FEDOR ROZHANSKIY
University of Tartu & Institute for Linguistic Studies RAS

ELENA MARKUS
University of Tartu & Institute of Linguistics RAS

A new resource for Finnic languages: The outcomes of the Ingrian documentation project

Abstract The report introduces a new digital resource on minor Finnic languages. This resource is the main outcome of the project "Documentation of Ingrian: collecting and analyzing fieldwork data and digitizing legacy materials" carried out by Fedor Rozhanskiy and Elena Markus at the University of Tartu in 2011–2013. The collected materials cover several minor Finnic languages with a special focus on varieties spoken in Western Ingria: the Soikkola, Lower Luga, and Heva dialects of Ingrian, the Lower Luga varieties of Votic, and Ingrian Finnish. The resource contains (a) legacy recordings of different genres made by previous researchers in 1968–2012; (b) new audio and video materials recorded mostly in 2011–2013 by the project participants; (c) transcriptions and translations into Russian and English synchronized with sound and video using the ELAN software. Altogether the resource presents 510 hours of audio recordings, 21 hours of video recordings, and 15 hours of ELAN annotations. All media files in the resource are provided with detailed metadata specifying the place and time of the recording, sociolinguistic data about the speaker, the contents of the recording, and the access rights. The resource is available on the websites of the Endangered Languages Archive (London, UK) and the Archive of Estonian Dialects and Kindred Languages of the University of Tartu (Estonia).

Multilingual Finnic. Language contact and change. 303–326.
Uralica Helsingiensia 14. Helsinki 2019.
‹https://doi.org/10.33341/uh.85039›

1. Introduction

This paper reports on the project "Documentation of Ingrian: collecting and analyzing fieldwork data and digitizing legacy materials" carried out by Fedor Rozhanskiy and Elena Markus at the University of Tartu in 2011–2013. The project was financed by the Endangered Languages Documentation Programme housed at the School of Oriental and African Studies of the University of London.

The main project outcome is a new digital resource on minor Finnic languages. Initially, the work was focused on the documentation and description of the Ingrian language, but in the course of the project the scope was broadened to include other Finnic varieties spoken in the area, the first of which were the Votic language and Ingrian Finnish. The amount of collected data exceeded all expectations and for this reason language documentation became the main project focus while descriptive tasks were partially postponed for future work. The resulting resource contains materials on minor Finnic languages with a special focus on varieties spoken in Western Ingria.

This paper is organized in the following way. Section 2 lists the main goals and tasks of the project. Section 3 gives some information on Ingrian and related languages. Section 4 addresses the project methodology. Section 5 describes the resulting data resource: the collections, organization of metadata, archiving institutions, etc. Section 6 discusses further prospective work and research tasks based on the collected materials.

2. Project goals

The main goal of the project was to create a language resource with various types of data. A major part of this resource consists of audio recordings. Two main sources of materials were 1) legacy recordings made by previous researchers and 2) new field materials recorded in the course of the project by the project participants.

The task of collecting legacy materials was aimed both at preserving the highly valuable data and making them accessible for researchers. For obvious reasons we left out legacy recordings that are

stored in institutional archives, are fully digitized, and are already provided with metadata. Many recordings that we managed to collect were stored on magnetic media (cassettes or tapes) and risked being lost or damaged. Most of the collections were not provided with metadata and therefore had low research value (the information on the contents, time, and place of the recording is essential for most tasks).

Recording of new field materials was aimed not only at expanding the amount of data but also at filling gaps in previously collected materials. In particular, one of our tasks was to provide high quality recordings that can be used for various types of linguistic work including experimental phonetic research. Also, we tried to fill the genre gaps and record both spontaneous speech and elicitations. Legacy collections often present samples of spontaneous speech with a narrow focus on a particular topic, and prove a rather limited data source for many grammatical and phonetic questions. Additionally, we did video linguistic documentation that was totally missing in previous collections.

Summing up, the main goals of the project were the following:

1) To track down, digitize and prepare for archiving potentially endangered collections of audio recordings;
2) To make contemporary field recordings of different genres: narratives on various subjects, dialogues, elicitations of grammar, and phonetic questionnaires;
3) To provide video linguistic documentation;
4) To create a corpus of transcribed and translated recordings synchronized with sound (and video);
5) To provide a data basis for compiling dictionaries, grammars, and annotated text collections.

3. Ingrian and related languages

Ingrian belongs to the northern group of Finnic languages. At present it comprises two dialects, Soikkola and Lower Luga, which differ significantly from each other.[1] Lower Luga Ingrian is a convergent variety with a high degree of subdialectal variation (Rozhanskiy & Markus 2013b, 2014a) spoken along the lower course of the Luga River. Soikkola Ingrian is spoken on the Soikkola peninsula. Both dialects are located in the Kingisepp region of the Leningrad oblast (Russian Federation).

By our estimation, the total number of Ingrian speakers is now less than 20 people. The average age of the speakers is more than 80 years. In the majority of cases, middle-aged Ingrians either do not know the language at all, or their language competence is limited to a few sentences on everyday topics. Children neither speak, nor understand Ingrian. All speakers are bilingual, and for most of them the main language of communication is Russian. Until recently, some fluent speakers used their native language in everyday communication, mostly to talk with neighbors or relatives, but by now communication in Ingrian is almost gone.

The Ingrian written variety was introduced in the 1930's. School education in Ingrian lasted for several years, but in the beginning of 1938, Ingrian was banned from schools and teachers were repressed. At present, the language prestige of Ingrian is very low. Most people see no benefit in studying Ingrian and only few native Ingrians show any interest in their ancestors' language. Ingrian language courses are taught by Nikita Diachkov at the Ingrian museum in Vistino, but these courses do not involve fluent speakers. Most speakers are too old and have health problems, so they cannot participate in the courses. Some speakers are very skeptical about the new learners. Skepticism comes either from a monopolistic attitude ("We have suffered but preserved the language, but where have you been all that time? Now let us die peacefully with our language."), or from the idea that new learners make many mistakes and are not able to speak a "pure language" (see more details in Rozhanskiy & Markus 2013a: 270, 294).

1. Two other Ingrian dialects – Heva and Oredeži – are already extinct.

From a linguistic point of view, the Ingrian language is valuable not only as contributing to the general picture of the history and development of the Finnic family, but also as possessing some typologically rare features. In particular, the Soikkola dialect demonstrates a ternary contrast of consonant lengths, while the Lower Luga dialect has voiceless vowel phonemes.

Ingrian is perhaps the least studied Finnic language. Porkka's (1885) grammatical description and a school grammar by Junus (1936) are obviously outdated and incomprehensive, and do not correspond to modern standards of language description. Works by Laanest (1966, 1978, 1986) mostly address the historical development of Ingrian and dialect variation. There is no synchronic description of Ingrian that would reflect the contemporary state of the language. It is therefore essential to document the language while it is still spoken and thus create a basis for future language description.

Ingrian lexicography is in a slightly better condition. There is a dictionary by Nirvi (1971) that represents all Ingrian dialects, although the data on Lower Luga Ingrian are very limited[2]. There is also a dictionary of the already extinct Heva dialect by Laanest (1997).

In addition to the Ingrian language, the two languages that constitute a considerable part of the audio materials in the resource are the Votic language and the Ingrian dialects of Finnish (see Table 1 in Section 5 for a full list of languages represented).

The Votic language is on the verge of extinction. No more than five people can currently be considered fluent speakers. They represent the Liivtšülä-Luuditsa and the Jõgõperä varieties of the Western Votic dialect[3]. Both varieties are located in the Kingisepp region. The Votic language has never had a written variety. In the 1930s Votic children attended schools in Ingrian together with their Ingrian neighbors. The Jõgõperä and Liivtšülä-Luuditsa villages had a mixed Votic-Ingrian population. For that reason both varieties experienced Ingrian

2. A highly specific group of southern varieties of Lower Luga Ingrian is not represented at all, see Muslimov 2005 regarding Ingrian dialectology.

3. It is not certain whether there are any fluent speakers of the Kukkuzi variety, a mixed Votic-Ingrian variety (see Suhonen 1985; Markus & Rozhanskiy 2012) traditionally listed as a Votic dialect. The last fluent speaker that we worked with passed away several years ago.

contact influence, but the degree of influence differs depending both on the variety and the concrete speaker. The language was described in five grammars (Ahlqvist 1856; Ariste 1968[4]; Agranat 2007; Tsvetkov 2008 [the original manuscript from 1922]; Markus & Rozhanskiy 2017), and four dictionaries (Posti 1980[5]; Kettunen 1986; Tsvetkov 1995 [compiled in the 1920's]; Grünberg 2013[6]).

Ingrian Finnish comprises a large heterogeneous group of varieties spread across the entire territory of Ingria (from the Narva River up to Karelia). The information on the Ingrian Finnish dialects is very scarce; the published works address either some particular aspects or some concrete varieties (Galahova 1974, 2000; Leppik 1975; Kirpu 1989; Lehto 1996; Kokko 2007; Riionheimo 2007; Muslimov 2009, 2014; and others), and there is no comprehensive description of Ingrian Finnish. Although at the end of the 19th century, the number of Ingrian Finnish speakers in Ingria was around 130 000 (Musaev 2004: 26), at present it is no more than 500 native speakers[7]. Most of these speakers were born in the 1930s; their average age approaches 80 (Kuznetsova et al. 2015: 24–25). These dialects have been influenced by Standard Finnish (the degree of influence depends greatly on the region and particular speaker). In the middle of the 1930s, the number of Finnish schools in the region was quite impressive (Musaev 2004: 248 indicates more than 300 schools), but the language of instruction was Standard Finnish (Kuznetsova et al. 2015: 26).

4. The grammar (Ariste 1968) is the English translation of Ariste 1948 written originally in Estonian.
5. Posti 1980 is the dictionary of the Kukkuzi variety.
6. Grünberg 2013 is the second edition. The first edition in seven volumes was published in 1990–2011 (1st–4th volumes edited by Elna Adler and Merle Leppik, 5th–7th volumes edited by Silja Grünberg).
7. It is difficult to estimate the number of Ingrian Finnish speakers in other regions (Karelia, Estonia, and Finland).

4. Project methodology

An audio corpus of legacy recordings was compiled in the following way.

1) Agreements were made between the University of Tartu and the owners of the original legacy collections. The agreements specified the conditions for processing the collections and the details of archiving.
2) Four out of six collections were stored on magnetic media and had to be digitized. This work was done partially by the owners of the collections (collections from Ilya Nikolaev and the Institute of Language, Literature, and History of the Karelian Research Centre) and partially by Fedor Rozhanskiy and Elena Markus (collections from Mehmet Muslimov and Enn Ernits). All the recordings were digitized as WAV files with a sampling rate 44,1 kHz, 16 bit, mono.
3) The resulting files corresponded to the original side of the tape or cassette and often contained several working sessions with different speakers. In order to provide proper metadata, such files were cut into smaller files to separate the sessions and named in a uniform way (see Section 5.2. on naming conventions).
4) Finally, metadata files were provided for each recording (see the details in 5.2.1.). The relevant details about the recordings were obtained either from the original owners or from the recording itself. In the latter case, the project participants listened through the recordings in order to find the necessary information.

New field materials were recorded in the course of the project from Ingrian and Votic speakers. These sessions were aimed at collecting data not represented in legacy collections and providing a better quality of the data. In particular:

1) The recordings represent the contemporary state of the two languages[8].
2) Apart from the narratives that are traditionally considered the main type of language data by Finno-Ugric researchers, we recorded many hours of elicitations. These included nominal and verbal paradigms and specially designed phonetic questionnaires. Many of the recorded narratives were transcribed and translated with the help of language consultants[9].
3) All the sessions were recorded with high quality equipment providing modern recording standards: audio files in WAV format with a sampling rate 48 kHz, 16 bit, stereo.
4) Video documentation was done for both the Ingrian and Votic languages.

All recorded data were provided with detailed metadata files. A part of the recorded narratives and grammar sessions was transcribed and translated with the help of language consultants. The transcriptions were time aligned with media files (either audio or video together with the extracted waveform) in the ELAN software.

5. Project results

The main result of the project is the language resource on Ingrian and other Finnic languages. The resource consists of six collections. Five collections come from private owners: Ilya Nikolaev, Enn Ernits, Natalia Kuznetsova, Mehmet Muslimov, Fedor Rozhanskiy & Elena Markus. One collection comes from the Institute of Language, Literature, and History of the Karelian Research Centre in Petrozavodsk. All collections were prepared for archiving by Fedor Rozhanskiy and Elena Markus.

Table 1 shows the amount of data (audio and video) for each language represented in the resource.

8. Although both Votic and Ingrian languages are on the verge of extinction, they demonstrate development processes typical of fully alive languages (see Rozhanskiy & Markus 2014b) and should not be treated as degraded.

9. We prefer the term "language consultant" to the synonymic "language informant" (see, e.g., Mosel 2012: 76 on the specific connotations for both terms).

Language	Audio (hrs)	Video (hrs)
Ingrian	302.7	15.2
Votic	83.7	6.2
Ingrian Finnish	51.0	
Veps	9.0	
Karelian	2.6	
South Estonian	1.0	
Estonian	0.2	
Mixed varieties	59.3	
TOTAL	**509.5**	**21.4**

Table 1. The amount of data (in hours) for the six languages represented in the resource. Under mixed varieties we list (a) instances when speakers of different varieties are involved in one session; (b) sessions with speakers of mixed varieties (these are mostly various mixtures of Ingrian, Votic, and Ingrian Finnish).

5.1. Resource contents

This section provides details about the contents of each collection.

5.1.1. Collection compiled by Ilya Nikolaev

The materials collected by Ilya Nikolaev (Saint Petersburg, Russian Federation) comprise about 20 hours of Ingrian recordings. A major part of the data represents Soikkola Ingrian; a few recordings represent the Lower Luga dialect. The recordings were made in 1996–2002 on a mini cassette recorder. Most of the recording sessions are the interviews of Nikolaev with the speakers on different topics (life in the village, deportation to Finland, fishing, traditional festivities, the Ingrian language, personal biographies, and other topics). Several recordings contain elicitations of basic grammar and vocabulary.

The materials were digitized by Ilya Nikolaev; metadata were provided by Ilya Nikolaev, Fedor Rozhanskiy, and Elena Markus.

5.1.2. Collection compiled by Enn Ernits

The collection compiled by Enn Ernits (Tartu, Estonia) contains about 23 hours of recordings of Finnic languages: Ingrian (Soikkola and

Lower Luga), Votic, Veps, and Karelian[10]. The recordings (1971–1986) were made on a reel-to-reel recorder and later on a cassette recorder. Most of the recordings are conversations between the interviewers and speakers on various topics, the main focus being folk medicine, folk astronomy, and folk music. The interviewers are Enn Ernits, Tiiu Ernits, and occasionally other Estonian researchers, among them Paul Ariste. The interviews are done in the speaker's language.

The materials were digitized by Fedor Rozhanskiy; metadata were provided by Enn Ernits, Fedor Rozhanskiy, and Elena Markus.

5.1.3. Collection compiled by Natalia Kuznetsova

The collection compiled by Natalia Kuznetsova (Saint Petersburg, Russian Federation) contains about 30 hours of Ingrian recordings (Soikkola and Lower Luga dialects). Additionally, there is one recording session with a speaker of the Kukkuzi variety. The recordings were made in 2008–2012 on a digital recorder. Most of the data are elicitations of nominal morphology and phonetic questionnaires. Some of the recordings are interviews in Ingrian and Russian.

Metadata were provided by Natalia Kuznetsova, Fedor Rozhanskiy, and Elena Markus.

5.1.4. Collection compiled by Mehmet Muslimov

The collection by Mehmet Muslimov (Saint Petersburg, Russian Federation) contains more than 245 hours of recordings made in the Lower Luga area. Muslimov recorded minor Finnic languages spoken in the area, including the Lower Luga dialect of Ingrian, the Lower Luga varieties of Votic, the Kukkuzi variety, and the Lower Luga varieties of Ingrian Finnish. There are also a few recordings of the Veps, South Estonian, and Estonian (Siberian variety) languages made in the Lower Luga region. The recordings were made in 2000–2005 on a cassette recorder. The materials include talks on various topics (mostly festivities, food, school education, local languages, and folk linguistics), and elicitations of dialect vocabulary, paradigms, and simple sentences.

10. Veps and Karelian data were recorded in Boksitogorsk and Prionezhsk regions.

Muslimov visited almost all the villages in the Lower Luga area that had remnants of Finnic population in the beginning of the 21st century, and worked with most of the speakers. This is probably the largest collection of recordings of the Lower Luga varieties, and it is extremely valuable for the study of language contact, dialectology, and the history and culture of the region. Unfortunately, the quality of the recordings is not always good.

The materials were digitized by Fedor Rozhanskiy; metadata were provided by Mehmet Muslimov, Natalia Kuznetsova, Fedor Rozhanskiy, and Elena Markus.

5.1.5. Collection from the Institute of Language, Literature, and History of the Karelian Research Centre, Russian Academy of Sciences (Petrozavodsk, Russian Federation)

This collection contains more than 65 hours of recordings made by folklore researchers (Eino Kiuru, Elina Kylmäsuu, and others) in the late 1960s–1970s. A major part of the collection represents the Ingrian language (Soikkola, Lower Luga, and the now extinct Heva dialects). There are also some recordings of Ingrian Finnish and several Votic recordings. The materials were recorded using a reel-to-reel (sometimes cassette) recorder and were digitized by the Institute under the agreement made in the course of the project. The collection covers various genres of Ingrian folklore including songs, laments, riddles, folk rhymes, and proverbs. There are also interviews with the speakers about traditional ceremonies and beliefs.

Metadata were provided by the Institute, Ilya Nikolaev, Fedor Rozhanskiy, and Elena Markus.

5.1.6. Collection compiled by Elena Markus and Fedor Rozhanskiy

The collection compiled by Elena Markus and Fedor Rozhanskiy (Tartu, Estonia) contains about 120 hours of Ingrian (Soikkola and Lower Luga dialects) and Votic (Lower Luga varieties) recordings. A major part of the collection was recorded in 2011–2013 (the earliest recordings are from 2006). There are three types of data in this collection:

- audio recordings,
- video recordings,
- ELAN annotations (audio and video recordings aligned with transcription and translation into Russian and English).

Most of the audio materials were made with high-quality recording equipment (digital recorders with external electret microphones, sampling rate 48 kHz).

The collection includes both samples of spontaneous speech (narratives and dialogues), and elicitations (phonetic and grammar questionnaires, nominal and verbal paradigms).

Figure 1 plots the six collections from the point of view of languages represented and the time of the recording.

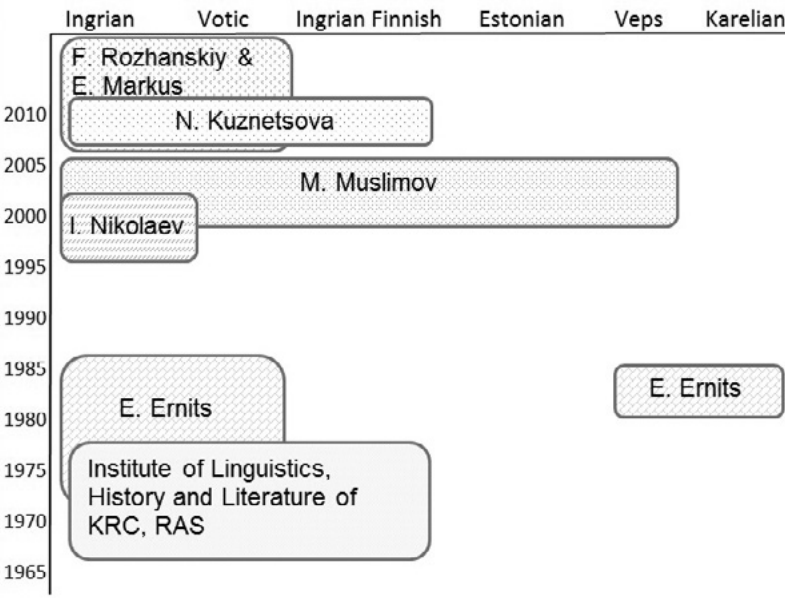

Figure 1. Languages and periods of data collection.

5.2. Representation of the data and structure of metadata files

In the resulting resource, the files are organized into bundles. Each bundle consists of files with the same name but different extensions that correspond to different data types. In the simplest variant, a bundle consists of two files: an audio file plus a metadata file, e.g.,

- ING_MAR_PAJA_OM110727.WAV (sound)
- ING_MAR_PAJA_OM110727.RTF (metadata)

In more complex cases, a bundle also contains a video file and an ELAN annotation, e.g.,

- ING_MAR_SKVORTSAD_EN120622.MPG (video)
- ING_MAR_SKVORTSAD_EN120622.WAV (sound)
- ING_MAR_SKVORTSAD_EN120622.EAF (ELAN transcription)
- ING_MAR_SKVORTSAD_EN120622.RTF (metadata)

File naming conventions are slightly different for each collection. A name always begins with the language code and the collection code, e.g., ING_NIK_VEN_SEV110502, FIN_KRC_01706A_SEPPANEN, where ING stands for Ingrian, FIN stands for Finnish, NIK and KRC denote the owners of the collection (Ilya Nikolaev and the Institute of Language, Literature, and History of the Karelian Research Centre, respectively). The remaining part of the name is a unique identifier that refers to the original recording on magnetic tapes and usually contains a reference to the native speaker.

Each file in the resource is provided with a metadata file. A metadata file is a table in the RTF format[11], see Table 2.

The metadata are presented in five subsections:

1) File (lists the file properties);
2) Recording (provides details about the recording, including the collector, recording date, processing stages, etc.);
3) Language consultant (provides sociolinguistic data about the speaker: the name, date, and place of birth of the speaker and his parents, the spoken variety, etc.);
4) Contents (briefly describes the contents of the recording);
5) Access (specifies the owner of the recording and access rights).

11. The archive of the University of Tartu has a special way of representing metadata: the metadata are given in special forms accompanying each audio file.

File	
File name	**ING_MAR_TORMI_ VV110729**
File type	**Sound**
File properties	**WAV, 48000 Hz, 16-bit, stereo**
Duration of sound	**00:02:19**
File size	**26.869.894 bytes**
Processing	**completed**
Recording	
Collector's name	**Fedor Rozhanskiy, Elena Markus**
Date of collection	**2011-07-29**
Recorded with	**Recorder: Edirol R-09HR, Microphone: Edirol CS-15**
Recording transcribed (by whom, when)	**No**
Recording digitized (by whom, when)	**N/A**
Recording annotated (by whom, when)	**N/A**
Recording prepared for archiving (by whom, when)	**Fedor Rozhanskiy, Elena Markus, March 2012**
Other file processing (by whom, when)	
Language consultant	
Language	**Ingrian**
Dialect	**Soikkola (South)**
Consultant name	**Petrova Valentina Vasilyevna**
Place of collection	**Slobodka [Säätinä]**
Consultant's date of birth	**1937**
Consultant's place of birth	**Krasnaja Gorka [Yhimägi]**
Consultant's parents (mother: native language, place of birth; father: native language, place of birth)	**Ingrian, Smenkovo [Otsave]; Ingrian, Krasnaja Gorka [Yhimägi], KIA in WWII**
Comments	
Contents	
Genre	**Narrative**
Detailed contents	**Story about fishing**
Intermediary language	**Ingrian, Russian**
Other comments	**Noise – thunder Interviewer: Elena Markus**

Access	
Rights to process the files	**Fedor Rozhanskiy, Elena Markus**
The owner of the recording	**Fedor Rozhanskiy, Elena Markus**
Location of original carrier	**N/A**
Access for listening	**No limitations**
Access for copying	**No limitations**
Comments	**Full or partial publication requires the owner's permit (handarey@yahoo.com)**

Table 2. A sample of a metadata file.

Additionally, all individual metadata files are combined in an Excel spreadsheet. The spreadsheet contains:

a) the main worksheet with a full list of files and their detailed characteristics (one line for one recording corresponds to one bundle);
b) a worksheet with the list of all the recorded speakers and basic sociolinguistic data;
c) a list of all mentioned settlements with parallel Russian and Finno-Ugric names and location.

The main worksheet lists the same metadata as the RTF files. Each line in the sheet corresponds to one individual RTF file. The spreadsheet allows scrolling through the whole dataset and applying all Excel possibilities to sort, filter, and search the data.

The worksheet with the list of speakers provides basic sociolinguistic data necessary for language analysis: the name, date, and place of birth, the native variety, place of recording, the native language and place of birth of the speaker's parents.

The third worksheet lists all settlements mentioned in the metadata and specifies the official Russian name, the original names in the local languages, the geographical coordinates, and the administrative region.

5.3. Archiving

The language resource created in the course of the project has been archived at four institutions:

- the Endangered Languages Archive (London, UK),
- the University of Tartu (Estonia),
- the Institute of Language, Literature, and History of the Karelian Research Centre (Petrozavodsk, Russian Federation),
- the Ingrian museum in the village of Vistino (the Kingisepp region of the Leningrad oblast, Russian Federation).

Each institution has some specific archiving conditions, so the amount of data deposited and the access conditions differ, see Table 3.

	Endangered Languages Archive	University of Tartu	Karelian Research Centre	Ingrian Museum
Location	London, UK	Tartu, Estonia	Petrozavodsk, RF	Vistino, RF
Online access	yes	yes	no	no
Audio files (WAV), hrs	510	510	510	240
Video files (mpg, mp4), hrs	21	–	21	–
ELAN annotation (eaf), hrs	15	–	12	–

Table 3. Archiving institutions.

The Endangered Languages Archive and the University of Tartu provide online access to the materials via the addresses ‹http://elar.soas.ac.uk/deposit/0147› and ‹http://www.murre.ut.ee/arhiiv/otsi.php›.

5.4. Access rights and restrictions

Under the agreements made with the owners of the original collections, the owners preserved full rights on their materials. The project participants were granted non-exclusive rights to process the recordings (for example, split and rename the files, provide metadata).

A major part of the resource is open access for scientific purposes. The conditions and a few restrictions on the data usage are the following:

- listening of the recordings does not require the owner's permit;
- copying of the materials can be free or can require the owner's permit. Each metadata file specifies access/usage conditions and restrictions, the owner of the particular recording, and the email address, which can be used for contacting the owner;
- full or partial publication of any materials requires the owner's permit.

The archives with online access (ELAR and UT) have some specific terms of use. ELAR has four access categories: free access, access for researchers, access for the members of the language community, and access for subscribers (for more details see http://www.elar-archive.org/using-elar/access-protocol.php). A major part of our resource allows free access. A small part of the materials is currently accessible only for subscribers, because the corresponding materials are being prepared for publication.

The UT archive has two access conditions: free access or access for authorized users. Most of the materials from our resource are freely accessible. Respecting the wishes of the owners, access to certain parts of the data requires authorization.

6. Further work

The materials accumulated in the course of the Ingrian documentation project open wide possibilities for further descriptive work. The main tasks intended to be carried out by the authors of this paper are the following:

1) Compiling a concise morphological dictionary of Soikkola Ingrian (around 2 000 entries). A major part of the materials used in the dictionary was collected during the documentation project. The first (electronic) version of the dictionary has been placed at ‹http://ingrian.org› (the electronic dictionary was designed by Fedor Rozhanskiy during the course of the project financed by the Kone Foundation in 2015–2018). This software builds inflectional paradigms and provides illustrative audio material for the dictionary entries. Preparation of the paper version of the dictionary is in progress. A similar dictionary is later planned for Lower Luga Ingrian.
2) Compiling a corpus of transcribed and annotated recordings. Within the documentation project, several hours of Ingrian narratives and dialogues were recorded, transcribed, translated, and aligned with the audio track in ELAN. The annotation work will be continued in order to enlarge the corpus.
3) Research on Ingrian prosody. As mentioned above, Ingrian has a non-trivial system of quantity relations. Based on the phonetic questionnaires recorded during the documentation project we intend to thoroughly investigate segmental and suprasegmental characteristics of the Soikkola Ingrian dialect.
4) Compiling a grammar of Soikkola Ingrian based on the data on phonetics, morphology, and syntax collected during the project.

Along with the descriptive tasks listed above, we intend to continue documentation and further expand the resource. In particular, we are planning to include more of our Ingrian and Votic field recordings. At the moment, our collection of Votic recordings comprises about 260 hours, and the Ingrian collection[12] approaches 950 hours.

12. The Ingrian collection includes recordings made by Fedor Rozhanskiy, Elena Markus, Natalia Kuznetsova, and some other participants of the Ingrian field trips.

References

Agranat, Tatjana B. 2007: *Zapadnyj dialekt vodskogo jazyka* [Western dialect of Votic]. Mitteilungen der Societas Uralo-Altaica, Heft 26. Moskva–Groningen.

Ahlqvist, August 1856: *Wotisk grammatik jemte språkprof och ordförteckning* [Votic grammar with language samples and vocabulary]. Helsingfors.

Ariste, Paul 1948: *Vadja keele grammatika* [A grammar of the Votic language]. Tartu: Teaduslik Kirjandus.

—— 1968: *A grammar of the Votic language*. Indiana University publications, Uralic and Altaic series vol. 68. Bloomington–the Hague: Indiana University.

Galahova, Lidia Ja. 1974: *Osnovnye osobennosti konsonantizma v finskix govorax Leningradskoj oblasti* [The main characteristics of consonantal inventory in Finnish varieties of the Leningrad oblast. PhD thesis]. Sankt-Peterburg: Leningradskij gosudarstvennyj universitet.

—— 2000: Čeredovanie stupenej soglasnyx v osnove slova v finskix govorax Leningradskoj oblasti [Grade alternation in Finnish varieties of the Leningrad oblast]. – *Kafedra finno-ugorskoj filologii: Izbrannye trudy k 75-letiju kafedry*. Sankt-Peterburg: Izdatel′stvo Sankt-Peterburgskogo universiteta. 115–133.

Grünberg, Silja (ed.) 2013: *Vadja keele sõnaraamat* [Dictionary of Votic]. 2., täiendatud ja parandatud trükk. Tallinn: Eesti Keele Instituut & Eesti Keele Sihtasutus.

Junus, Väino I. 1936: *Ižoran keelen grammatikka. Morfologia opettaijaa vart* [A grammar of the Ingrian language. Morphology for teachers]. Leningrad–Moskva: Ucpedgiz.

Kettunen, Lauri 1986: *Vatjan kielen Mahun murteen sanasto* [Vocabulary of the Votic Mahu dialect]. Castrenianumin toimitteita 27. Helsinki: Castrenianumin laitokset & Suomalais-Ugrilainen Seura.

Kirpu, Lilia 1989: O nekotoryx fonetičeskix osobennostjax markovskogo govora finskogo jazyka Leningradskoj oblasti [On some phonetic characteristics of the Markovskij variety of the Ingrian Finnish of the Leningrad Oblast]. – *Fenno-Ugristica* 15: 80–87.

Kokko, Ossi 2007: *Inkerinsuomen pirstaleisuus. Eräiden sijojen kehitys murteen yksilöllistymisen kuvastajana* [Scattered Ingrian Finnish. The development of selected cases as reflectors of the individualization of a dialect. PhD thesis]. University of Joensuu Publications in the Humanities 48. Joensuu. Available at: ‹http://urn.fi/URN:ISBN:978-952-219-036-9›

Kuznetsova, Natalia, Elena Markus & Mehmed Muslimov 2015: Finno-Ugric minorities of Ingria: the current sociolinguistic situation and its background. – Heiko Marten, Michael Rießler, Janne Saarikivi & Reetta Toivanen (eds), *Cultural and linguistic minorities in the Russian Federation and the European Union*. Multilingual Education 13: Comparative studies on equality and diversity. Berlin: Springer. 127–167.

Laanest, Arvo 1966: *Ižorskie dialekty. Lingvogeografičeskoe issledovanie* [Ingrian dialects. A linguistic-geographical study]. Tallinn: Akademija nauk Estonskoj SSR.

—— 1978: *Istoričeskaja fonetika i morfologija ižorskogo jazyka* [Historical phonetics and morphology of Ingrian. PhD thesis]. Tallinn: Institut jazyka i literatury.

—— 1986: *Isuri keele ajalooline foneetika ja morfoloogia* [Historical phonetics and morphology of Ingrian]. Tallinn: Valgus.

—— 1997: *Isuri keele Hevaha murde sõnastik* [Vocabulary of the Ingrian Heva dialect]. Tallinn: Eesti Keele Instituut.

Lehto, Manja Irmeli 1996: *Ingrian Finnish: Dialect Preservation and Change*. [PhD thesis.] Acta Universitatis Upsaliensis. Studia Uralica Upsaliensia 23. Uppsala: Uppsala University.

Leppik, Merle 1975: *Ingerisoome Kurgola murde fonoloogilise süsteemi kujunemine* [The formation of the phonological system of the Ingrian Finnish Kurgola dialect]. Tallinn: Eesti NSV Teaduste Akadeemia.

Markus, Elena & Fedor Rozhanskiy 2012: Votic or Ingrian: new evidence on the Kukkuzi variety. – *Finnisch-Ugrische Mitteilungen* 35: 77–95.

Markus, Elena B. & Fedor I. Rozhanskiy 2017 [2011]: *Sovremennyj vodskij jazyk. Teksty i grammatičeskij očerk. 2-e izdanije, ispravlennoje i dopolnennoje* [Contemporary Votic language. Texts and grammar. 2nd edition]. Sankt-Peterburg: Nestor-Istorija. Available at: ‹http://ingrian.org/Votic-grammar/›

Mosel, Ulrike 2012: Morphosyntactic analysis in the field: a guide to the guides. – Nicholas Thieberger (ed.), *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press. 72–89.

Musaev, Vadim I. 2004: *Političeskaja istorija Ingermanlandii v konce XIX–XX veke* [Political history of Ingria at the end of the 19th–20th century]. 2nd ed. Sankt-Peterburg: Nestor-Istorija.

Muslimov, Mehmed Z. 2005: *Jazykovye kontakty v Zapadnoj Ingermanlandii (nižnee tečenie reki Lugi)* [Language contacts in Ingria (the lower course of the Luga River). PhD thesis]. Sankt-Peterburg: Institute for Linguistic Studies of the Russian Academy of Sciences.

—— 2009: K klassifikacii finskix dialektov Ingermanlandii [On classification of the Finnish dialects in Ingria]. – Sergej Myznikov & Igor′ Brodskij (eds), *Voprosy uralistiki* 2009. *Naučnyj al'manax*. Sankt-Peterburg: Nauka. 179–204.

—— 2014: Zametki o moloskovickom ingermanlandskom dialekte [Notes on the Moloskovitskij variety of Ingrian Finnish]. – Valentin F. Vydrin & Natalia V. Kuznetsova (eds), *From Bikin to Banbaluma, from the Varangians to the Greeks. Field-inspired essays in honour of Elena V. Perekhvalskaya*. Sankt-Peterburg: Nestor-Istorija. 277–287.

Nirvi, Ruben Erik 1971: *Inkeroismurteiden sanakirja* [Dictionary of Ingrian dialects]. Lexica Societatis Fenno-Ugricae XVIII. Helsinki: Suomalais-Ugrilainen Seura.

Porkka, Volmari 1885: *Über den ingrischen Dialekt mit Berücksichtigung der übrigen finnisch-ingermanländischen Dialekte* [On the Ingrian dialect with respect to other Ingrian Finnish dialects]. Helsingfors: J. C. Frenckell & Sohn.

Posti, Lauri 1980: *Vatjan kielen Kukkosin murteen sanakirja* [Dictionary of the Votic Kukkuzi dialect]. Ainekset kerännyt Lauri Posti. Painokuntoon toimittanut Seppo Suhonen Lauri Postin avustamana. Lexica Societatis Fenno-Ugricae XIX, Kotimaisten kielten tutkimuskeskuksen julkaisuja 8. Helsinki: Suomalais-Ugrilainen Seura & Kotimaisten kielten tutkimuskeskus.

Riionheimo, Helka 2007: *Muutoksen monet juuret. Oman ja vieraan risteytyminen Viron inkerinsuomalaisten imperfektinmuodostuksessa* [Multiple roots of change. Mixing native and borrowed influence in the past tense formation by Ingrian Finns. PhD thesis]. Suomalaisen Kirjallisuuden Seuran Toimituksia 1107. Helsinki: Suomalaisen Kirjallisuuden Seura.

Rozhanskiy, Fedor I. & Elena B. Markus 2013a: Ižora Sojkinskogo poluostrova: fragmenty sociolingvističeskogo analiza [Soikkola Ingrians: a sociolinguistic study]. – *Acta linguistica Petropolitana. Transactions of the Institute for Linguistic Studies*. Vol. IX, part 3. Sankt-Peterburg: Nauka. 261–298.

—— 2013b: O statuse nižnelužskogo dialekta ižorskogo jazyka sredi rodstvennyx idiomov [On the status of Lower Luga Ingrian among related varieties]. – Tatjana B. Agranat, Olga A. Kazakevich & Egor V. Kashkin (eds), *Lingvističeskij bespredel – 2. Sbornik naučnyx trudov k jubileju A. I. Kuznecovoj*. Moskva: Izdatel′stvo Moskovskogo Universiteta. 219–232.

Rozhanskiy, Fedor & Elena Markus 2014a: Lower Luga Ingrian as a convergent language. – FINKA Symposium: On the Border of Language and Dialect. University of Eastern Finland, Joensuu, 4–6 June, 2014. Joensuu. [Abstracts.] 36–37.

Rozhanskiy, Fedor I. & Elena B. Markus 2014b: Dinamika morfologii vodskogo jazyka s načala XX veka [Dynamics in the morphology of Votic since the beginning of the 20th century]. – *Finno-ugorskie jazyki i kultury v sociokulturnom landšafte Rossii. Materialy V Vserossijskoj konferencii finno-ugrovedov*. Petrozavodsk, 25–28 ijunja 2014 g. Petrozavodsk: Karelskij naučnyj centr RAN. 118–121.

Rozhanskiy, Fedor & Elena Markus 2019: *Ingrian morphologial dictionary*. Electronic dictionary. Available at: ‹http://ingrian.org/Ingrian-dictionary/›

Suhonen, Seppo 1985: Wotisch oder Ingrisch? [Votic or Ingrian?] – Wolfgang Veenker (ed.), *Dialectologia Uralica: Materialen des ersten Internationalen Symposions zur Dialektologie der uralischen Sprachen 4.–7. September 1984 in Hamburg*. Veröffentlichungen der Societas Uralo-Altaica. Band 20. Wiesbaden: Harrassowitz. 139–148.

Tsvetkov, Dmitri 1995: *Vatjan kielen Joenperän murteen sanasto* [Vocabulary of the Votic Jõgõperä dialect]. Toimittanut, käänteissanaston ja hakemiston laatinut Johanna Laakso. Lexica Societatis Fenno-Ugricae XXV, Kotimaisten kielten tutkimuskeskuksen julkaisuja 79. Helsinki: Suomalais-Ugrilainen Seura & Kotimaisten kielten tutkimuskeskus.

—— 2008: *Vadja keele grammatika* [Grammar of the Votic language] (Эсимейн′ ваддя чээле грамаатикк. Первая грамматика водьского языка, 1922). Jüri Viikberg (ed.). Tallinn: Eesti Keele Sihtasutus.

Результаты проекта по документации ижорского языка: новый интернет-ресурс по прибалтийско-финским языкам

Федор Рожанский & Елена Маркус

Статья посвящена описанию нового интернет-ресурса, который стал главным результатом работы над проектом по документации ижорского языка. Проект выполнялся в 2011–2013 годах в Тартуском университете Ф. Рожанским и Е. Маркус при финансовой поддержке Программы по документации языков под угрозой исчезновения (Лондон, Великобритания). Исходно проект был нацелен в равной степени на документацию и описание ижорского языка, однако, впоследствии на первый план вышли задачи по документации. При этом производился сбор материала не только по ижорскому, но и по соседствующим малым языкам (прежде всего, водскому). Двумя основными направлениями проекта стали: (а) оцифровка и архивация записей ижорского и других малых прибалтийско-финских языков, сделанных предшествующими исследователями; (б) сбор нового полевого материала в соответствии с современными стандартами документации.

К архивации было подготовлено пять коллекций предшествующих исследователей: коллекция Ильи Николаева (около 20 часов записей ижорской речи и некоторое количество элицитаций, по большей части сойкинский диалект, 1996–2002 гг.); коллекция Эна Эрнитса (23 часа записей сойкинского и нижнелужского ижорского, водского, карельского и вепсского языков, 1971–1986 гг.); коллекция Натальи Кузнецовой (около 30 часов записей сойкинского и нижнелужского ижорского, а также куровицкого идиома, в основном анкеты по фонетике и грамматике, 2008–2012 гг.); коллекция Мехмеда Муслимова (245 часов записей нижнелужского ижорского, водского, куровицкого идиома, ингерманландского диалекта финского, а также других языков, интервью на разные темы и анкеты по грамматике и диалектологии, 2000–2005 гг.); коллекция Института языка, литературы и истории Карельского научного центра (более 65 часов записей сойкинского, нижнелужского и хэваского ижорского, ингерманландского финского и

водского, содержащих фольклорные тексты и интервью на тему обрядов и ритуалов, 1968–1977 гг.).

Шестую коллекцию составили полевые материалы исполнителей проекта, включающие в себя около 120 часов аудиозаписей ижорского и водского языка (образцы спонтанной речи и анкеты по грамматике и фонетике), а также видеозаписи и аннотации в программе ELAN (транскрипция, сопровождаемая русским и английским переводами). Основная часть записей сделана в 2011–2013 гг.

Для всех файлов были подготовлены подробные метаданные, содержащие информацию о медиафайле, месте и времени записи, носителе языка, содержании записи и правах доступа.

В общей сложности ресурс содержит 510 часов аудиозаписей, 21 час видеозаписи и 15 часов аннотаций в ELAN.

Ресурс размещен на сайте Архива языков под угрозой исчезновения (Лондон, Великобритания) и на сайте Архива эстонских диалектов и родственных языков (Тарту, Эстония). Также копии ресурса переданы в архив Института языка, литературы и истории Карельского научного центра (Петрозаводск, Россия) и в Ижорский музей (Вистино, Ленинградская область, Россия).