LIINA LINDSTRÖM, PÄRTEL LIPPUS &
TUULI TUISK
University of Tartu

The online database of the University of Tartu Archives of Estonian Dialects and Kindred Languages and the Corpus of Estonian Dialects

1. Introduction

This paper introduces the online database of the University of Tartu Archives of Estonian Dialects and Kindred Languages (AEDKL; in Estonian, Tartu Ülikooli eesti murrete ja sugulaskeelte arhiiv), which is freely accessible and open to researchers at ‹https://murdearhiiv.ut.ee/›, and as an independent part of the archives, also the Corpus of Estonian Dialects (CED; in Estonian, Eesti murrete korpus, ‹https://www.keel.ut.ee/et/keelekogud/murdekorpus›). Both sources have been developed at the University of Tartu. These sources are related, as the recordings and transcribed texts of the CED are held in the archives, while the materials of the AEDKL are used in the CED. The first half of this paper introduces the history and materials held in the AEDKL and how they can be used online. The second half of the paper gives an overview of the dialect corpus. This paper describes the state of the archives as of December 2019.

***Multilingual Finnic**. Language contact and change*. 327–350.
Uralica Helsingiensia 14. Helsinki 2019.
‹https://doi.org/10.33341/uh.85040›

2. The Archives of Estonian Dialects and Kindred Languages (AEDKL)

The archives are a collection of Finno-Ugric linguistic materials and are located at the University of Tartu. The collection contains four types of materials: 1) sound recordings of Estonian dialects and other Uralic languages; 2) unpublished manuscripts, including student coursework and theses defended at the Institute of Estonian and General Linguistics, fieldwork diaries, transcriptions and written notes on Estonian and Finno-Ugric languages; 3) photos from fieldwork expeditions and linguistic events; 4) video recordings.

The organization of the archives began in 2000 with the digitization of the sound recordings of Estonian dialects and the creation of the Corpus of Estonian Dialects. The work then moved on to digitizing the recordings of other Finno-Ugric languages, scanning the written materials, and organizing the metadata into an online database. By now, most of the materials are digitized and accessible through the online database. The majority of the recordings are of Estonian dialects but another significant portion contains recordings of other Uralic languages.

2.1. History and background

The history of the archives dates back to the post-World War I period early in the history of the Republic of Estonia. In 1920, a year after the University of Tartu was opened as an Estonian language university, the Mother Tongue Society (*Emakeele Selts*) was established. The Society initiated systematic work in collecting Estonian dialect materials and conducting research into these dialects (Rätsep 2003, Erelt 2010). The collection containing mostly manuscripts but also sound recordings on wax cylinders was organized as the Archives of the Estonian Language and Kindred Languages and due to the fact that the Mother Tongue Society was related to the University of Tartu, the archives were stored in the rooms of the university phonetics lab. Near the end of World War II when Tartu was heavily bombed, most of the archives were evacuated, but the collection of wax cylinders was destroyed in a fire (Ahven 1955). After the war the manuscripts were returned to Tartu.

In 1947, the Institute of Language and Literature at the Estonian SSR Academy of Sciences (*Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut*[1]) was established. The Mother Tongue Society along with its archives was taken over by the Estonian SSR Academy of Sciences (see EF 1997: 10, Rätsep 2003: 169). This was part of a policy to organize research into institutes under the umbrella of the Academy of Sciences with the university instead focusing on teaching. At first, the two institutions were both located in Tartu and used the common archives, but in 1952 the institute together with the archives moved to Tallinn. In order to carry on with research in Tartu, the university had to build a new collection of linguistic data. The basis of the current archives is formed from the seminar papers and MA theses of the 1920s and 1930s which the institute returned to the university. Step by step more materials were added with students making handwritten copies of manuscripts of the Mother Tongue Society collection as well as through the process of collecting new data on annual fieldwork expeditions. A new era dawned in 1957, when the first battery-powered tape recorder was obtained and the first sound recordings were made. Since then, fieldwork expeditions have been organized every summer (Kingisepp 1967). Therefore, the amount of materials held in the archives increases every year.

2.2. Sound recordings

The archives consist of about 2 800 hours of sound recordings. The majority of the recordings are of Estonian dialects with the remainder composed of recordings of other Finnic languages (Livonian, Votic, Ingrian, Veps, Karelian, Olonets Karelian, Lude, and Ingrian Finnish) and Finno-Ugric (Inari Saami, Erzya, Moksha, Komi, Udmurt, Khanty, Hungarian) and Samoyedic (Kamas, Nenets) languages (see Table 1).

1. Beginning in 1993: *Eesti Keele Instituut*, the Institute of the Estonian Language.

Family	Language	Total duration (h:mm:ss)
Finnic	Estonian	1 866:00:12
Finnic	Ingrian	390:18:37
Finnic	Votic	192:26:58
Finnic	Livonian	145:25:27
Finnic	Ingrian Finnish	64:37:21
Finnic	Veps	103:27:21
Finnic	Karelian	34:44:02
Finnic	Olonets Karelian	10:12:28
Finnic	Lude	36:41:30
Finno-Ugric	Khanty	8:11:24
Finno-Ugric	Inari Saami	7:16:23
Finno-Ugric	Erzya	6:16:04
Finno-Ugric	Udmurt	5:17:03
Finno-Ugric	Komi	3:39:52
Finno-Ugric	Hungarian	2:28:44
Finno-Ugric	Moksha	0:24:22
Samoyedic	Kamas	11:00:38
Samoyedic	Nenets	0:16:23

Table 1. Uralic languages represented in the archives.

Depending on the period to which the recordings date, they are found on reel-to-reel tapes, cassettes, or digital media. The first reel-to-reel tape recordings date back to 1959; between 1980 and 2000 most of the recordings are on cassettes, and digital recordings have been made from 2000. Most of the reel-to-reel tapes and cassettes have been digitized. If undocumented fieldwork materials from earlier periods are donated to the archives, these recordings are digitized and added to the database. The main series that contain sound recordings are presented in Table 2.

The sound recordings contain mainly interviews on different subjects, such as biographic facts, descriptions of ethnographic household labor, customs, everyday life, traditional events, etc. There are also linguistic questionnaires (e.g., phonetic questionnaires, different word lists), folk songs and tales, lectures and seminars held at the university.

Series	Content	Total duration (h)
F	reel-to-reel tape and cassette recordings of Estonian dialects (this series also contains some recordings of other Finno-Ugric languages)	1 044
SU	reel-to-reel tape and cassette recordings of other Finno-Ugric languages	133
DS	digital sound recordings	695
EMH	copies of recordings of Estonian dialects from the Institute of the Estonian Language	396
SUHK	Finno-Ugric languages from the Institute of the Estonian Language	77
IHF	collection of Ingrian sound recordings (see Rozhanskiy & Markus 2019 in this volume)	480

Table 2. Series of sound recordings in the AEDKL.

In the online database the metadata include information about the participants, recording time and place, content of the recording, and also technical details, e.g., the type of recording equipment and resolution of digital files.

2.3. Manuscripts

There are a total of 396 000 pages of written manuscripts in the archives (~ 270 000 pages are digitally available). Written materials include student coursework and theses defended at the Institute of Estonian and General Linguistics, fieldwork diaries, transcriptions and written notes on Estonian and other Finno-Ugric languages. The earliest written material (a description of the Coastal dialect) dates back to 1910. There are 16 series of materials in the archive (see Table 3).

Some of the manuscripts were lost during the fire in the university main building in 1965. The most well-preserved manuscripts are the student papers on Estonian dialects and other Finno-Ugric languages. There are BA and MA theses on the Estonian language dating from 1946 or later and other Finno-Ugric languages dating from 1956 or later defended at the Institute of Estonian and General Linguistics of the University of Tartu (~ 215 000 pages).

Series	Content
C	various topics (including professors' lecture notes and teaching materials, phonetic data: palatograms and sonograms, etc.)
D	BA and MA theses on Estonian language defended at the Institute of Estonian and General Linguistics at the University of Tartu (since 1946)
H	student papers on sound systems of Estonian dialects
K	fieldwork questionnaires for collecting data on Estonian dialects
L	student papers on Estonian lexicology
LFS	Livonian folklore collection by Oskar Loorits (copies from the Estonian Literary Museum)
M	student papers on Estonian morphology
MKT	phonetic transcriptions of texts used in the Corpus of Estonian Dialects
MT	Mihkel Toomse manuscript of Estonian dialects
P	fieldwork diaries from Estonian dialect expeditions
S	student papers on various topics, mainly from the 1920s and 1930s
SUD	BA and MA theses on Finno-Ugric languages defended at the University of Tartu Institute of Estonian and General Linguistics
SUKD	seminar papers and theses on Finno-Ugric languages from 1956–2011
SUPP	fieldwork diaries from Finno-Ugric language expeditions
T	transcriptions of Estonian dialects
Tx	dialectology exams
Y	student papers describing particular Estonian dialect areas

Table 3. Written materials found in the archives.

2.4. Photos and videos

The collection holds about 3 000 photos from fieldwork expeditions and linguistic events (e.g., conferences and seminars). Photos are divided into two series based on media type: paper and digital photos. There are around 1 300 paper photos that are digitized, and digitization is still in progress. Around 1 700 digital photos are from recent years of fieldwork and different linguistic events.

Video recordings are from fieldwork conducted during recent years. Also, old film rolls from the 1970s and 1980s have been

digitized (these include recordings made during fieldwork and at various university events). There are 73 hours of video recordings and in the new version of the database viewing these videos is integrated into the online archive system (see Section 2.5 for update information).

2.5. Using the AEDKL

The online user interface of the AEDKL database was launched in 2012. It is available at ‹https://murdearhiiv.ut.ee/›. The metadata are arranged into a relational MySQL database, keeping the speaker information and the different media information in separate cross-linked tables. In this way, even if there is more than one recording and/or text transcription from the same consultant, then there is only a single database entry containing the consultant's personal information (name, date of birth, etc.) linked to all the recording and manuscript entries where (s)he is participating.

In 2019, updating of the online database was finished and the user interface was renewed. The whole database was relocated and the software was upgraded. As a result, viewing the videos is integrated on the online archive system, searching results appear on the map, sound files are presented with HTML5 player, etc.

The user should note that only the user interface has an English translation, which means that only the field names are translated. The database is monolingual with most entries in Estonian. Instructions for searching in the database and a small Estonian-English dictionary are presented on the homepage (https://murdearhiiv.ut.ee/abi.php?t=otsi).

2.5.1. Simple search

A simple search in the database is carried out on all fields of the database. The results give all database entries containing the search term within any field. For example, if one searches for "kala" ('fish'), one gets over 900 matches (see Figure 1), where the search term can either be found within the topic or the name of the speaker. Of course, this search also returns irrelevant matches, for example where a field contains the word "foneetikalabor" ('phonetics lab'). In this case it would be better to use the detailed search option.

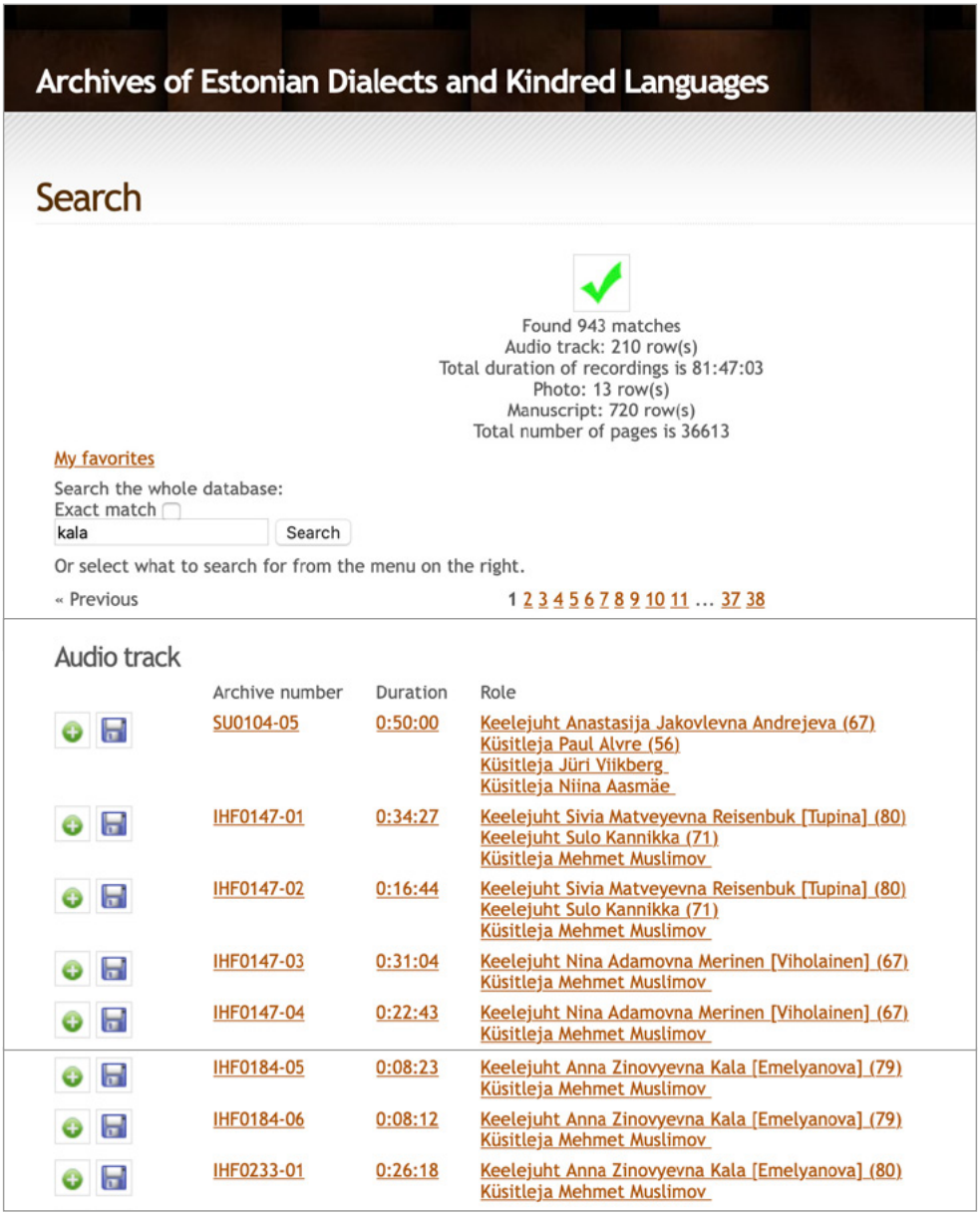


Figure 1. Example of results of a simple search with the search term "kala" ('fish').

If the search term consists of multiple words, it is split by its spaces and all individual words are searched within all fields of the database separately. For example, if one searches for all the work involving Prof. Paul Ariste, one simply searches for "Paul Ariste". This will give over 400 matches where either there was a field containing the string "Paul Ariste" or there was one field containing "Paul" and another containing "Ariste".

2.5.2. Detailed search

In case the simple search gives too many irrelevant results, the more detailed search provides another option. First of all, one has to select whether to search within audio or video recordings, photos, or manuscripts. Secondly, it is necessary to know a little about the database structure to select the database fields to search. In the detailed search option, the search is carried out only on the specified fields. For example, if one wants to find all the recordings of Votic made by Prof. Ariste in the year 1975, one has to 1) select "Search audio track" and enter 2) Language: "vadja" 3) Recording time: "1975" to "1975", 4) First name: "Paul", 5) Last name: "Ariste" (see Figure 2).

2.5.3. User access

The AEDKL user interface has two types of database users (see Table 4). An anonymous user without a user account has limited access to the database while the authorized users have full access to the archives. Information for obtaining a user account can be found at the AEDKL website ‹https://murdearhiiv.ut.ee/abi.php?t=otsi#parool›.

Anonymous user	Authorized user
Limited access	Can see all database entries
Can listen to recordings online	Can download .wav files
Can see reduced images	Can download full size images
Can see the name and village of the consultants	Can see the full personal info of the consultant

Table 4. AEDKL user rights.

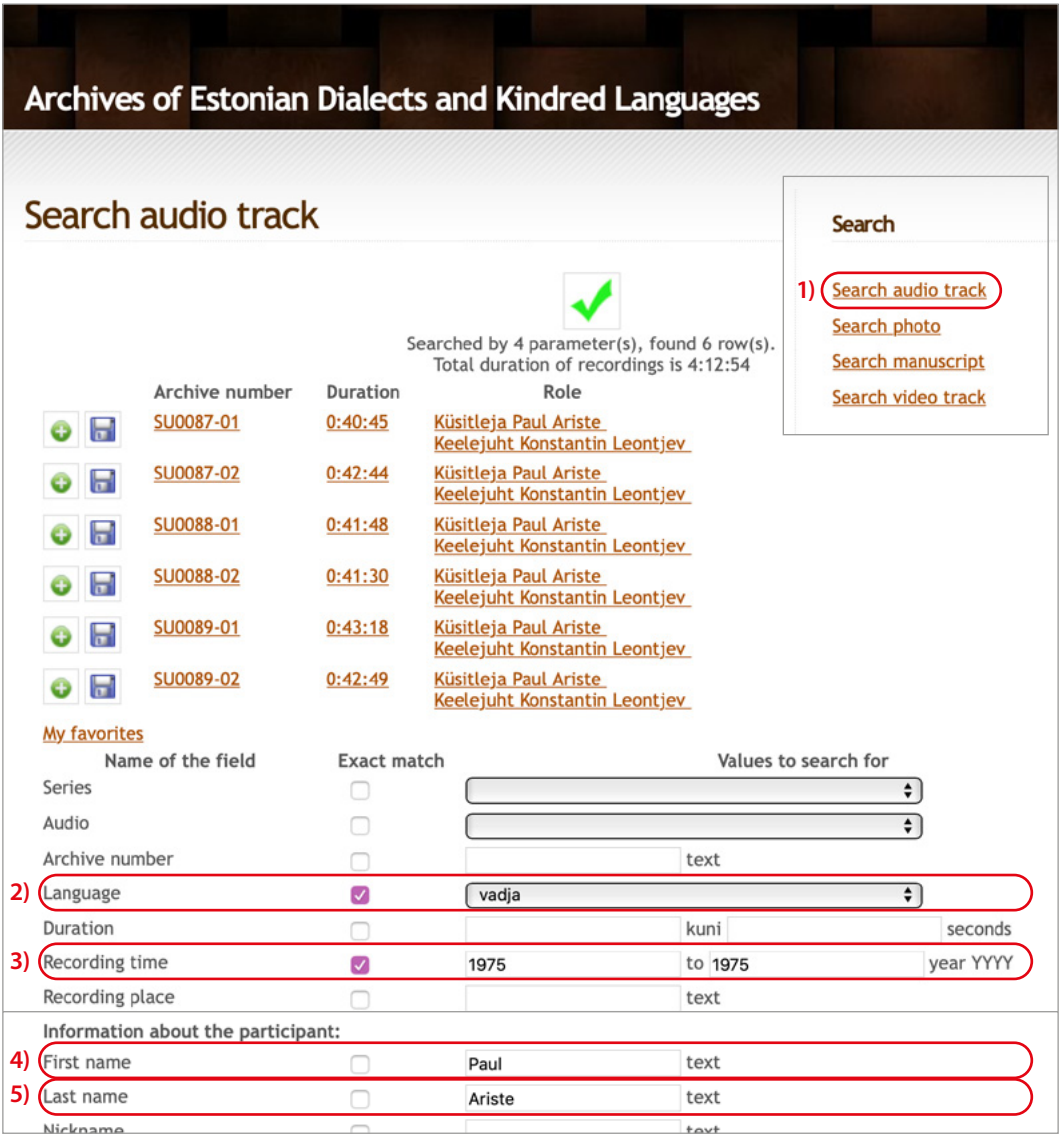


Figure 2. Example of a detailed search within audio recordings of Votic recorded by Paul Ariste in 1975. The numbers 1–5 illustrate the steps explained in the text.

3. The Corpus of Estonian Dialects

The Corpus of Estonian Dialects (CED) is a collection of electronic data containing authentic dialect texts from all Estonian dialects. This project was initiated in 1998 by the University of Tartu in cooperation with the Institute of the Estonian Language. Its main aim is to provide access to a carefully chosen collection of accurately transcribed dialect materials. These materials allow one to compare the phonological and grammatical features of Estonian dialects.

An approximately equal amount of data is provided for each Estonian dialect in the corpus; in addition to Estonian dialects, data from Votic and Livonian also are included.

The CED consists of

- sound recordings,
- transcribed texts utilizing Finno-Ugric phonetic transcription,
- dialect texts in simplified transcription,
- morphologically annotated texts,
- syntactically parsed texts,
- a database containing information about consultants and recordings.

Figure 3 illustrates the workflow of the CED: first, the sound recordings are phonetically transcribed, later these texts are converted to a simplified transcription and are morphologically annotated. After the morphological annotation, they can be syntactically parsed. During this process, relevant information about speakers and recordings is added to the metadata database.

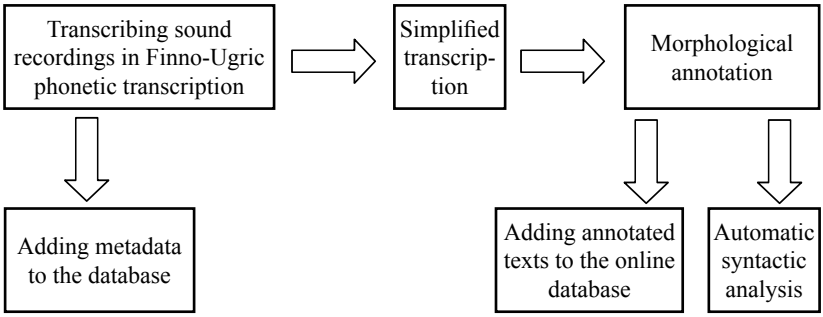


Figure 3. Workflow of the Corpus of Estonian Dialects.

3.1. Materials

The corpus is based on the materials of the AEDKL and on the materials of the Institute of the Estonian Language, which has a large collection of dialect recordings that are also transcribed in Finno-Ugric phonetic transcription. For more information about this collection, see Ermus et al. 2019 in this volume.

The interviewed consultants were chosen on the basis of their social background: they are typically elderly people with little formal education, who have lived their entire lives in the same place in the countryside, and whose parents have these same characteristics. In the past, the dialect research tradition of Estonia has considered such consultants to be good representatives of the old local dialect.

Typically, the recordings contain traditional dialect interviews. In these a linguist interviews the consultant in surroundings familiar to him or her (the consultant's home or backyard). The topics of the interviews include life in earlier times, folk traditions, traditional work, the speaker's biography, etc.

The materials in Votic and Livonian, however, differ from Estonian dialects since these originate from more heterogeneous sources. In addition to sound recordings, there are also earlier published texts (on Votic, Elna Adler: *Vadjalaste endisajast* (1960), Paul Ariste: *Vadjalane kätkist kalmuni* (1974) and *Vadja muistendeid* (1977); on Livonian, E. N. Setälä: *Näytteitä liivin kielestä* (1953)).

3.2. Sound recordings

The oldest recordings of Estonian dialects date back to 1938 but the majority of the interviews were recorded during the 1960s and 1970s (see Figure 4). Although older materials were recorded in the studio by Paul Ariste in 1938, the nature of the interviews is the same compared with later recordings.

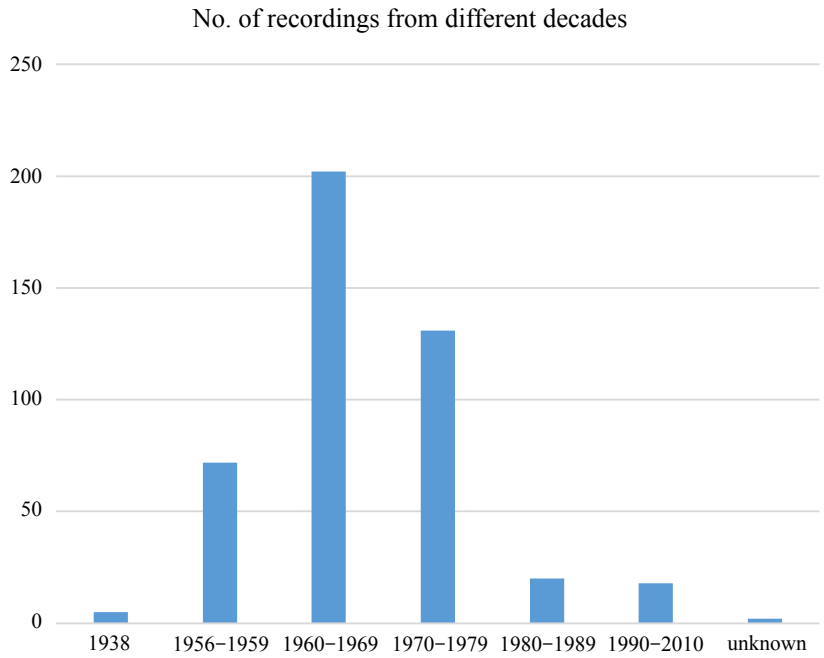


Figure 4. Number of recordings from different decades.

3.3. Phonetic transcription

In the first stage, the recordings have been transcribed using the Finno-Ugric phonetic transcription, utilizing Unicode fonts and a virtual keyboard that allows for easily combining characters with diacritics (SUT, compiled by Esko Oja). The aim has been to transcribe the texts as accurately as possible; the features of spontaneous speech (e.g., discourse particles, corrections, repetitions), which traditionally have not been considered important in dialect research, have been added to the texts. The speech of the interviewer has been transcribed as well.

The Finno-Ugric phonetic transcription has been a standard for transcribing dialect texts for a long time in traditional Estonian dialectology and most of the older transcriptions are available only in Finno-Ugric phonetic transcription. Since we have preferred the recordings that were already transcribed by earlier researchers as much

as possible, we also have preserved the transcription of these materials. However, the Finno-Ugric phonetic transcription is problematic for many reasons: 1) it is not an international standard and therefore it is hard to follow for many users; 2) it requires the use of specific non-standard fonts, a large number of diacritical marks and combinations of different diacritics for the same character; 3) it is redundant for modes of linguistic analysis other than phonetics and phonology (e.g., for analyzing dialect syntax).

To solve the problem, we automatically convert phonetically transcribed texts into a simplified transcription for futher use.

Sound recordings and transcriptions are stored at the AEDKL and can be searched via the AEDKL online database (Search: Manuscripts; Series = MKT; the links to sound recordings can be found in the manuscript description field "Related materials").

3.4. Dialect texts in simplified transcription

All phonetically transcribed texts have been converted into a simplified transcription. There are some differences in the simplified transcriptions as compared to the standard ortography; for example, the geminate voiceless consonants in Quantity 2 words have been written with two letters (*kattus* 'roof', *kadakkad* 'junipers', in standard Estonian ortography correspondingly: *katus*, *kadakad*), and the acute accent inserted before the word helps to differentiate Quantity 3 words from Quantity 2 (the acute is added only to Quantity 3 words: *`katta* 'to cover', *`kas'si* 'cat, sg partitive').

In the case of an ambiguously long quantity degree, the symbol * has been used (mainly in the Northeastern Coastal dialect group where there is no distinction between Quantity 2 and Quantity 3, e.g., **kassi* 'cat' in genitive or partitive).

Palatalization is marked with an apostrophe (e.g., *palk* 'salary' vs. *pal'k* 'log').

A more detailed overview of simplified transcription principles can be found in Lindström 2015.

3.5. Morphological annotation

As the speech of Estonian dialect varieties shows remarkable variation on all linguistic levels, the morphological annotation cannot be done automatically. In the CED, it is done semi-manually, using the program Liivike (compiled by Külli Prillop). This program automatically creates an XML file which is later uploaded to the online corpus search engine ‹www.murre.ut.ee/mkweb›. The XML files can also be downloaded and used independently from the online database.

For every word the following fields have been tagged:

- Word (*sõne*). The original form of the token as it occurs in the text (in simplified transcription), e.g., *t's'ibõrdõl'l'i* 'fidget' (past sg 3), *`vaesõq* 'poor' (pl nominative), *sääl* 'there'.
- Keyword (*märksõna*). The keyword (lemma) as it occurs in the literary language, e.g., *hüva* 'good' (using standard orthography without vowel harmony). If the word occurs only in dialects, the dialect form is used as the keyword, e.g., *tsiberdelema* 'to fidget'. If the same stem with the same meaning exists in standard Estonian, the standard Estonian word has been given as the keyword, e.g., *vaene* 'poor', *seal* 'there'. For Estonian verbs, the keyword (lemma) form is the 2nd infinitive (e.g., *õppima* 'to learn'), for Votic and Livonian, the 1st infinitive (e.g., Votic: *õppia* 'to learn', Livonian: *oppõ* 'to learn') is used.
- Meaning (*tähendus*). This field is used only when the meaning of the word is different from the literary language or when there is no equivalent word in the literary language, e.g., *tsiberdelema* 'siplema' ('to fidget'). For Votic and Livonian, the field is always filled. The meaning is provided in Estonian.
- Word class (*sõnaliik*). In the dialect corpus, the main aim of identifying word class has been to use a classification which would be sufficiently understandable and detailed for researchers working with particular dialects, while also clear-cut enough for those who do the morphological tagging. Words are divided into 24 word classes according to their morphological inflections, syntactic characteristics, and semantics. The classification is based on the system of word classes presented in Estonian grammars

(EKG I: 14–41). However, since the language in the dialect corpus is spoken, more subclasses are distinguished than in traditional grammars, referring to phenomena related to spoken language use (discourse particles, communicative words). The word classes used in the CED are represented in Table 5. For more details on the issue of word classes in the dialect corpus see Lindström et al. 2006.

- Morphological information (*vorm*). Morphological information has been added to inflected words (nouns, verbs, pronouns, adjectives, numerals, etc.).

The headers of the morphologically annotated XML files include geographic information (longitude, latitude), as well as some additional data about the recording (year of the recording, names of the recorders), and speakers (age, gender, year and place of birth, education), if available. The recordings in mp3 format are linked with the annotation.

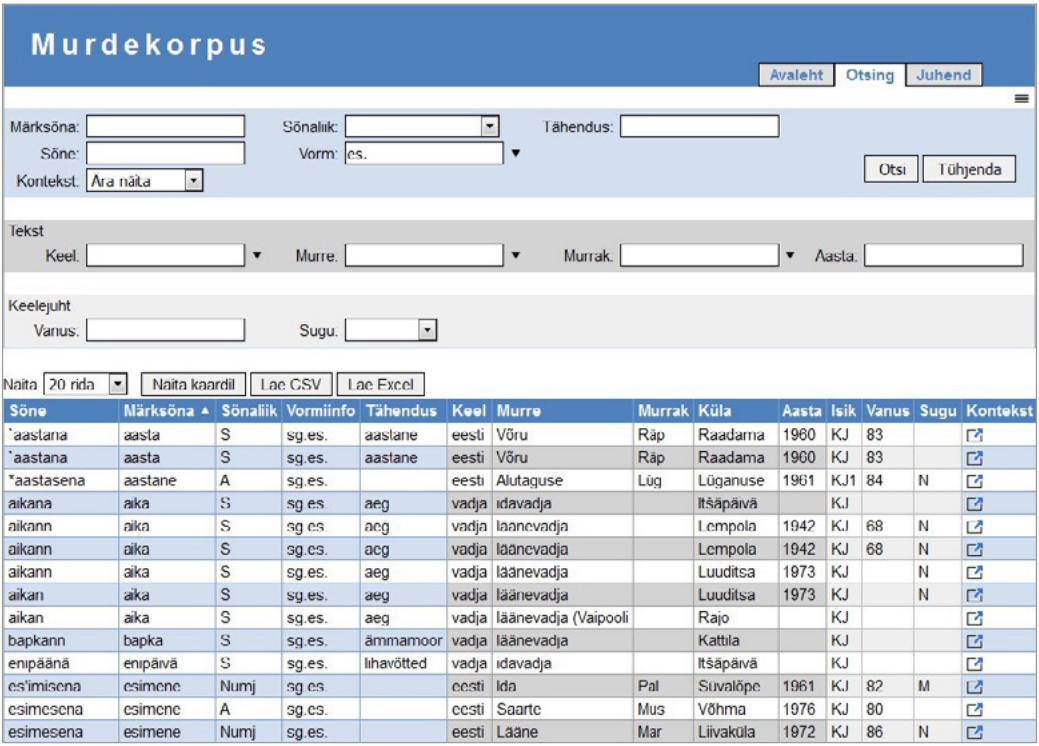
All the texts are in XML format (UTF8) and can be used as an independent source for different studies. For this purpose, a number of scripts for R and Python have been developed (by Kristel Uiboaed), which enable one to search for various information, to manipulate data, to collect and handle frequency data, and to visualize the results on maps. (See Uiboaed & Kyröläinen 2015, visualization options can be seen also in Lindström et al. 2019 in this volume.)

The online database of the morphologically annotated texts is available at ‹www.murre.ut.ee/mkweb›; it is freely accessible and updated regularly. The user interface of the search engine is only available in Estonian. In order to search within the corpus, the following fields are relevant: *Märksõna* (keyword), *Sõne* (word), *Sõnaliik* (word class), *Vorm* (morphological information), *Tähendus* (meaning), *Keel* (language), *Murre* (dialect), *Murrak* (parish, sub-dialect), *Aasta* (year of the recording), *Vanus* (age of the consultant), *Sugu* (gender of the consultant).

Search results are presented as a table (see Figure 5). The results can be sorted by columns. All search results can be downloaded either in CSV or Excel format (buttons *Lae CSV*, *Lae Excel*). The results can be viewed in context and the original recording can be played by

Word class		Abbrev.	Example
Noun (substantive)		S	*kas's* 'cat', *hommik* 'morning'
Proper name		H	*Jüri*, *Pärnumaa*
Verb		V	*ostma* 'buy'
	Auxiliary	Va	*olema* 'be' in compound tenses (*oli olnud* 'have been', *oli tehtud* 'was done')
Adverb		Adv	*täna* 'today', *kiiresti* 'quickly'
	Verb particles	Adva	particles belonging to verbs, e.g., *välja* (*mõtlema)* '(think) out', ära (*jooksma*) '(run) away'
	Modal adverb	ModAdv	*ilmselt* 'probably' *
Numerals			
	Cardinals	Nump	*kaks* 'two', *viis* 'five'
	Ordinals	Numj	*teine* 'second', *viies* 'fifth'
Adjective		A	*vana* 'old', *kole* 'ugly'
Pro-words			
	Pronoun	ProS	*see* 'this, it', *too* 'that', *tema* '(s)he', *mina* 'I'
	Proadjective	ProA	*niisuke*, *sihuke* 'such'
	Proadverb	ProAdv	*siin* 'here', *seal* 'there', *siis* 'then'
	Pronumeral	ProNum	*mitu* 'several, many'
Adpositions			
	Postposition	Post	(*maja*) *taga* 'behind the house'
	Preposition	Pre	*pärast* (*hommikusööki*) 'after breakfast'
Discourse particle		Par	*noh*, *jah*, *no*, *oi*
Communicative word		Suht	*aitäh* 'thanks', *palun* 'please', *tere* 'hi'
Onomatopoetic word		Ono	*mürts*, *pauh*
Interrogative word		Intr	*kas* 'whether', *kes* 'who', *mis* 'what', *kus* 'where'
Conjunction		Konj	*ja* 'and', *et* 'that'
Negation word		Mn	*ei*, *mitte* 'not'
Comparative word		Ms	*kõige* (*ilusam*) 'the most (beautiful)'
Interjection		Intj	*oh*, *oi* 'oh' *

Table 5. Word classes in the Corpus of Estonian Dialects. (* = Not used for Estonian dialects.)



Sõne	Märksõna	Sõnaliik	Vormiinfo	Tähendus	Keel	Murre	Murrak	Küla	Aasta	Isik	Vanus	Sugu	Kontekst
`aastana	aasta	S	sg.es.	aastane	eesti	Võru	Räp	Raadama	1960	KJ	83		
`aastana	aasta	S	sg.es.	aastane	eesti	Võru	Räp	Raadama	1960	KJ	83		
*aastasena	aastane	A	sg.es.		eesti	Alutaguse	Lüg	Lüganuse	1961	KJ1	84	N	
aikana	aika	S	sg.es.	aeg	vadja	idavadja		Itšäpärvä		KJ			
aikann	aika	S	sg.es.	aeg	vadja	läänevadja		Lempola	1942	KJ	68	N	
aikann	aika	S	sg.es.	aeg	vadja	läänevadja		Lempola	1942	KJ	68	N	
aikann	aika	S	sg.es.	aeg	vadja	läänevadja		Luuditsa	1973	KJ		N	
aikan	aika	S	sg.es.	aeg	vadja	läänevadja		Luuditsa	1973	KJ		N	
aikan	aika	S	sg.es.	aeg	vadja	läänevadja (Vaipooli		Rajo		KJ			
bapkann	bapka	S	sg.es.	ämmamoor	vadja	läänevadja		Kattila		KJ			
enipäänä	enipäivä	S	sg.es.	lihavõtted	vadja	idavadja		Itšäpärvä		KJ			
es'imesena	esimene	Numj	sg.es.		eesti	Ida	Pal	Suvalõpe	1961	KJ	82	M	
esimesena	esimene	A	sg.es.		eesti	Saarte	Mus	Võhma	1976	KJ	80		
esimesena	esimene	Numj	sg.es.		eesti	Lääne	Mar	Liivaküla	1972	KJ	86	N	

Figure 5. Search for the essive case in the CED online search engine.

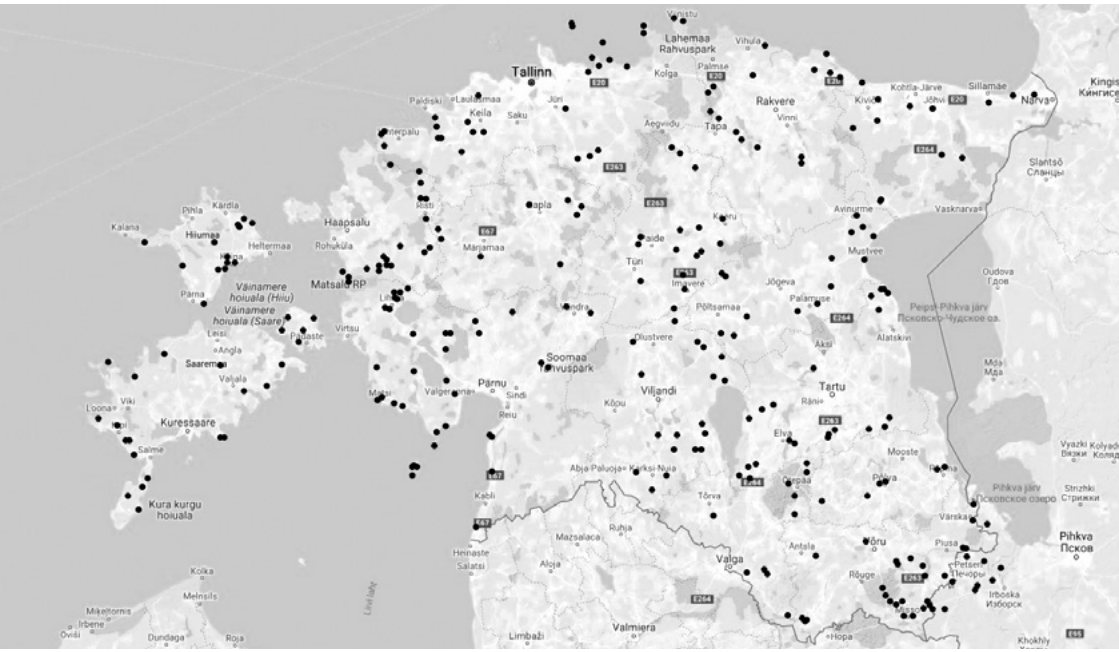
clicking on the button . The website has a simple Google map application to visualize the results on a map (button *Näita kaardil*).

Adding the original recordings to annotated texts started in 2014. By December 2019, 38% of morphologically annotated files (178 out of 467) have a sound recording available in the online database.

The number of texts included in the CED has continued to increase but still there are "white spaces" without data on the map of Estonian parishes (see also maps in Lindström et al. 2019 in this volume). Table 6 shows the content currently found in the CED. In Map 1, the data points (villages where the materials of the corpus are collected) are shown.

Dialect	No. of words in phonetic transcription	No. of morphologically annotated words
Coastal	141 390	99 372
Northeastern	83 185	63 173
Insular	317 040	220 747
Western	369 716	265 489
Mid	347 981	249 065
Eastern	66 180	50 570
Mulgi	83 335	65 432
Tartu	113 351	81 101
Võru	167 312	112 700
Seto	126 156	85 114
South Estonian enclaves in Latvia (Lutsi & Leivu)	24 268	0
Estonian dialects, Total	**1 839 914**	**1 292 763**

Table 6. Number of phonetically transcribed words and morphologically annotated words in the CED (as of December 2019).



Map 1. Data points (villages) of the corpus materials (as of December 2019). Map data ©2015 Google, Maarja-Liisa Pilvik.

The data from Votic and Livonian come from more heterogeneous sources than the Estonian data (i.e., from spoken interviews and earlier published texts). Currently, the online database contains 34 331 morphologically annotated words from Votic and 60 321 annotated words from Livonian.

The morphologically annotated texts form the core of the corpus, as they enable one to conduct different analyses on Estonian dialects as well as on Votic and Livonian. It is also the input for syntactic parsing.

3.6. Syntactically parsed texts

Syntactic parsing of the CED has been done automatically by using the parser developed by Kaili Müürisep. The parser is based on the Constraint Grammar framework and it was first developed for written texts in Standard Estonian, later adapted for spoken Estonian and Estonian dialects. (Lindström & Müürisep 2009.) For dialect texts, it uses morphological annotation as an input; the rules of the parser were adapted to the morphological annotation and XML used in the CED. Importantly, clause boundaries also had to be added by the parser, as they are missing in other layers of the corpus. (For more details, see Lindström & Müürisep 2009.) As a result, only part of the corpus is syntactically parsed (approx. 650 000 words). For searching within parsed texts by combining syntactic and morphological information, a Python script is available.

It must be noted that after syntactic parsing was carried out on the CED in 2009–2010, both the technical basis of the parser and also the XML of the corpus have changed, which means that the syntactic parsing requires technical improvements in order to continue this work.

4. Conclusions

The Archives of Estonian Dialects and Kindred Languages and the Corpus of Estonian Dialects form a collection of materials in Estonian and related languages that are freely available online. The AEDKL consists of materials that have been systematically collected in fieldwork trips by various researchers over a long period. Collecting and digitizing the data will go on and researchers continue to be encouraged to deposit their unarchived fieldwork materials in the AEDKL.

The CED offers a balanced data set from the "golden age" of Estonian dialectology, i.e., materials recorded mainly in the 1960s and 1970s. The sound recordings are annotated on various linguistic levels: phonetic transcription, morphological annotation, and syntactic parsing are provided. The corpus enables one to apply various methods that are used in corpus linguistics and corpus-based dialectology, thereby opening up new horizons in the study of certain aspects of Estonian dialects such as dialect syntax.

Acknowledgements

We would like to thank Eva Liina Asu for the first proofreading of this paper. The work of the AEDKL and CED is supported by the national program "Estonian Language and Cultural Memory" (project "Database of Estonian Dialects and Kindred Languages II"), the national program "Collections of Humanities and Natural Sciences", and the European Regional Development Fund (The Centre of Excellence in Estonian Studies).

References

Adler, Elna 1968: *Vadjalaste endisajast I. Idavadja murdetekste* [The Votes in former times I. Dialect texts in eastern Votic]. Ed. M[erle] Leppik. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Ahven, Heino 1955: Emakeele Seltsi tegevusest sõjajärgseil aastail (1945–1954) [Activities of the Mother Tongue Society during the post-war years (1945–1954)]. – *Emakeele Seltsi aastaraamat* I: 5–26.

Ariste, Paul 1974: *Vadjalane kätkist kalmuni* [Vote from cradle to grave]. Emakeele Seltsi toimetised 10. Ed. T[iit]-R[ein] Viitso. Tallinn: Eesti NSV Teaduste Akadeemia.

—— 1977: *Vadja muistendeid* [Votic stories]. Emakeele Seltsi toimetised 12. Tallinn: Eesti NSV Teaduste Akadeemia Emakeele Selts, Valgus.

EF 1997 = *Eesti filoloogia poolsajand Teaduste Akadeemias* [A half century of Estonian philology at the Academy of Sciences]. Tallinn: Eesti Keele Instituut.

EKG I = Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael & Silvi Vare 1995: *Eesti keele grammatika I. Morfoloogia. Sõnamoodustus* [Estonian grammar I. Morphology. Word formation]. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

Erelt, Mati 2010: 90 years of the Mother Tongue Society. – *Linguistica Uralica* 46 (2): 144–153.

Ermus, Liis, Mari-Liis Kalvik & Tiina Laansalu 2019: The Archive of Estonian Dialects and Finno-Ugric Languages at the Institute of the Estonian Language. – Sofia Björklöf & Santra Jantunen (eds), *Multilingual Finnic. Language contact and change*. Uralica Helsingiensia 14. Helsinki: Finno-Ugrian Society. 351–366. Available at: ‹https://doi.org/10.33341/uh.85041›

Kingisepp, Valve-Liivi 1967: Tartu Riikliku Ülikooli eesti keele kateedri murdearhiiv [The dialect archive of the Department of the Estonian language at the State University of Tartu]. – *Kodumurre* 8. Tallinn: Eesti NSV Teaduste Akadeemia, Emakeele Selts. 12–17.

Lindström, Liina 2015: Ülevaade eesti murrete korpusest [Overview of the Corpus of Estonian Dialects]. Unpublished manuscript. Available at: ‹http://www.keel.ut.ee/sites/default/files/www_ut/emk_teejuht2015.pdf›

Lindström, Liina, Liisi Bakhoff, Mari-Liis Kalvik, Anneliis Klaus, Rutt Läänemets, Mari Mets, Ellen Niit, Karl Pajusalu, Pire Teras, Kristel Uiboaed, Ann Veismann & Eva Velsker 2006: Sõnaliigituse küsi-

musi eesti murrete korpuse põhjal [Questions on word categorization explored using the Corpus of Estonian Dialects]. – E[llen] Niit (ed.), *Keele ehe*. Tartu Ülikooli eesti keele õppetooli toimetised 30. Tartu: Tartu Ülikool. 154–167.

Lindström, Liina & Kaili Müürisep 2009: Parsing corpus of Estonian dialects. – E. Bick, K. Hagen, K. Müürisep & T. Trosterud (eds), *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing, Odense, Denmark; 14.05.2009*. NEALT Proceedings Series. Tartu: Tartu University Library.

Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma & Kristel Uiboaed 2019: On the use of perfect and pluperfect in Estonian dialects: Frequency and language contacts. – Sofia Björklöf & Santra Jantunen (eds), *Multilingual Finnic. Language contact and change*. Uralica Helsingiensia 14. Helsinki: Finno-Ugrian Society. 155–193. Available at: ‹https://doi.org/10.33341/uh.85035›

Rätsep, Huno 2003: Tartu Ülikooli eesti keele arhiivi saamisloost ja saatusest [On the origin and destiny of the Estonian language archive at the University of Tartu]. – *200 aastat eesti keele ülikooliõpet. Juubelikogumik*. Tartu Ülikooli eesti keele õppetooli toimetised 25. 153–170.

Rozhanskiy, Fedor & Elena Markus 2019: A new resource for Finnic languages: The outcomes of the Ingrian documentation project. – Sofia Björklöf & Santra Jantunen (eds), *Multilingual Finnic. Language contact and change*. Uralica Helsingiensia 14. Helsinki: Finno-Ugrian Society. 303–326. Available at: ‹https://doi.org/10.33341/uh.85039›

Uiboaed, Kristel & Aki-Juhani Kyröläinen 2015: Keeleteaduslike andmete ruumilisi visualiseerimisvõimalusi [Spatial visualizing possibilities for linguistic information]. – *Eesti Rakenduslingvistika Ühingu aastaraamat* 11: 281–295.

Tartu Ülikooli eesti murrete ja sugulaskeelte digitaalne arhiiv ja Eesti murrete korpus

Liina Lindström, Pärtel Lippus & Tuuli Tuisk

Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivi on koondatud helisalvestised eesti murretest ja sugulaskeeltest, käsikirjalised materjalid, keele kogumise ja keeleteadusega seotud fotod ning videosalvestused. Arhiiv sisaldab nelja tüüpi materjali: 1) helisalvestisi eesti murretest ja sugulaskeeltest (alates 1950. aastatest); 2) käsikirjalisi materjale, sh 1920.–1980. aastatel eesti keele kateedris tehtud kursuse- ja seminaritöid ning lõputöid alates 1940. aastatest kuni tänapäevani, murdetekstide transkriptsioone ja kuuldelisi kirjapanekuid, murdepäevikuid jms; 3) keele kogumise ja keeleteadusega seotud fotosid; 4) videosalvestusi. Arhiivi kogusid on digitaliseeritud alates 2000. aastast. Praeguseks on arhiivis u 2800 tundi helisalvestusi, käsikirju u 396 000 lehekülge ja fotosid ligi 3000. Arhiivi maht kasvab iga-aastaste välitööde ja murdepraktika materjalide ning lõputööde lisandudes. Aastal 2012 valmis arhiivi veebipõhine andmebaas, mis asub aadressil ‹https://murdearhiiv.ut.ee/› ning on avatud kõigile huvilistele ja uurijatele.

Arhiivi eraldiseisev osa on Eesti murrete korpus – elektrooniline andmekogu, mis sisaldab morfoloogiliselt märgendatud murdetekste ja on kasutatav iseseisvana. Selle põhieesmärk on teha hoolikalt valitud ja täpselt litereeritud materjalid kõigist eesti murretest uurijatele elektrooniliselt kättesaadavaks. Lisaks on murdekorpuse tekstid märksõnastatud ja morfoloogiliselt märgendatud. Märgendatud korpus paikneb aadressil ‹www.murre.ut.ee/mkweb› ning sisaldab 1 241 233 tekstisõna eesti murretest, 34 331 vadja keelest ja 60 321 sõna liivi keelest. Korpuse põhjal on võimalik uurida võrdlevalt eesti murdeid nii häälikulisel, morfoloogilisel kui süntaktilisel tasandil ning võrrelda eesti murrete andmeid liivi ja vadja keelega.