In all such cases, technical solutions must be applied, such as treating compounds as two separate words and combining them after conversion to ELAN or entering syncretic forms into the FLEX lexicon as separate lexemes.

Stage 5 – Corpus publication

The conceptual challenge at this stage is choosing the format. For the Ingrian language, the key issue is to find a format that allows the corpus to be used both by researchers and by members of the language community. The proposed solution is to (a) publish the corpus both

online and in the form of a printed text collection, and (b) accompany the texts with translations into both English (for researchers) and Russian (for the language community). In the printed version, the texts are presented both in a multi-line format with morphological annotation and as continuous text. A grammatical index is attached to the texts to facilitate the search for grammatical forms in the printed collection.

References

Ariste P. 1960. Isuri keelenäiteid. In P. Ariste (ed.), *Soome-ugri keelte küsimusi*, 7–116. Tallinn.

Laanest A. 1966. *Isuri murdetekste*. Tallinn.

Markus E. & Rozhanskiy F. 2022. Ingrian. In M. Bakró-Nagy, J. Laakso & E. Skribnik (eds.), *The Oxford Guide to the Uralic Languages*, 308–329. Oxford.

Porkka V. 1885. *Über den Ingrischen Dialekt: mit Berücksichtigung der übrigen finnisch-ingermanländischen Dialekte*. Helsingfors.

Rožanskij F. I. & Markus E. B. 2012. «Zolotaja ptica» (publikacija ižorskoj skazki, zapisannoj v XIX veke). *Acta linguistica petropolitana* 8(1), 448–503.

Sovijärvi A. 1944. *Foneettis-äännehistoriallinen tutkimus Soikkolan inkeroismurteesta*. Helsinki.

Virtaranta, Pertti. 1967. *Lähisukukielten lukemisto*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Virtaranta, Pertti & Seppo Suhonen (eds.). 1978. *Lähisukukielet. Finnic languages: Karelian, Ludic, Vepsian, Ingrian, Votic, Livonian*. Helsinki: Suomalaisen kirjallisuuden seura.