The Creation of Siberian Ingrian Finnish and Siberian Tatar Speech Corpora

Ivan Ubaleht
Department of Automated Systems for Information Processing and Management, Omsk State Technical University, Omsk, Russia
ubaleht@gmail.com

Abstract

In this paper we present the first results of the creation of speech corpora for Siberian Ingrian Finnish and Siberian Tatar. The speech data of these languages were published, are accessible to the public, and are licensed under a Creative Commons Attribution 4.0 license. These corpora will be used to the Lexeme system. Lexeme – is new application for managing speech corpora for under-resourced languages. Now, the Lexeme system is under development.

1 Introduction

At present, there are enough software solutions which allow working with speech corpora. There are following stand-alone applications: IrcamCorpusTools (Veaux and Beller, 2008), EXMARaLDA (Schmidt and Wörner, 2009), LaBB-CAT (Fromont and Hay, 2012) and newer systems such as SPPAS (Bigi 2015). The following modern software solutions are based on a client-server model: the EMU Speech Database Management System (Winkelmann et al., 2017) and ISCAN (McAuliffe, et al., 2019). For under-resourced languages LingSync and the Online Linguistic Database (Dunham et al., 2014) can be note.

2 The Lexeme System

Lexeme is a new system provides following features: the storage of audio data, data processing, representing of speech information to users. This system will have special features for the documentation and revitalization of endangered languages. Lexeme is based on the following key principles:

- Openness and transparency (all code and data (including primary audio data) will be accessible on GitHub and licensed under CC BY)
- Universality (the system will consist of independent levels, users can use artifacts irrespective to level for own projects)
- Targeted at different users (linguists, computational linguists, speakers of endangered languages, language activists).

Nowadays, the Lexeme system is under development. We describe current status of the creation of two first corpora for the Lexeme system in Section 3.

3 Speech Corpora for the Lexeme System

3.1 Speech Corpus of Siberian Ingrian Finnish

Language context: Siberian Ingrian Finnish – is a language (dialect) based on the Lower Luga Ingrian Finnish and Lower Luga Ingrian (Izhorian) varieties (Kuznetsova et al., 2015). This language used by the descendants of the settlers from the Lower Luga area. The ancestors of the speakers of Siberian Ingrian Finnish came from the Lower Luga area in the early 19th century. They came from the Rosona river area, to be exact. This region is also called Estonian Ingria. They have been living in Omsk Oblast (Russia) for more than 200 years (previously they lived also in other regions of the Siberia).

Siberian Ingrian Finnish (Russian: Сибирский ингерманландский идиом) is the term introduced by D. V. Sidorkevich. D. V. Sidorkevich with the participation of M. Z. Muslimov and N. V. Kuznetsova from the Institute for Linguistic Studies of the Russian Academy of Sciences researched and documented Siberian Ingrian Finnish in 2008-2014 (Sidorkevich, 2011; Sidorkevich, 2013). Several expeditions were undertaken to Omsk oblast (Ryzhkovo and Mikhailovka settlements) in 2008-2011.

The current status of the creation of Siberian Ingrian Finnish Speech Corpus: For the first time a speech data of the Siberian Ingrian Finnish language has been published and are accessible to the public[1]. These speech data are available on GitHub and licensed under CC BY 4.0. Currently, has been published larger part of the audio data from our expeditions. At present, we are creating metadata describing speakers and characteristics of speech audio files then we'll annotate this speech data. Siberian Ingrian Finnish speech corpus will be created based on this audio data and new speech audio data. We recorded 10 hours of audio from 8 speakers from four our expeditions and from the interviews via phone in 2019-2020. Approximately 5 hours of the audio data were published on GitHub. The structure these primary audio data is shown in Table 1.

The Code of the Speaker and Gender	The Year of Birth	Speech Data (Duration, In Minutes)
AAK-47 (M)	1947	12
IAI-33 (F)	1933	75
JuMS-28 (M)	1928	54
KKM-34 (M)	1934	31,5
MAP-49 (F)	1949	30,5
MMM-39 (M)	1939	43
PGM-56 (F)	1956	11
SVM-29 (M)	1929	33

Table 1: Speech data of Siberian Ingrian Finnish: breakdown by speaker.

3.2 Speech Corpus of Siberian Tatar

Language context: The language of Siberian Tatars is a Turkic language. This language is relatively well-studied, around 100,000 people are spoken in this language but nonetheless this language is under-resourced language. The Siberian Tatar language was given the code "sty" (ISO 639-3) by ISO in 2012. The language of Siberian Tatars has three dialects: Tobol-Irtysh, Tom and Baraba. Tobol-Irtysh dialect of the language of Siberian Tatars consists of following subdialects: Tyumen, Tobol, Zabolotny, Tevriz and Tara. The speech data of our first expedition were recorded in Tevriz subdialect area.

The current status of the creation of Siberian Tatar Speech Corpus: The our first expedition was undertaken to Siberian Tatar village Ilchebaga (Ust'-Ishimsky District, Omsk Oblast, Siberia, Russia) in 2020. We recorded the speech data of 9 speakers in this first expedition. These primary speech data already were published and are accessible to the public[2]. These speech data are available on GitHub and licensed under CC BY 4.0. The structure these primary audio data is shown in Table 2. We started creating the Siberian Tatar speech corpus based on this data. We plan to collect speech material of all dialects and accents of Siberian Tatars for this speech corpus.

The Code of the Speaker and Gender	The Year of Birth	Speech Data (Duration, In Minutes)
AVN-69 (M)	1969	2,5
GMG-67 (M)	1967	2,5
GNSh-29 (F)	1929	24,5
KMM-63 (M)	1963	49
MKhU-50 (F)	1950	32
MRCh-60 (M)	1960	34
NGA-45 (F)	1945	12
NIA-53 (M)	1953	9
SGL-61 (M)	1961	5

Table 2: Speech data of Siberian Tatar: breakdown by speaker.

4 Conclusion

In this paper, we have presented our current results of the creation of speech corpora for Siberian Ingrian Finnish and Siberian Tatar. The speech data of these languages were published and are accessible to the public. Furthermore, we briefly reviewed key principles of the Lexeme system.

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

[2] https://github.com/ubaleht/SiberianTatar

References

Brigitte Bigi. 2015. SPPAS-multi-lingual approaches to the automatic annotation of speech. *The Phonetician* 111(112): 54-69.

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.

Robert Fromont, and Jennifer Hay. 2012. LaBB-CAT: An annotation store. In Proceedings of the Australasian Language Technology Association Workshop, pages 113-117.

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.

Michael McAuliffe, et al. 2019. ISCAN: A system for integrated phonetic analyses across speech corpora. Pages 1322-1326.

Thomas Schmidt, and Kai Wörner. 2009. EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19(4): 565-582.

Daria V. Sidorkevich. 2013. *Yazyk ingermanlandskih pereselentsev v Sibiri*. ILIRAN. (In Russian)

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

Christophe Veaux, Grégory Beller, and Xavier Rodet. 2008. *IrcamCorpusTools: an extensible platform for speech corpora exploitation*.

Raphael Winkelmann, Jonathan Harrington, and Klaus Jänsch. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, 45: 392-410.