

Siberian Ingrian Finnish Talking Dictionary

19th Annual Conference of Applied Linguistics 2022

Ivan Ubaleht
Omsk State Technical University

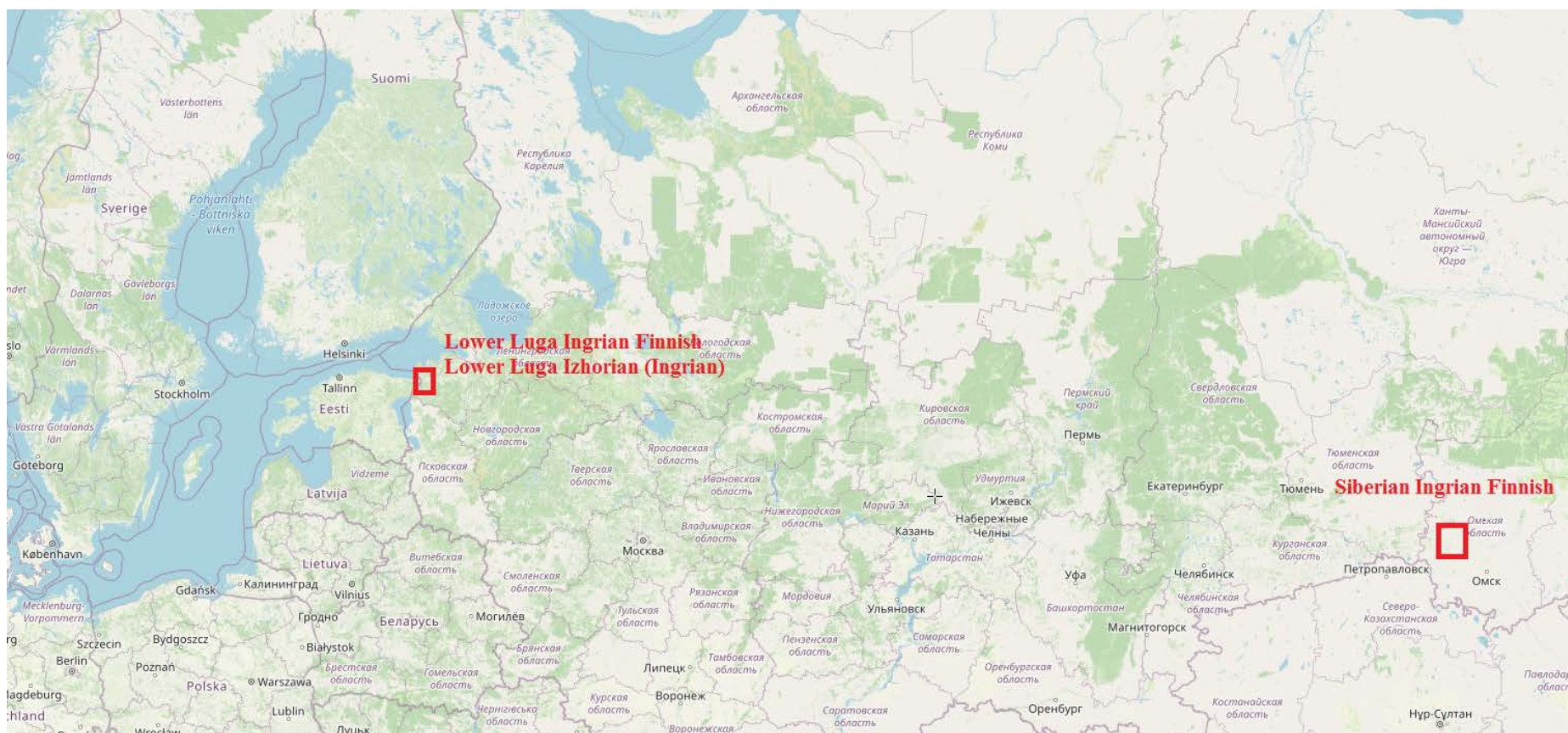
Tallinn 2022

Contents

1. An overview of the Siberian Ingrian Finnish language
2. Our goals and related work
3. The design and development of the talking dictionary
 - 3.1 The general structure of the talking dictionary
 - 3.2 The data collection
 - 3.3 The annotations
 - 3.4 A parser for ELAN annotation files *.eaf (XML)
 - 3.5 The fieldwork database for the talking dictionary
 - 3.6 The web application for the Siberian Ingrian Finnish talking dictionary
4. Conclusion and future work

1. An overview of the Siberian Ingrian Finnish language

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. They migrated from the Lower Luga area to Siberia in 1803-1804.

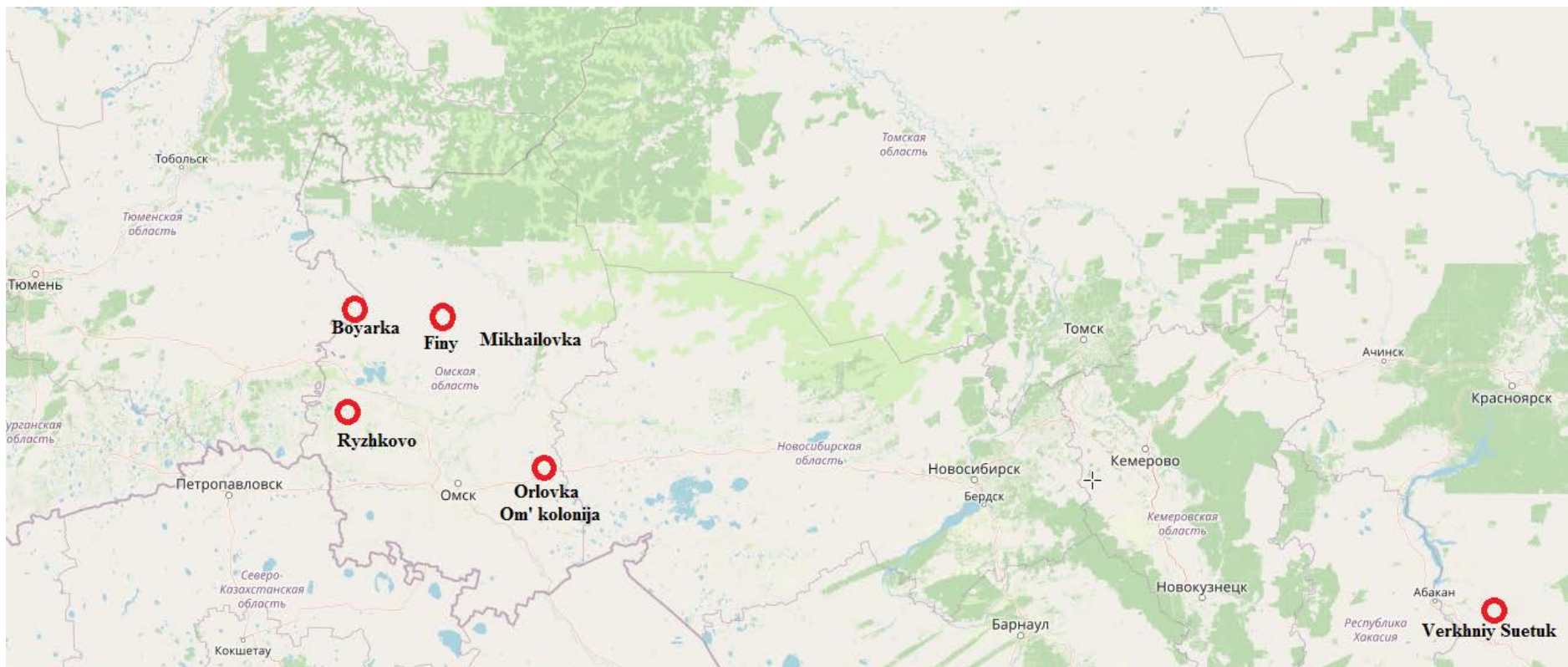


1. An overview of the Siberian Ingrian Finnish language

Lower Luga Dialect area at the end of the 20th century



1. An overview of the Siberian Ingrian Finnish language



Siberian Ingrian Finnish in Siberia late 19th century - early 20th century

At the beginning of the 21st century, communities of speakers of this language remained only in Ryzhkovo village and Omsk.

About 100 native speakers and semi-speakers of this language now live in Russia and Estonia, this information is relevant for 2022.

1. An overview of the Siberian Ingrian Finnish language

Siberian Ingrian Finnish has a number of distinctive features such as word-final vowel reduction and the emergence of a large number of consonant phonemes. The language has a distinctive features in morphology and lexicon.

Example of nominal inflectional paradigm, word **kukk** (flower):

Nouns have 5 stems: main stem (**kukk**), indirect stem SG (**kuka**), stem PART.SG (**kukka**), stem ILL.SG (**kukka**), indirect stem PL (**kukki**)

SG.:	Nom.	kukk	PL.:	Nom.	kuka+t
	Gen.	kuka+n		Gen.	kukki+n
	Part.	kukka		Part.	kukki+j
	Ill.	kukka		Ill.	kukki+s
	Ines.	kuka+s		Ines.	kukki+s
	Elat.	kuka+st		Elat.	kukki+st
	Ad./All.	kuka+l (kuka+n pääl)		Ad./All.	kukki+l
	Abl.	kuka+lt		Abl.	kukki+lt
	Transl.	kuka+ks		Transl.	kukki+ks
	Com.	kuka+n'ka		Com.	kukki+n'ka

1. An overview of the Siberian Ingrian Finnish language

Sample text in Siberian Ingrian Finnish:

Ei kaukal küläst ennempi oli müllü. Müllü pitivæt suamet. Suomalaiset. Sinn müllü piti pall' neit püh'kij, mašina, pall' räpakkoj. Bor miun mamm korjas neit räpakkoj lapsiŋka, veivæt sill' mullü, sis heil' antovat siel' vähä rahha, neist räpakkost. Kosk tuli küllä suur базар, sis hüö sel' bazarill ostivæt prännikkoj itsel' siel'. Hy, mitä saivat ostivæt? A kui viel' mänivæt neit räpakkoj viämäsel' tulti kovatsen poikset vassta. Tahtovat ottanet räpakot vällä. No miul mamm oli oike juulke, hä mäni iess, toiset tütset mänivæt jäless sinn, et poikset juoksivat takko.

Kui mie viel' olin pien mie mammonka tihti kävim metsäs marjass. Ei liika kauk, ono kilometra viis, oli külä Uunihh. Hä nüt sitä küllä ennä siel' ei ole. I mie kosk olin pien, hänt ennä ei ol. Bor naiset männöt iess, haastat kaikillaisij, kaikillaisist kummituksist. A mie olin tüttenen, mään heil' jäles, pöl'kän, haastat et ennep sinn olti hu... huntin nura. Huntit olivat. I kummituksij nä... näkivæt metsäs näi setä. Mie ni kovast mänin hejjen jäles i pöl'ksin. Katso naen ümper rinti näitses et, bor nüt hunt' i hüppiija metsäst vällä. A karut ennempi ei olt, karuj. Olivat lossit, huntit. Karut nüt tulivat, metsät pallat ümpär rintij i karut tullos tähä kera. Nu bor müö, isä viel' oli elos, müö löim hänenkä heinä. Hä vei viikkattenkä heinä, a mie haravoitin. Mie lüvv liika en ossa. Bor haravoitin, haravoitin, nossen pään ülläl', katson, kenelle seiso. Dumsin, hy, karu! Juaksin isän juure, hä: määm katsoma. Hä, a siel' ss... palant puolessa koivu, must. A mie dumsin et karu seiso. Kummituksest en muis mittäkk. Aha. Naiset mänivät, haastovæt et kummituksij näkivät, a... mitä hüö näkivät hüö ei haastanet miul. Hüö если б its keskel' mitäle haastovæt...

Kosk oli sota aik, a mamm piti lehmä. Hein loppu sü... kevväl' i hä mäni kelkaŋka, lyyžonka i häl' oli suur koir. Neit pitivät sinn, ei el'... ei kaukal külä Partizan oli, hüö pitivät koirij suurij, suan jaoks. Afčar, neit, afčarkat suuret olivat. Nu i üks oli miksle, hy, выбракованный. Hä sen penikan häl' kenelle ant. Hä kazv suureks koiraks. Hy, mamm mäni metsä otsan koiran ühess. Вотъ, koiran ühess, hy, пока hä sel' pani sitä heinä kelkan pääl', hy, kuulə mikäle siel' шy(y)m mokomain. I kits ajakminentle (неразб.). A hä katso, koir püis' kitsin kii. Hy, tapp hänen, tapp se koer sen kitsin. Hy и nüt noisti kotti mänemä, mamm viep sitä heinä kelkaŋka, nuar ümper vatsa i lyyžon pääl' kelkaŋka. I se koer tempija sitä kitsl, lunt müö kera, koti puole. Mamm pani sen kitsin kuarman pääl', toi kotti. Hy i siz viel' kenelle pelot' et' pannot häl' štraafin seest kitsist, hä sen kitsin peit' i aitta. I sis vähä havolt veitsel' koiral, süöt sitä koira sen kitsiŋkä. I its miksle ei süänt hänt, vai olis van its sü(ä)t, a... koer hänen repi, hy, tapp hänen. Hä jo i pöl'kü, sitä tapp hänen koer. Sota aik-to ei olt mitä süvv. Hä, se koer oli viel' perra soan, jo miun isä tuli kotti, siel't Kalymast laagerist. Se koer viel' hel' oli ja kuus vist hä sait hänen. Kaukan eli, se koer oli.

Recorded and transcribed by Daria Sidorkevich

1. An overview of the Siberian Ingrian Finnish language

The review of research studies dedicated to Siberian Ingrian Finnish

Siberian Ingrian Finnish (Russian: Сибирский ингерманландский идиом) is the term introduced by Daria Sidorkevich. Daria Sidorkevich with the participation of Mehmet Muslimov and Fedor Rozhanskiy researched and documented Siberian Ingrian Finnish in 2008-2014. Several expeditions were undertaken to Omsk oblast (Ryzhkovo, Mikhailovka, Orlovka settlements), Tyumen oblast and Estonia in 2008-2011. Approximately 80 hours of the field audio data were recorded during these expeditions. Ph.D. thesis was written by Daria Sidorkevich in 2014. The Siberian Ingrian Finnish phonology, morphology as well as certain other aspects were described in detail in this Ph.D. thesis. At this moment, Natalia Kuznetsova is also researching this language.

I started researching this language in 2019. At this moment, I am researching this language together with the Estonian Academic Society of Ingria. This research was supported by Estonian Non-Profit Organisation Fenno-Ugria in 2021 (“Lääne-Siberi ingerisoomlaste keele dokumenteerimine ja turgutamine”).

2. Our goals and related work

Since we plan to document and revitalize several endangered languages from our region, we are developing software – the Lexeme.Net system, at www.lexeme.net that is adapted to our requirements and goals.

Our goals and requirements:

- All source code and data should be accessible on GitHub and licensed under one of a free license;
- Speech corpora should be available to users via the Internet without installing additional software;
- Speech corpora should be convenient not only for linguists, but also for speakers of endangered languages, language activists and software developers;
- Speech corpora should have a powerful system of requests to data

2. Our goals and related work

At present, there are solutions that meet these requirements: the “Tsakorpus” corpus platform, Kwaras and Namuti (Caballero et al., 2019), LingSync & the Online Linguistic Database (Dunham et al., 2014), Kratylos (Kaufman and Finkel, 2018), the IATH ELAN Text-Sync Tool (Dobrin and Ross, 2017). Since we wanted a very flexible solution, we decided to develop own project. We choose the .NET Framework and Microsoft SQL Server for the implementation of the project.

We started to create the open Siberian Ingrian Finnish speech corpus in 2019. This corpus based on our own audio data from our expeditions and telephone interviews with speakers. A talking dictionary is the first step to address the creation of the speech corpus for the Siberian Ingrian Finnish language. Talking dictionaries are used to revitalizing and linguistic research of endangered languages.

Gabriela Caballero, Lucien Carroll, and Kevin Mach. 2019. Accessing, managing, and mobilizing an ELAN-based language documentation corpus: The Kwaras and Namuti tools. *Language Documentation & Conservation*, 13:63-82.

Tsakorpus: <https://github.com/timarkh/tsakorpus>

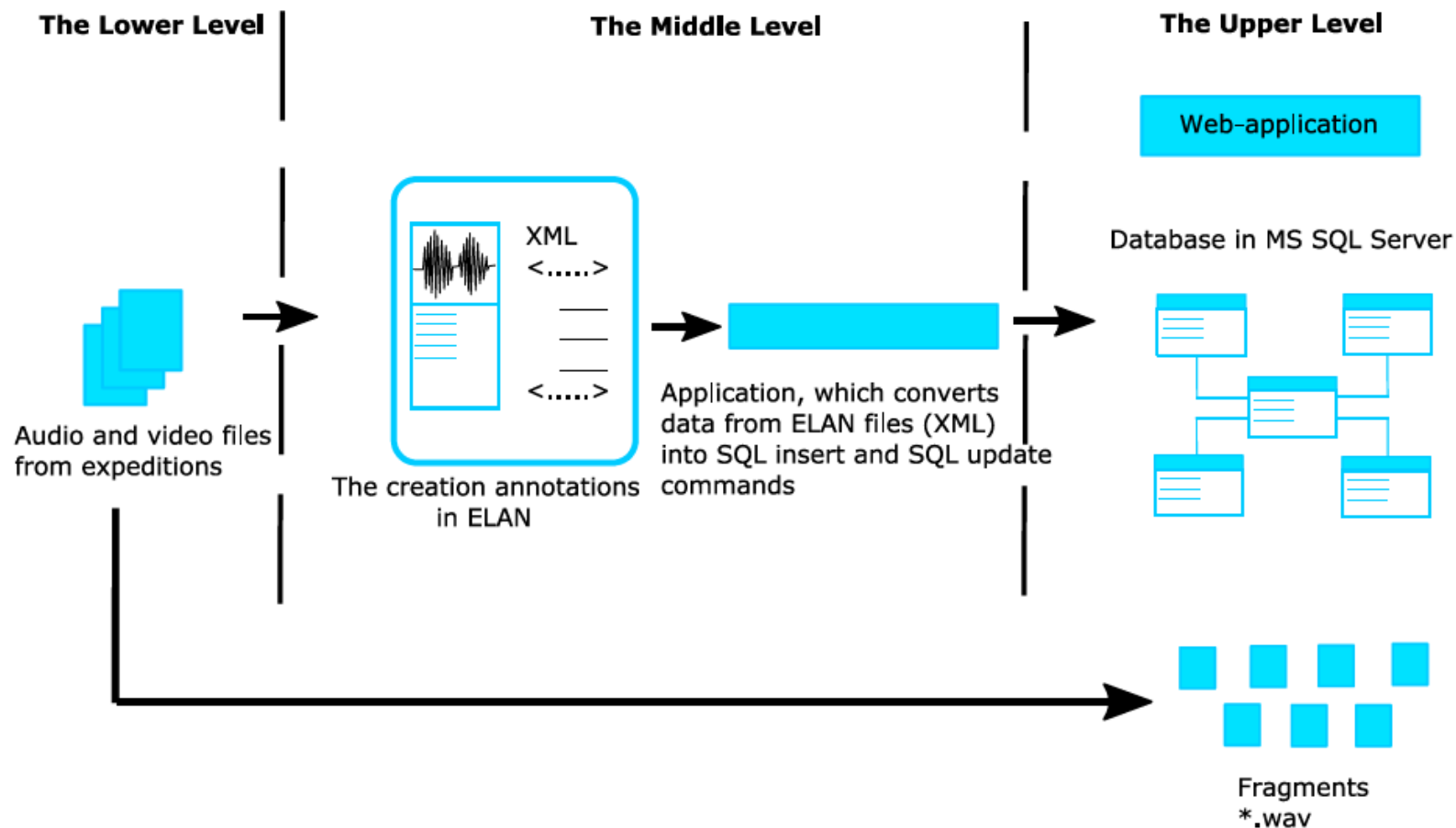
Joel Dunham, Gina Cook, and Joshua Horner. 2014. LingSync & the Online Linguistic Database: New models for the collection and management of data for language communities, linguists and language learners. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 24-33. <https://doi.org/10.3115/v1/W14-2204>

Daniel Kaufman and Raphael Finkel. 2018. Kratylos: A tool for sharing interlinearized and lexical data in diverse formats. *Language Documentation & Conservation*, 12:124–146.

Lise M. Dobrin and Douglass Ross. 2017. The IATH ELAN Text-Sync Tool: A Simple System for Mobilizing ELAN Transcripts On- or Off-Line. *Language Documentation & Conservation*, 11:94 – 102.

3. The design and development of the talking dictionary

3.1 The general structure of the talking dictionary



3.2 The data collection

1. Our speech data and annotations of Siberian Ingrian Finnish has been published and are accessible to the public on **GitHub** (5 hours):

<https://github.com/ubaleht/SiberianIngrianFinnish>

CC BY 4.0 License

Total amount of audio data from our own expeditions (2019-21): 15 hours.

2. At the moment we have all the data from Daria Sidorkevich's expeditions (2008 –2011). That's 80 hours of audio data, audio with synchronous texts - 2 hours and transcribed texts. At the moment, this data is under the copyright of Daria Sidorkevich but Daria allows to use her data in work.

The screenshot shows the GitHub repository page for 'Siberian Ingrian Finnish Talking Dictionary' by user 'ubaleht'. The repository has 1 branch and 0 tags. It was created 17 hours ago and has 77 commits. The README file is visible, titled 'The Siberian Ingrian Finnish Language'. The README text describes the project's goal: to create a talking dictionary for the Siberian Ingrian Finnish language, which is a dialect used by descendants of settlers from the Lower Luga area in the early 19th century. It mentions that the project is based on data from Daria Sidorkevich's expeditions (2008-2011) and the author's own expeditions (2019-21). The README also includes a list of references, a section for speech data (a Google Drive link), a section for video data (a YouTube link), and a license section (CC BY 4.0). At the bottom of the README, there is a table titled 'Description of the Speakers' with columns for Code of the Speaker and Gender, Year of Birth, Current Place of Residence, Place of the Birth, Birthplace of parents, and Speech Data (Duration). The table lists 10 speakers with their respective details.

Code of the Speaker and Gender	Year of Birth	Current Place of Residence	Place of the Birth	Birthplace of parents	Speech Data (Duration)
AAK-47 (M)	1947	Ryzhkovo	Syade	mother: Ryzhkovo, father: no data	40 min 57 s
IAI-33 (F)	1933	Oglukhino	Ryzhkovo	both parents: Ryzhkovo	33 min 14 s
JuMS-28 (M)	1928	Ryzhkovo (died)	Ryzhkovo	both parents: Ryzhkovo	77 min 53 s
KKM-34 (M)	1934	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	31 min 29 s
MAP-49 (F)	1949	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	30 min 36 s
MMM-39 (M)	1939	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	62 min 20 s
PGM-56 (F)	1956	Omsk	Finy	both parents: Finy	8 min 20 s
SVM-29 (M)	1929	Mikhailovka (died)	Larionovka	both parents: Yamburgsky Uyezd, Saint Petersburg Governorate	10 min 36 s

3.2 The data collection

The speech data of Siberian Ingrian Finnish are available in our repository on GitHub and licensed under a Creative Commons Attribution 4.0 license (CC BY 4.0). Currently, the part of the audio data from our expeditions has been published. We recorded 15 hours of audio and 2 hours of video from 9 speakers during our expeditions to Ryzhkovo and Mikhailovka settlements (Omsk oblast, Russia) and interviews via phone in 2019-2022. About 5 hours of the audio data were published on GitHub.

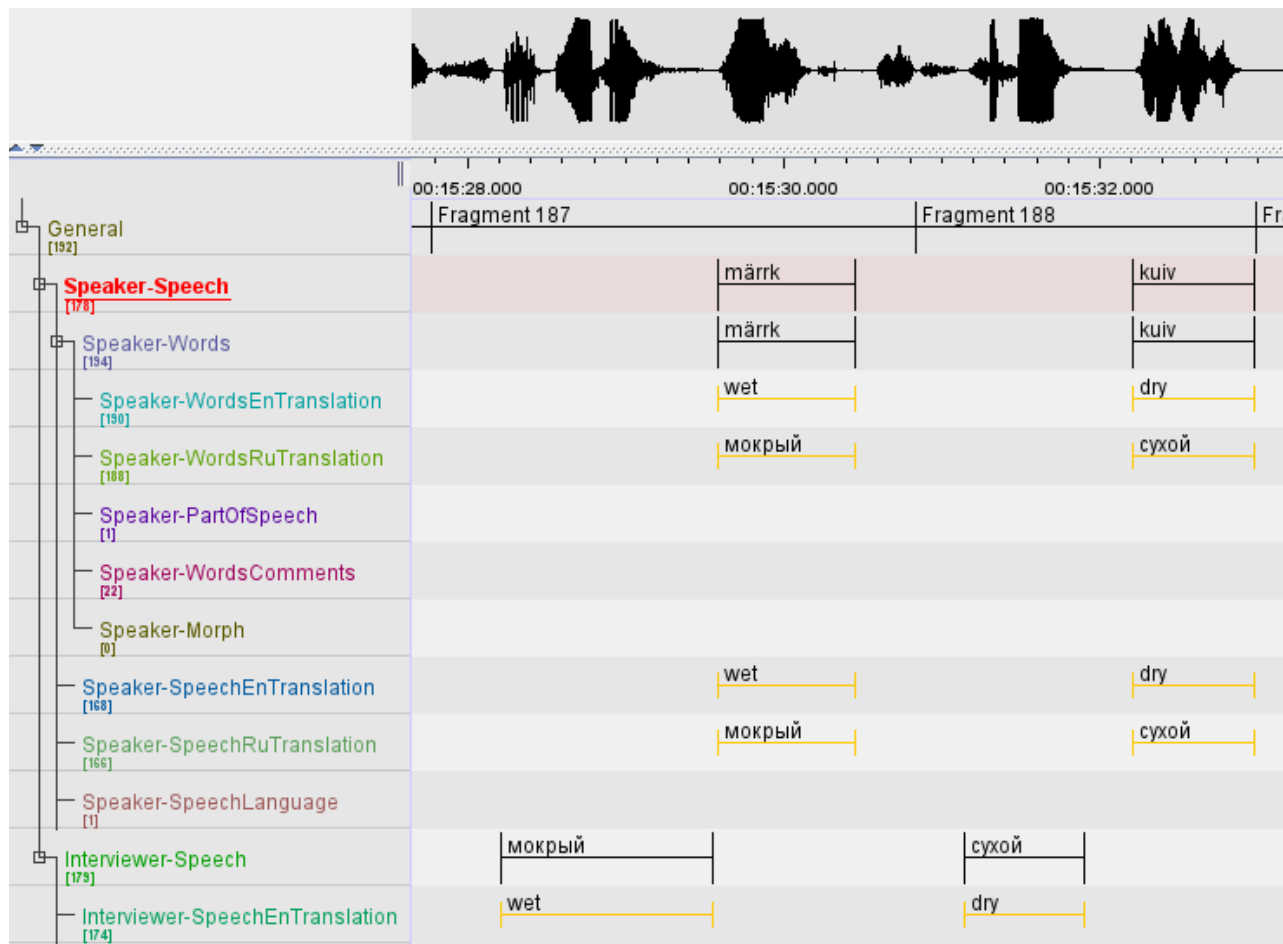
Code of Speaker and Gender	Year of Birth	Current Place of Residence	Place of Birth	Birthplace of Parents	Speech Data (Duration, In Minutes)
AAK-47 (M)	1947	Ryzhkovo	Syade	mother: Ryzhkovo father: no data	41
IAI-33 (F)	1933	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	33
JuMS-28 (M)	1928	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	78
KKM-34 (M)	1934	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	31.5
MAP-49 (F)	1949	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	30.5
MMM-39 (M)	1939	Ryzhkovo	Ryzhkovo	both parents: Ryzhkovo	62
PGM-56 (F)	1956	Omsk	Finy	both parents: Finy	8
SVM-29 (M)	1929	Mikhailovka	Larionovka	both parents: Yamburg-sky Uyezd	10.5

The structure of the audio data is shown in this table

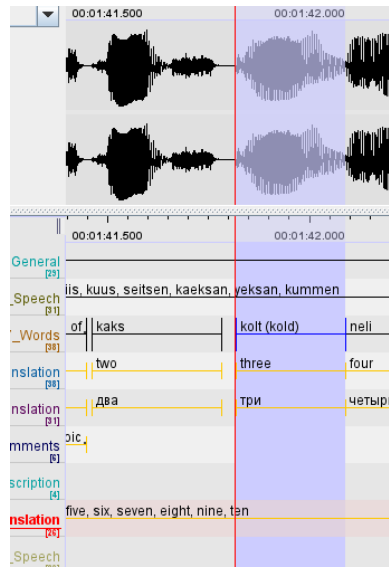
3.3 The annotations

We use ELAN media annotation tool for annotating speech data. Annotation files are stored in our project repository on GitHub:

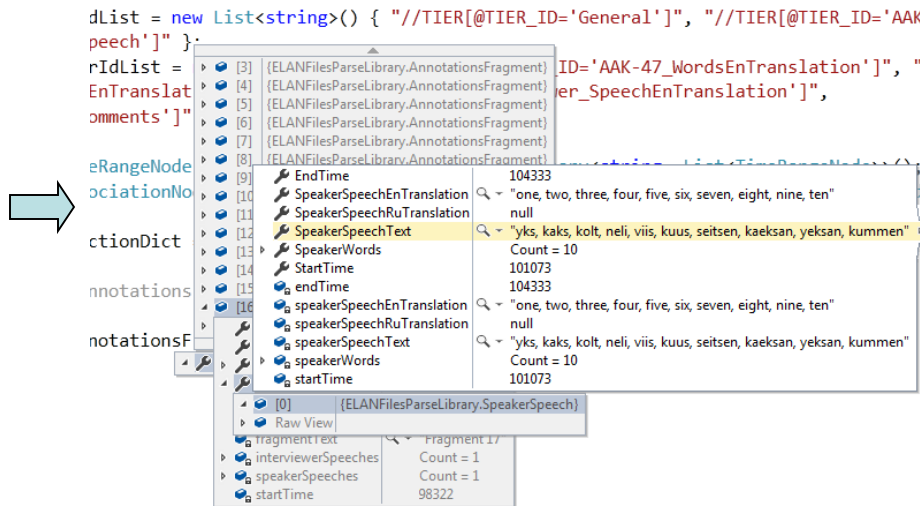
<https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/annotations>



3.4 A parser for ELAN annotation files *.eaf (XML)



Annotations and audio data



The parser forms an object tree

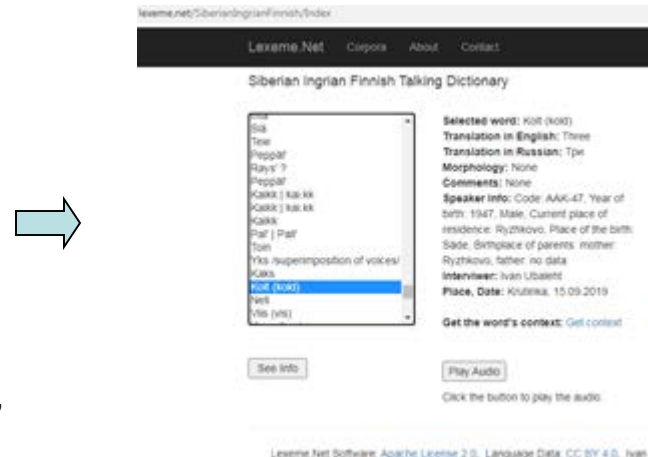
SQL-insert commands to apply to the database

The diagram shows a list of database tables on the left: `dbo.FieldWorkFile`, `dbo.GeneralFragment`, `dbo.Interviewer`, `dbo.InterviewerSpeech`, `dbo.RecContext`, `dbo.RecDevice`, `dbo.Session`, `dbo.Speaker`, `dbo.SpeakerSpeech`, and `dbo.Word`. An arrow points to a table with the following data:

id	start	end	start	end	word	translation	count	start	end
271	457	393	86176	86510	Pal' Pall'	many			
272	458	396	97398	97782	toin	other			
273	459	397	101073	101447	yks /superi...	one			
274	460	397	101460	101781	kaks	two			
275	461	397	101818	102091	kolt (kold)	three			
276	462	397	102091	102336	neli	four			
277	463	397	102336	102606	viis (vis)	five			
278	464	397	102606	102840	kuus (kus)	six			
279	465	397	102840	103141	seitsen	seven			

The start and the end of the audio segment for the player, in milliseconds. Audio files are stored in the file system.

Data in the fieldwork database (MS SQL Server)



The web application displays this data

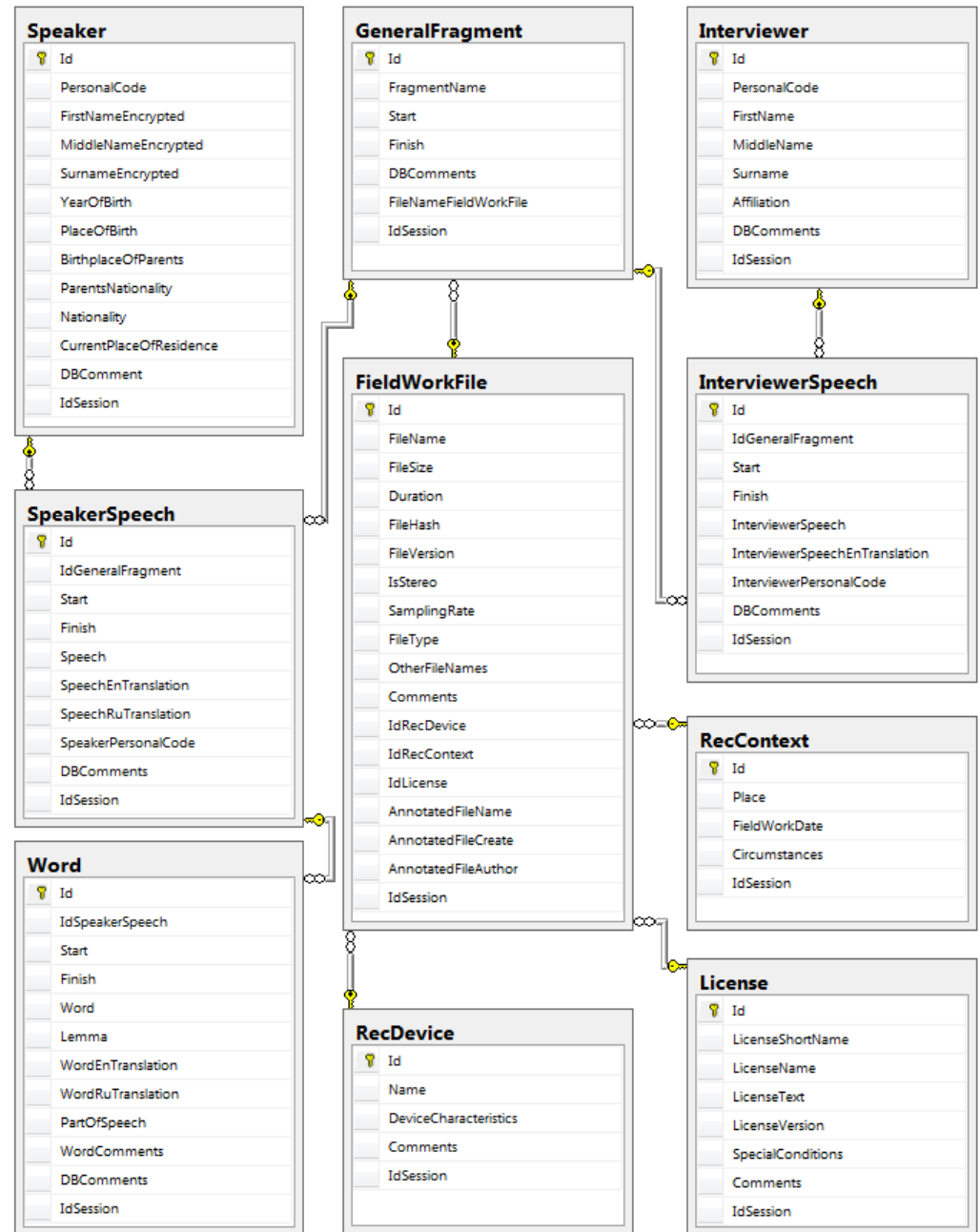
The source code of the parser is available on GitHub and licensed under the Apache 2.0 License: <https://github.com/ubaleht/LexemeELAN>

3.5 The fieldwork database for the talking dictionary

The fieldwork database stores the characteristics of the recorded fragments of speech and timestamps from annotations as well as characteristics of the speakers.

The Schema of the fieldwork database:

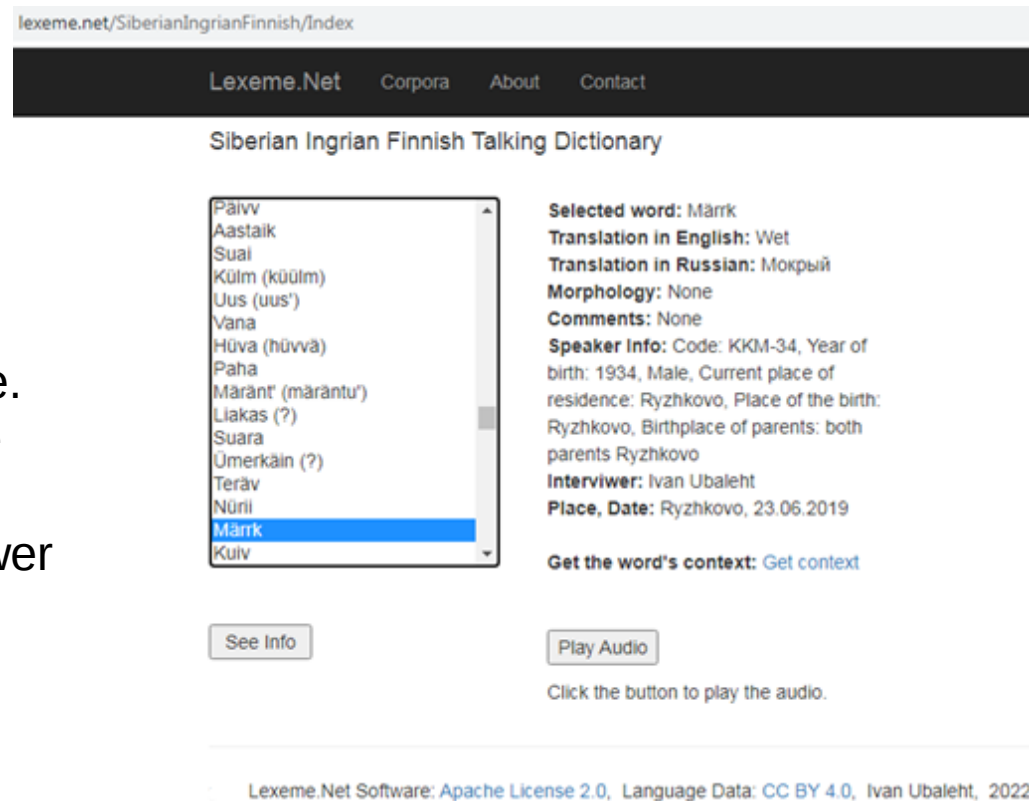
<https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/SpeechDatabase/Scheme>



3.6 The web application for the Siberian Ingrian Finnish talking dictionary

We have developed a web application that, in accordance with user requests, can display information taken from annotation files, which is stored in the database. This web application can play audio fragments according to timestamps obtained from the database. Depending on the user's request, these timestamps can correspond to such fragments as: words, phrases, interviewer questions.

The source code of the web application is open and available on GitHub: <https://github.com/ubaleht/Lexeme>
At the moment, web application is available via Internet: <http://lexeme.net/sif>



3.6 The web application for the Siberian Ingrian Finnish talking dictionary

In addition to words in the talking dictionary there are phrases:

lexeme.net/SiberianIngrianFinnish/Context?word=397

Lexeme.Net

Corpora

About

Contact

Word context

Speaker's speech: Miä kirjutan kirja, ketta en kuul

Translation into English of the speaker's speech: I am writing a letter without listening to anyone

Interviewer's speech: Я пишу письмо, никого не слушая

Translation into English of the interviewer's speech: I am writing a letter without listening to anyone

Audio file Name: MAP-49-002.WAV

Fragment start time (in seconds): 265

Fragment end time (in seconds): 277

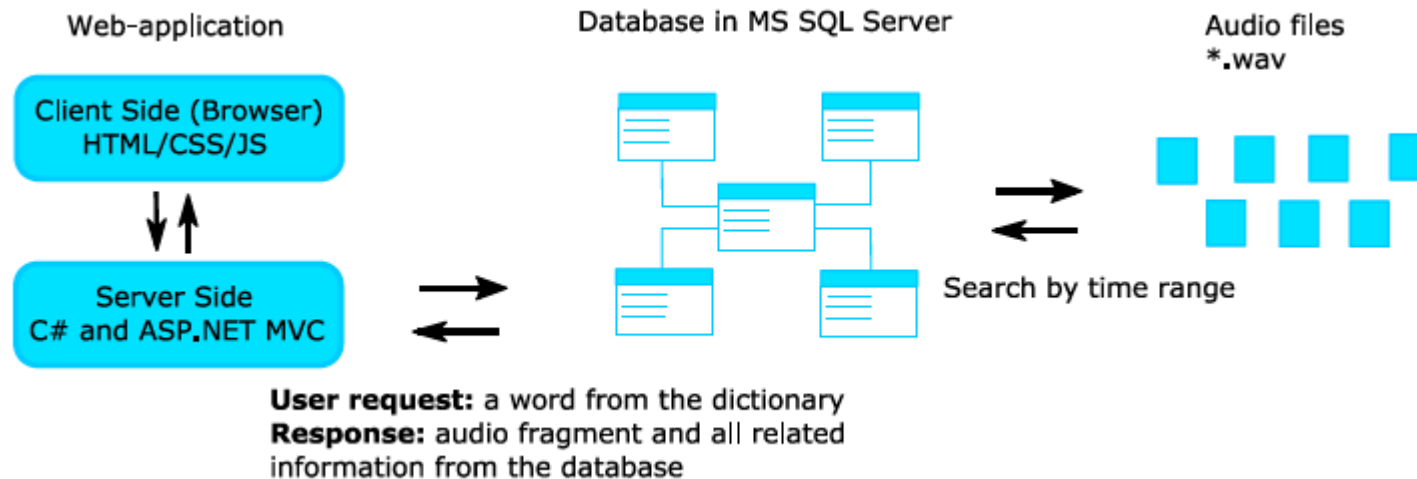
Play Audio

[Back](#)

Lexeme.Net Software: [Apache License 2.0](#), Language Data: [CC BY 4.0](#), Ivan Ubaleht, 2022

3.6 The web application for the Siberian Ingrian Finnish talking dictionary

The Siberian Ingrian Finnish talking dictionary is based on a client-server model. The following technology stack is used on the server side: .NET Framework, C#, ASP.NET MVC. The standard approach is used on a client side. A user with a help of web-browser sends requests to a server and receives a response from a server. HTML, CSS, JavaScript are used to display information.



The schema of work of the web-application for the Siberian Ingrian Finnish talking dictionary

4. Conclusion and future work

Current results:

- A software tool for parsing annotation files and feeding a database was created. This software is available on GitHub and licensed under the Apache 2.0 License.
- The structure of the fieldwork database has been developed and this database has been filled. Now this database contains information about 300 words.
- The web application had been created. The source code of the web application is open and available in GitHub. At the moment, the web application is available via Internet: www.lexeme.net/sif

We plan:

- To develop a morphological analyzer for this language.
- Since we have 2 hours of audio with synchronous text, we plan to use this data to create a training dataset for a neural network and we plan to create an automatic speech recognition (ASR) system for automatic/semi-automatic annotation of audio data for this language and maybe Lower Luga Izhorian and Kurkola dialect of Ingrian Finnish.
- We need this talking dictionary, ASR system for revitalize (as heritage language) this language in our region. At the moment I started to develop a language course for this language. People in our region have an interest in this language.



Ivan Ubaleht

Omsk State Technical University

ivan@ubaleht.com

ubaleht@gmail.com

+7 905 923 8134

GitHub: <https://github.com/ubaleht>