Lexeme: the Concept of System and the Creation of Speech Corpora for Two Endangered Languages

Ivan Ubaleht
Omsk State Technical University, Russia
ivan@ubaleht.com

Abstract

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

1 Introduction

At present, there are enough software solutions which allow working with speech corpora. There are the 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). The special linguistic tools, such as FieldWorks[1] from SIL International, allow users to document endangered languages. However, we see the need to use tools with special features for working with endangered languages. There are relatively few such tools, for example LingSync and the Online Linguistic Database (Dunham, 2014; Dunham et al., 2014). The need to develop new solutions in this area remains an important challenge. We briefly review our solution in Section 2. We describe current status of the creation of two first corpora of endangered languages for the Lexeme system in Section 3.

2 The Lexeme System

2.1 The Principles of 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 source code and data (including primary audio data) will be accessible on GitHub and licensed under one of a free license)
- Universality (the system will consist of independent levels, users can use artifacts irrespective of other levels for their own projects)
- Targeted at different users (linguists, computational linguists, speakers of endangered languages, language activists).

At this moment, the Lexeme system is under development.

2.2 The concept of the Lexeme system

The lower level of the Lexeme system: The lower level would bring together all collected primary data from speakers of endangered languages (except speech data that violate the ethical principles). These primary speech data will be accessible to the public under a Creative Commons Attribution 4.0 license (CC BY 4.0).

[1] https://software.sil.org/fieldworks/

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Resource Papers and Extended Abstracts, pages 20–23,
Online, March 2–3, 2021.

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[2]	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[2]	both parents: Finy[2]	8
SVM-29 (M)	1929	Mikhailovka	Larionovka[2]	both parents: Yamburg-sky Uyezd	10.5

Table 1: The speech data of the Siberian Ingrian Finnish language.

The examples of the lower level of the Lexeme system can be the speech data for Siberian Ingrian Finnish and Siberian Tatar corpora (see Section 3).

The middle level of the Lexeme system: Documentation and annotation of speech data are being conducted on this level. We use ELAN (Wittenburg et al., 2006) for annotating speech data and special methods for bridging "annotation bottleneck" on this level. There are annotations (for example, the annotations[3] in our repository of the corpus of Siberian Ingrian Finnish), schemes of databases, structured data for databases, scripts for conversion to different formats on this level. Users can use annotated speech data freely in own projects not only in the Lexeme system.

We plan to use crowdsourcing for annotation of speech data (work collaboratively with linguists, speakers of endangered languages and language activists). All data in this level will be licensed under a free license too.

The upper level of the Lexeme system: The upper level is the level of applications and services. All language recourses will be accessible through user-friendly applications and services on this level. A powerful system of requests to data and a user-friendly interface for representation of data will be implemented in this level. The Lexeme application will be available via Internet using the lexeme.net domain name. The source code of these applications and services will be accessible on the GitHub under a free license.

[2] These settlements used to exist in Omsk Oblast.

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

3 The Speech Corpora for the Lexeme System

3.1 The 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 is 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 Siberia).

Siberian Ingrian Finnish (Russian: Сибирский ингерманландский идиом) is the term introduced by D. Sidorkevich. D. Sidorkevich with the participation of M. Muslimov and N. 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, 2014; Kuznetsova, 2016). Several expeditions were undertaken to Omsk oblast (Ryzhkovo and Mikhailovka settlements) in 2008-2011. Ph.D. thesis was written by D. V. Sidorkevich in 2013 (Sidorkevich, 2014). The Siberian Ingrian Finnish phonology, morphology as well as certain other aspects were described in detail in this Ph.D. thesis.

In 2020, there is still a group of people of elder generation who use Siberian Ingrian Finnish in

Code of Speaker and Gender	Year of Birth	Current Place of Residence	Place of Birth	Birthplace of Parents or/and Ethnic of Parents	Speech Data (Duration, In Minutes)
AVN-69 (M)	1969	Ilchebaga	Ilchebaga	both parents: the Siberian Tatars	2.5
GMG-67 (M)	1967	Ilchebaga	Ilchebaga	both parents: the Volga Tatars	2.5
GNSh-29 (F)	1929	Ilchebaga	Ilchebaga	three grandparents: the Volga Tatars, one grandparent: the Siberian Tatars	24.5
KMM-63 (M)	1963	Ilchebaga	Tavinsk[4]	both parents: Tavinsk[4] (the Siberian Tatars)	49
MKhU-50 (F)	1950	Ilchebaga	no data	both: the Siberian Tatars, father: Tavinsk[4], mother: Tebendya[4],	32
MRCh-60 (M)	1960	Ilchebaga	no data	three grandparents: Erbagul[4], one grandparent: Ilchebaga	34
NGA-45 (F)	1945	Ilchebaga	Yarkovo[4]	father: the Volga Tatars, mother: the Siberian Tatars	12
NIA-53 (M)	1953	Ust'-Ishim	Kuchum[5]	father: Kuchum[5] (the Siberian Tatars), mother: the Volga Tatars	9
SGL-61 (M)	1961	Ilchebaga	no data	no data	5
Anonym Speaker (M)	no data	Ilchebaga	no data	mother: the Siberian Tatars, father: the Bukharian Tatars	4

Table 2: The speech data of the Siberian Tatar language.

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 and Oglukhino settlement in Omsk Oblast by phone.

The current status of the creation of the Siberian Ingrian Finnish Speech Corpus: For the first time speech data of the Siberian Ingrian Finnish language has been published and are accessible to the public. These speech data are available on GitHub and licensed under a Creative Commons Attribution 4.0 license (CC BY 4.0). Currently, the larger part of the audio data from our expeditions has been published[6]. We recorded 10 hours of audio from 8 speakers from four our expeditions to Ryzhkovo and Mikhailovka settlements and from the interviews via phone in 2019-2020. Approximately 5 hours of the audio data were published on GitHub. The structure of the primary audio data is shown in Table 1. At present, we are annotating speech data from our expeditions and creating a database for storing structured data. The database and structured data are essential to the work of the upper level of the Lexem system (web-application and services).

3.2 The 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 an endangered language. The Siberian Tatar language was given the code "sty" (ISO 639-3) by ISO in 2012. The language of the Siberian Tatars has three dialects: Tobol-Irtysh, Tom and Baraba. The Tobol-Irtysh dialect of the language of the Siberian Tatars consists of the following subdialects: Tyumen, Tobol, Zabolotny, Tevriz and Tara. The speech data of our first expedition were recorded in a Tevriz subdialect area.

The current status of the creation of the Siberian Tatar Speech Corpus: 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 10 speakers in this first expedition. These primary speech data have already been published and are accessible to the public[7]. These

[4] These villages are located in Omsk Oblast.

[5] The village used to exist in Omsk Oblast.

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

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

speech data are available on GitHub and licensed under a Creative Commons Attribution 4.0 license (CC BY 4.0). The amount of the primary audio data and characteristics of the speakers are shown in Table 2. We started creating the Siberian Tatar speech corpus based on this data. We plan to collect speech material of all the dialects and the accents of Siberian Tatars for this speech corpus. In 2020, we couldn't record more speech data because of the coronavirus COVID-19 pandemic.

4 Conclusion

In this paper, we have presented our current results of the creation of the speech corpora for Siberian Ingrian Finnish and Siberian Tatar. These languages are endangered languages. For the first time the speech data of these languages were published and are accessible to the public. Furthermore, we briefly reviewed key principles and concept of the Lexeme system. Lexeme is a new application for managing speech corpora for endangered languages.

References

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

Joel Dunham. 2014. *The Online Linguistic Database: software for linguistic fieldwork.* Diss. University of British Columbia.

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.

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.

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. 2014. *Yazyk ingermanlandskih pereselentsev v Sibiri*. Diss. 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.

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 1557–1559.