Development of the Siberian Ingrian Finnish Speech Corpus

Ivan Ubaleht
Omsk State Technical University
ubaleht@gmail.com

Taisto-Kalevi Raudalainen
Estonian Academic Society of Ingria
taika.rauta@gmail.com

Abstract

In this paper we present the speech corpus for the Siberian Ingrian Finnish language. The speech corpus includes: audio data, annotations, software tools for data processing, two databases and a web application. We have published part of the audio data and annotations. The software tool for parsing annotation files and feeding a relational database is developed and published under a free license. A web application is developed and available. At this moment, about 300 words and 200 phrases can be displayed using this web application.

1 Introduction

Many of endangered languages have the following specific features: (i) there are no writing system and stable orthography; (ii) there are very few available speech data and texts. These features should be considered when working on speech corpora. 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 requirements: (i) all source code and data should be accessible on GitHub and licensed under one of a free license; (ii) speech corpora should be available to users via the Internet without installing additional software; (iii) speech corpora should be convenient not only for linguists, but also for speakers of endangered languages, language activists and software developers; (iv) speech corpora should have a powerful system of requests to data (for example, search by grammatical categories, regular expression search). At present, there are solutions that meet these requirements: the "Tsakorpus" corpus platform[1] (for example, the project INEL uses the "Tsakorpus" corpus platform (Arkhangelskiy et al., 2019)), 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 chose the .NET Framework[2] and Microsoft SQL Server[3] for the implementation of the project. Siberian Ingrian Finnish was chosen as the first endangered language for the Lexeme.Net system.

We briefly review the Siberian Ingrian Finnish in section 2. We describe the design of the Siberian Ingrian Finnish speech corpus in section 3. In section 3 we describe the general structure of the speech corpus, annotation tiers, the data model of the fieldwork database and the web application.

2 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 (so-called the dialect of the Kurkola peninsula) and Lower Luga Ingrian varieties (Kuznetsova et al., 2015). They migrated from the Lower Luga area to Siberia in 1803-1804.

Siberian Ingrian Finnish (Russian: Сибирский ингерманландский идиом) is the term introduced by D.V. Sidorkevich. D.V. Sidorkevich who researched and documented Siberian Ingrian

[1] https://github.com/timarkh/tsakorpus
[2] https://dotnet.microsoft.com/
[3] https://www.microsoft.com/sql-server

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 1 - 4

Finnish (Sidorkevich, 2011; Sidorkevich, 2014) in 2008-2014. The language was also studied by N.V. Kuznetsova (Kuznetsova, 2016) and M.Z. Muslimov.

In 2022, there is still a group of people of elder generation who use Siberian Ingrian Finnish in the domestic sphere of communication in Ryzhkovo settlement (Krutinsky District of Omsk Oblast). The villagers of Ryzhkovo also use Siberian Ingrian Finnish for communication with their relatives from Estonia by phone. There is also a small group of people in Omsk who use this language occasionally. According to our estimates, about 15 native speakers of this language now live in Russia and Estonia. The number of semi-speakers is about 30-60.

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 no writing system, stable orthography and texts.

3 The design and development of the speech corpus

3.1 The general structure of the speech corpus

In this subsection, we briefly describe all components of the Siberian Ingrian Finnish speech corpus. We use ELAN media annotation tool (Wittenburg et al., 2006) to annotate speech data. The structure of the annotation files is shown in subsection 3.3.

Annotation files are XML files, therefore we have developed a software tool to read annotations (see subsection 3.3). After parsing annotations, the object tree with data from annotations is stored in a relational database. Two relational databases are part of the speech corpus (see subsection 3.4). The fieldwork database stores annotations of speech data, timestamps and the attributes of speakers, interviewers, audio files, equipment. The lexical database will store information about grammatical categories, parts of speech, synonyms of words. The lexical database will be used for the Siberian Ingrian Finnish dictionary and for the work of the morphological analyzer. Both of these databases can exchange information.

The next part of the speech corpus is the web application which allows users to play audio fragments according to timestamps from annotation files, display information according to annotations, and also will allow users to make complex queries to database. The web application also will display information about word-forms, morphology and grammatical categories.

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[4].

3.3 The annotations

We use ELAN media annotation tool for annotating speech data. Annotation files are stored in our project repository on GitHub [5] . The structure of the annotation files is shown in Figure 1. On the “Speaker-Speech” tier are the phrases spoken by the speakers of Siberian Ingrian Finnish. Tier “Speaker-Words” displays the words spoken by the speakers. Layers “Speaker-WordsEnTranslation”, “Speaker-WordsRu Translation”, “Speaker-SpeechEnTranslation” “Speaker-SpeechRuTranslation” display translations of phrases and words into English and Russian. Parts of speech and morphological aspects are described on the tiers: “Speaker-PartOfSpeach”, “Speaker-Morph”. Questions and phrases of an interviewer are annotated on tiers: “Interviewer-Speech”, “Interviewer-SpeechEnTranslation”.

ELAN file format is an XML format. We have developed a software tool (the desktop application for Windows) for parsing these XML files (*.eaf files) and transforming annotations into an object tree. Then in accordance with this object tree, our program library generates SQL-insert commands for adding these objects to the relational database. We tested this software tool by parsing an ELAN file with 10,000 lines and tested running about a

[4] https://github.com/ubaleht/SiberianIngrianFinnish

[5] https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/annotations

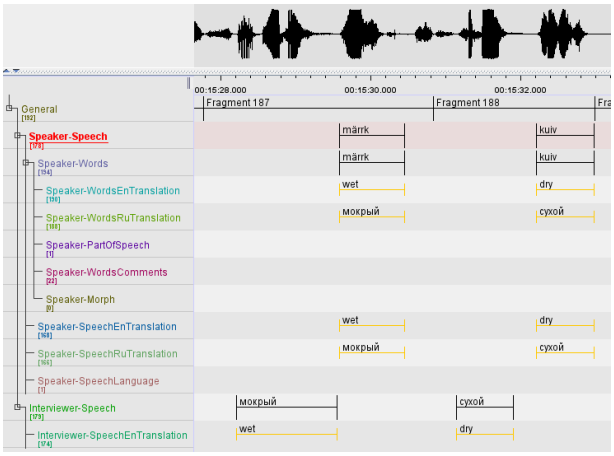

Figure 1: An example of annotated fragment of speech data of Siberian Ingrian Finnish and the list of tiers.

thousand SQL-insert commands [6] to add information from annotations to a relational database in Microsoft SQL Server. This software tool is stable and can be used for other endangered languages. This software is available on GitHub[7] and licensed under the Apache 2.0 License.

3.4 The databases for the speech corpus

The speech corpus uses two relational databases. The fieldwork database stores the characteristics of the recorded fragments of speech and timestamps from annotations as well as characteristics of the speakers. The data model[8] of this database is shown in Figure 2.

The lexical database stores the data for the Siberian Ingrian Finnish dictionary, more precisely, word-forms according to inflectional paradigms. Since the language is under-resourced, we build a part of word-forms hypothetically according to our knowledge of the grammar. We verify the data from the lexical database, using the data from the fieldwork database based on annotations. A key field for linking the two databases is the field “Lemma”. The lexical database is necessary for the work of the rule-based morphological analyzer for Siberian Ingrian Finnish.

[6] https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/SpeechDatabase/Data

[7] https://github.com/ubaleht/LexemeELAN

[8] https://github.com/ubaleht/SiberianIngrianFinnish/tree/master/SpeechDatabase/Scheme

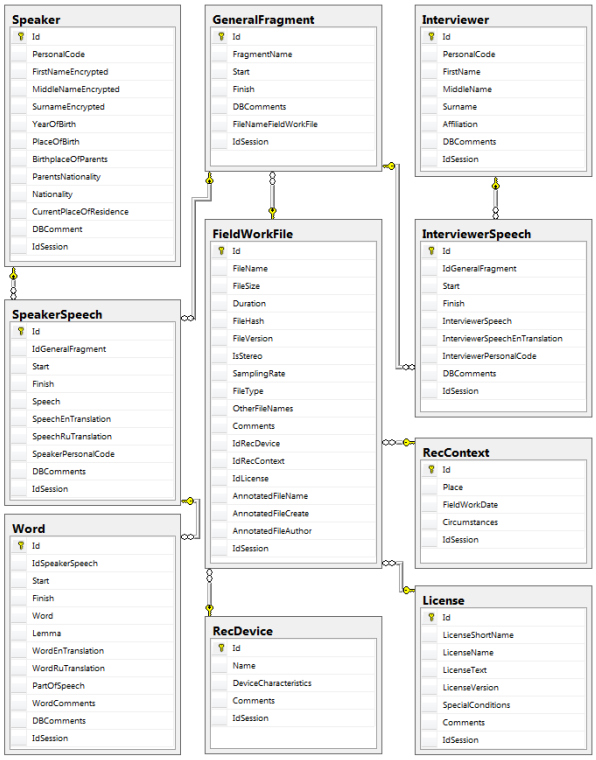

Figure 2: The data model of the fieldwork database.

3.5 The web application

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 (in order to better understand the context of words). The source code of the web application is open and available on GitHub[9]. At the moment, web application is available via Internet[10], see Figure 3.

4 The current status of the creation of the speech corpus and conclusion

At this moment, about 300 words (the number of individual pronunciations) and 200 phrases in audio files have been annotated. All these words and phrases were collected from 4 speakers of Siberian Ingrian Finnish. These words are mostly from the 200-word Swadesh list as well as the other basic lexicon. These 300 words and 200

[9] https://github.com/ubaleht/Lexeme

[10] http://lexeme.net/sif

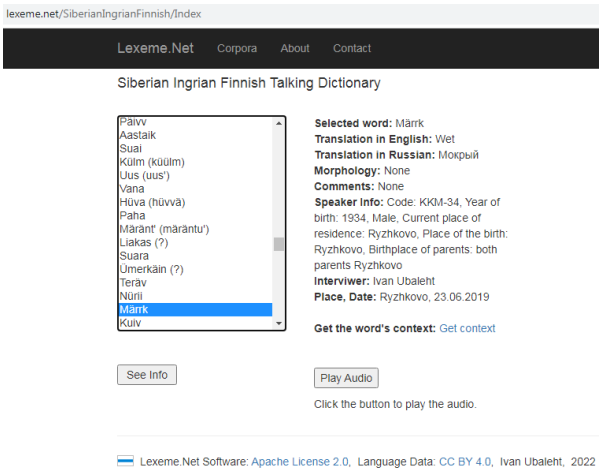

Figure 3: The first version of the web application for the Siberian Ingrian Finnish speech corpus.

phrases can be played in our web-application, and the web-application also displays information from the annotations associated with these audio fragments.

The following results have been achieved:

- Audio data of the Siberian Ingrian Finnish language has been published and licensed under a Creative Commons Attribution 4.0 license (CC BY 4.0).
- The annotations of audio data have been published.
- A software tool for parsing annotation files and feeding a database was created.
- The structure of the fieldwork database has been developed and this database has been filled. Now this database contains information about 300 words and 200 phrases.
- 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.
- The rule-based morphological analyzer and the lexical database of Siberian Ingrian Finnish is under development.

After creating the speech corpus for Siberian Ingrian Finnish, we plan to start creating a speech corpus for the Siberian Tatar language using the software described above.

References

Timofey Arkhangelskiy, Anne Ferger, and Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*. Association for Computational Linguistics, pages 115-124. https://doi.org/10.18653/v1/W19-0310.

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.

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.

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.

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

Natalia Kuznetsova, Elena Markus, and Mehmet Muslimov. 2015. Finnic minorities of Ingria. *Cultural and linguistic minorities in the Russian Federation and the European Union*, 13: 127-167. https://doi.org/10.1007/978-3-319-10455-3_6.

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. https://doi.org/10.3176/lu.2016.1.01.

Daria V. Sidorkevich. 2014. *Yazyk ingermanlandskih pereselentsev v Sibiri.* Diss. ILIRAN.

Daria V. Sidorkevich. 2011. On domains of adessive-allative in Siberian Ingrian Finnish. In *Proceedings of Institute for Linguistic Studies* 7(3): 575-607.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alexander Klassmann, and Han Sloetjes. 2006. ELAN: A Professional Framework for Multimodality Research. In *Proceedings of Language Resource and Evaluation 2006*, pages 1556–1559.