Applying spatial data in linguistics

Kristel Uiboaed, Siim Antso, Liina Lindström,
Maarja-Liisa Pilvik, Mirjam Ruutma

University of Tartu

Overview

- Project
- Resources and data
- Workflow
- Conclusions

Project

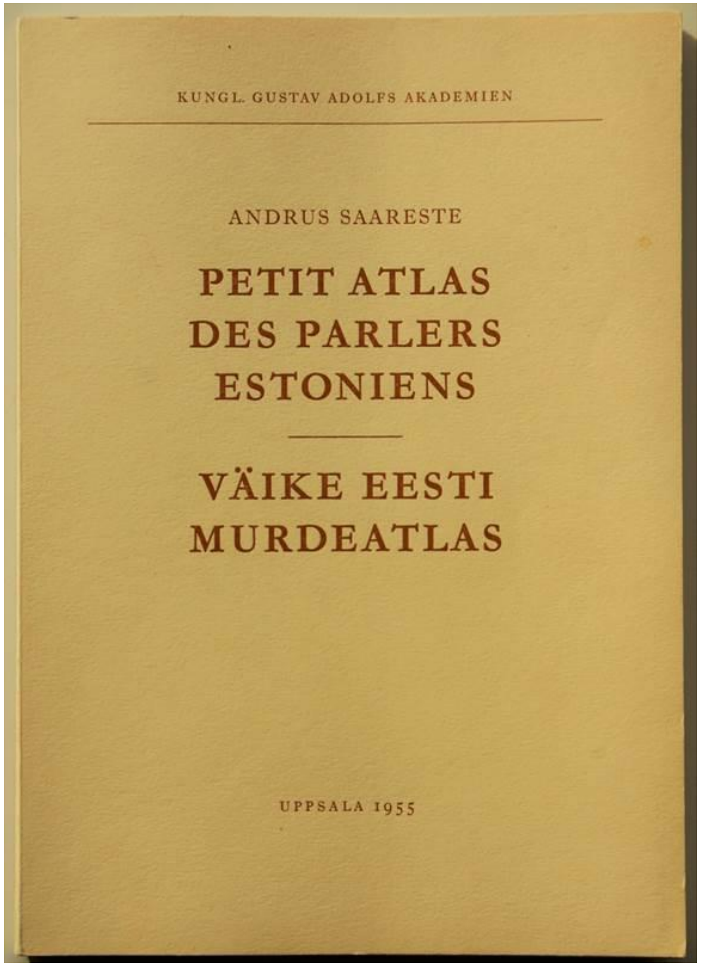
- Applying spatial data in linguistic research (01.04.2015–31.03.2017)
- Objectives:
  - digitalizing (scanning and georeferencing) data in Andrus Saareste’s dialect atlases and his unpublished manuscript maps
  - integrating map applications with data from the Corpus of Estonian Dialects (CED)
  - representing linguistic distances in CED by building interactive applications for multidimensional analyses
  - making the above-mentioned resources freely available

Andrus Saareste’s maps

- Mainly based on the data A. Saareste and his colleagues collected in Estonia during 1915–1944 (and that he managed to take with him when he emigrated to Sweden), to a lesser extent also on the material he collected from over 200 Estonian refugees while staying in Uppsala in 1945–1954.

Andrus Saareste’s maps

- Atlases
  - Petit atlas des parlers estoniens. Väike eesti murdeatlas (1955)
  - Eesti murdeatlas. I vihik (Atlas des parlers estoniens) (1938)
  - Eesti murdeatlas. II vihik (Atlas des parlers estoniens) (1941)
  - Altogether 215 maps
  - Based on Estonia’s map during 1920–1940ies
- Manuscripts
  - Maps in Uppsala University archives
  - 2500 maps scanned and to be digitalized



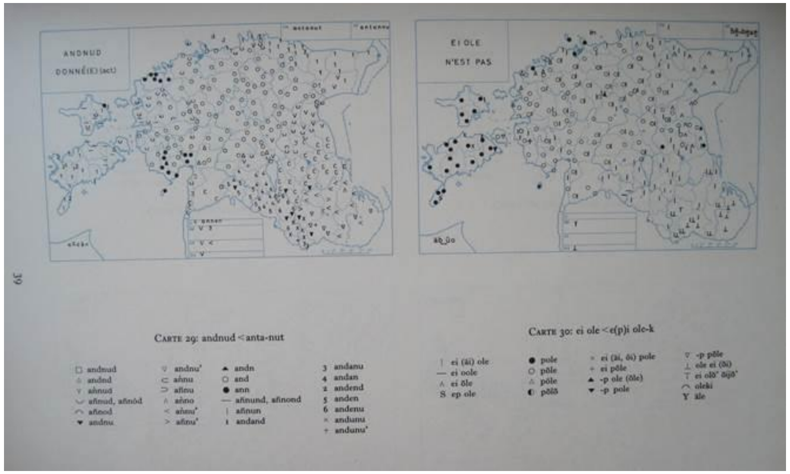
KUNGL. GUSTAV ADOLFS AKADEMIEN

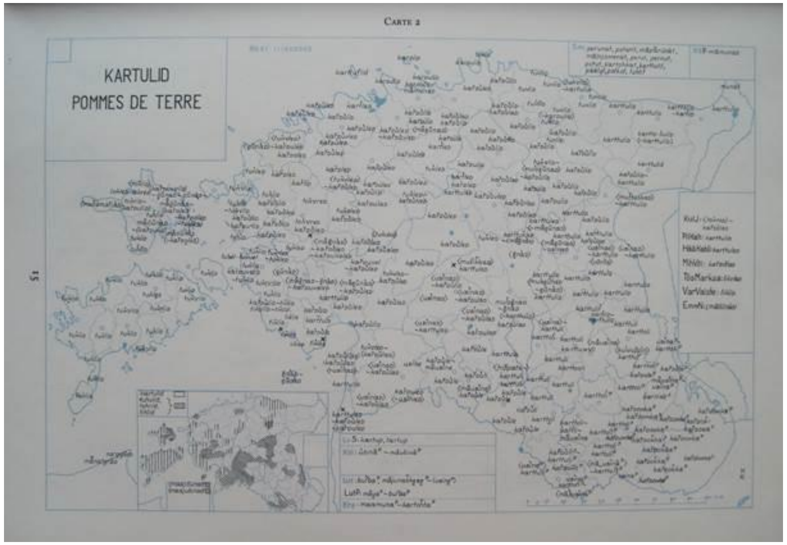
ANDRUS SAARESTE

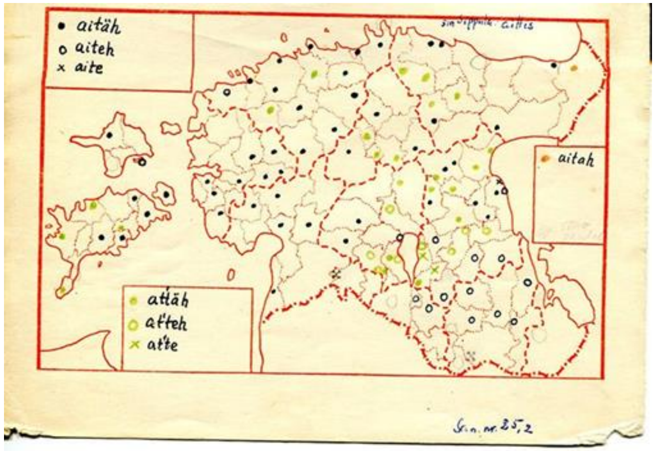
PETIT ATLAS DES PARLERS ESTONIENS

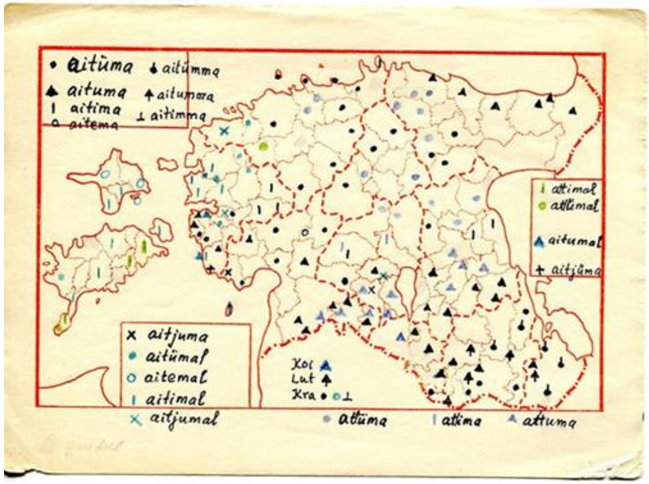
VÄIKE EESTI MURDEATLAS

UPPSALA 1955






• aitäh
○ aiteh
× aite
aitah
attäh
atteh
atte
Gr.n.nr. 25,2

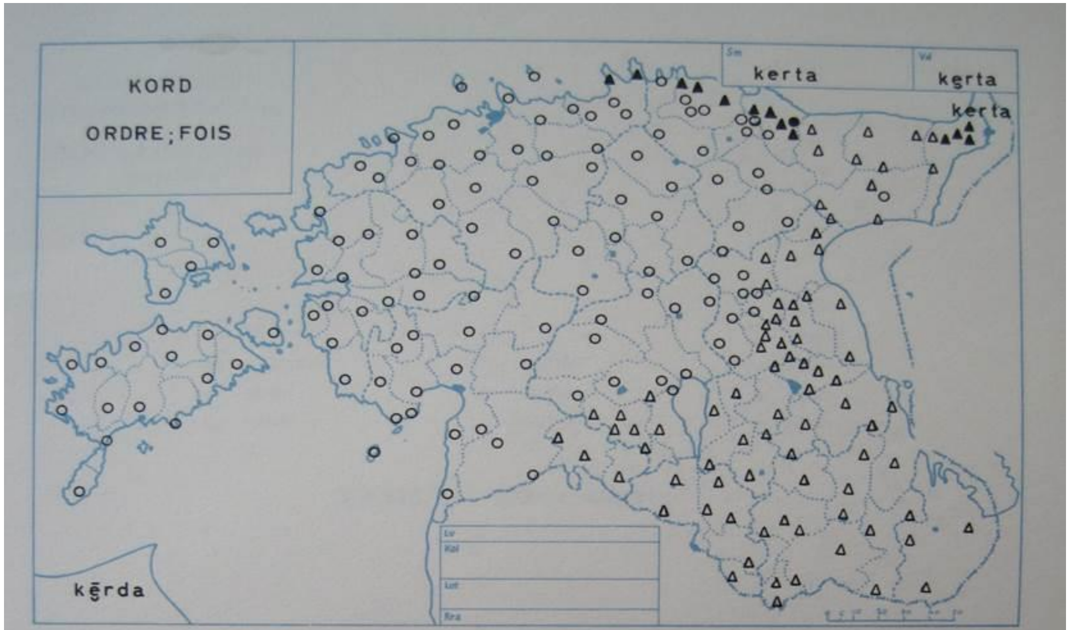

• aitüma
aitümma
▲ aituma
aitumma
| aitima
⊥ aitimma
○ aitema
× aitjuma
• aitümal
○ aitemal
| aitimal
× aitjumal
| attimal
• attümal
▲ aitumal
+ aitjüma
Koi
Lut
Kra
• attüma
| attima
▲ attuma

Workflow: resources

- GIS for map digitalization
- R for analysis and visualization
- Free map resources (Maa-amet Geoportal, Google maps)

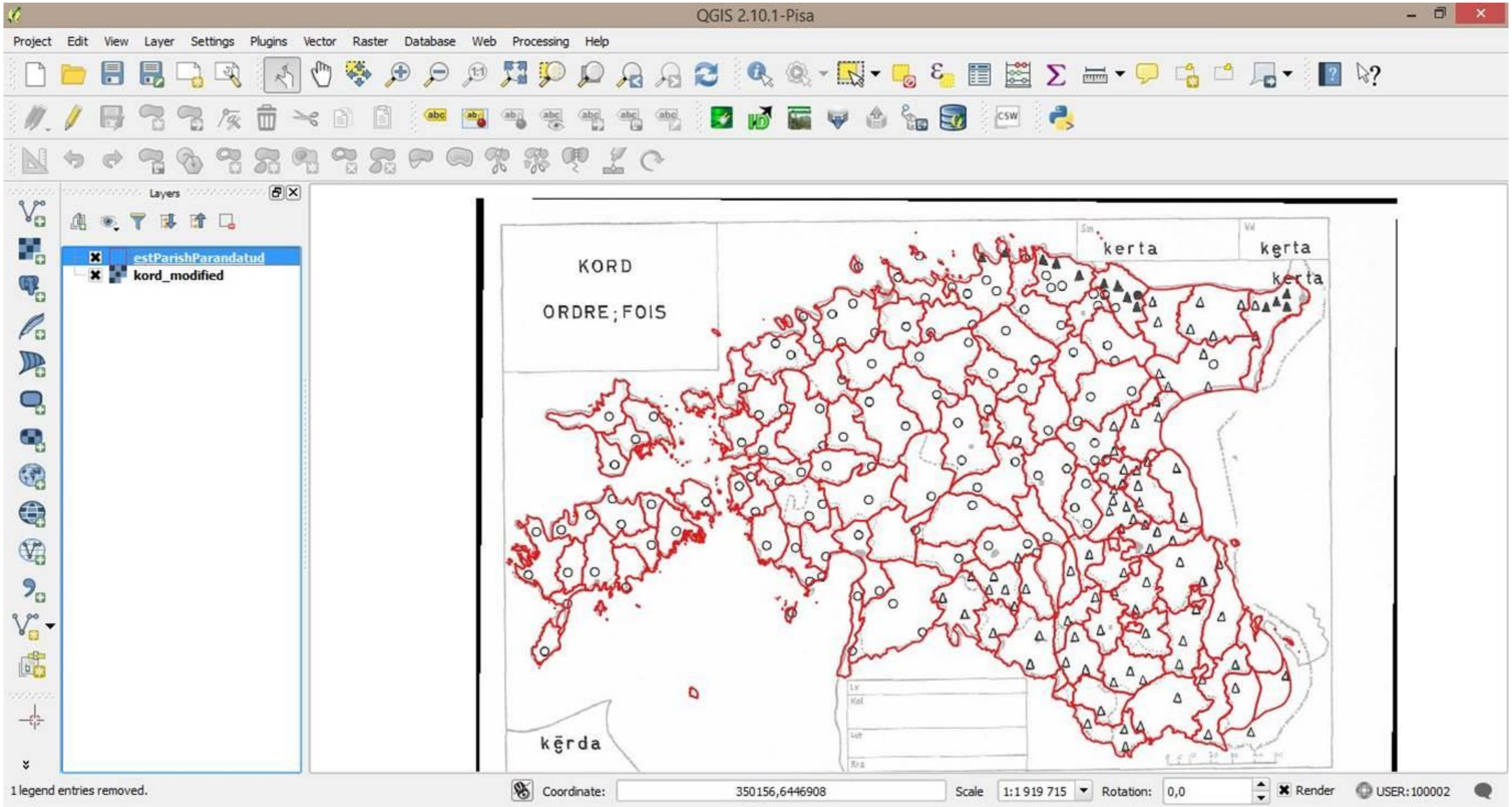
Data insertion

- Geographical data collection for old data, aligning and verifying.
- Construct a shapefile map with the data collection points using the Maa-amet settlements base map (contemporary administrative division).
- Several villages in older data no longer exist → nearest settlement to mark that data collection point.
- Scan the maps from the *Petit Atlas des parlers estoniens* and georeference them.
- Insert the data on the map into the shapefile’s attribute table.
- Design the final map (choose matching symbology, correct the position of the labels).



Carte 56: kord < ke̮rta

○ kord △ kõrd ▲ kerd ● kerda



☒ estParishParandatud
☒ kord_modified

1 legend entries removed.

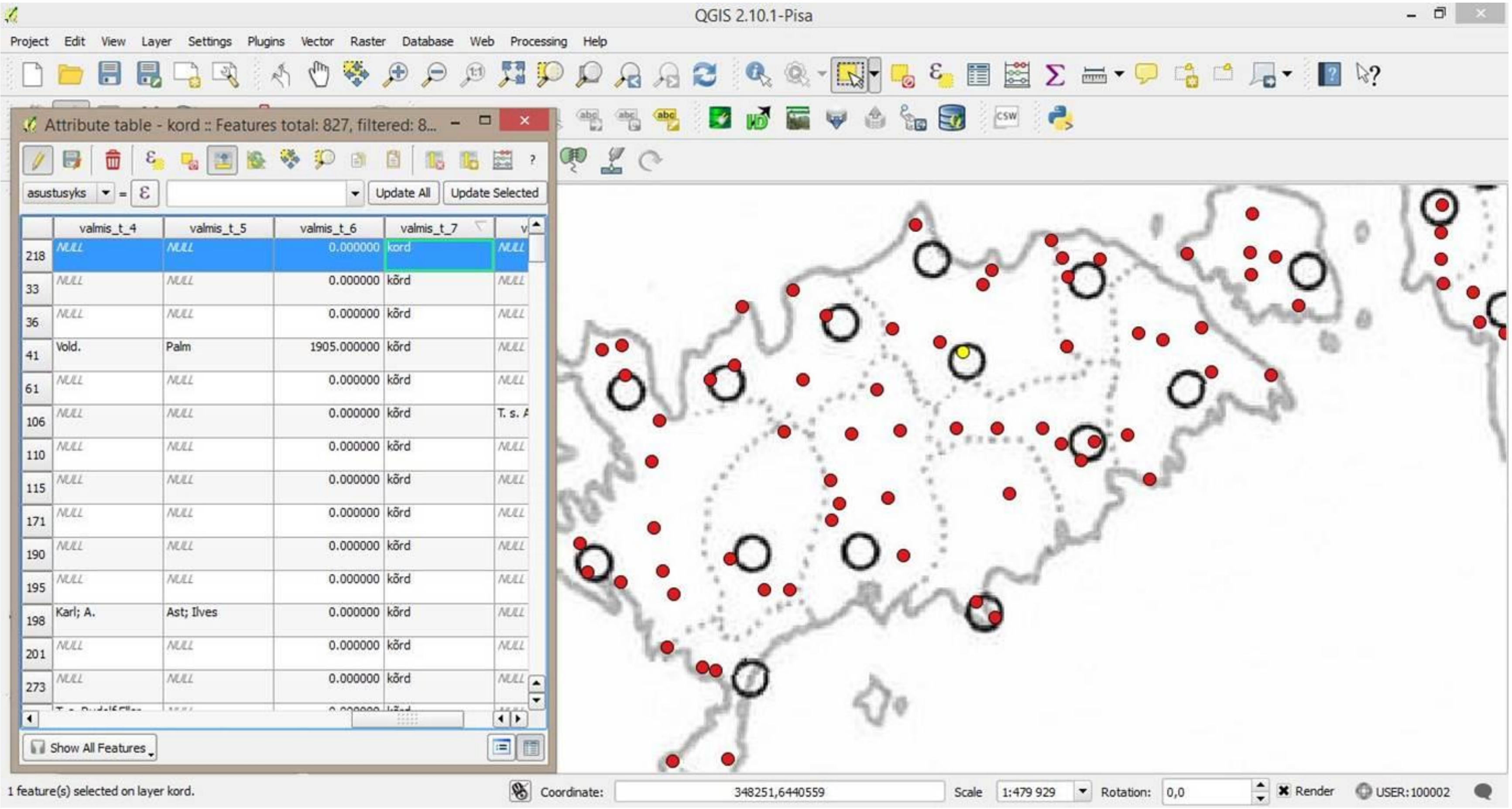
Coordinate: 350156,6446908

Scale 1:1 919 715

Rotation: 0,0

☒ Render

USER:100002



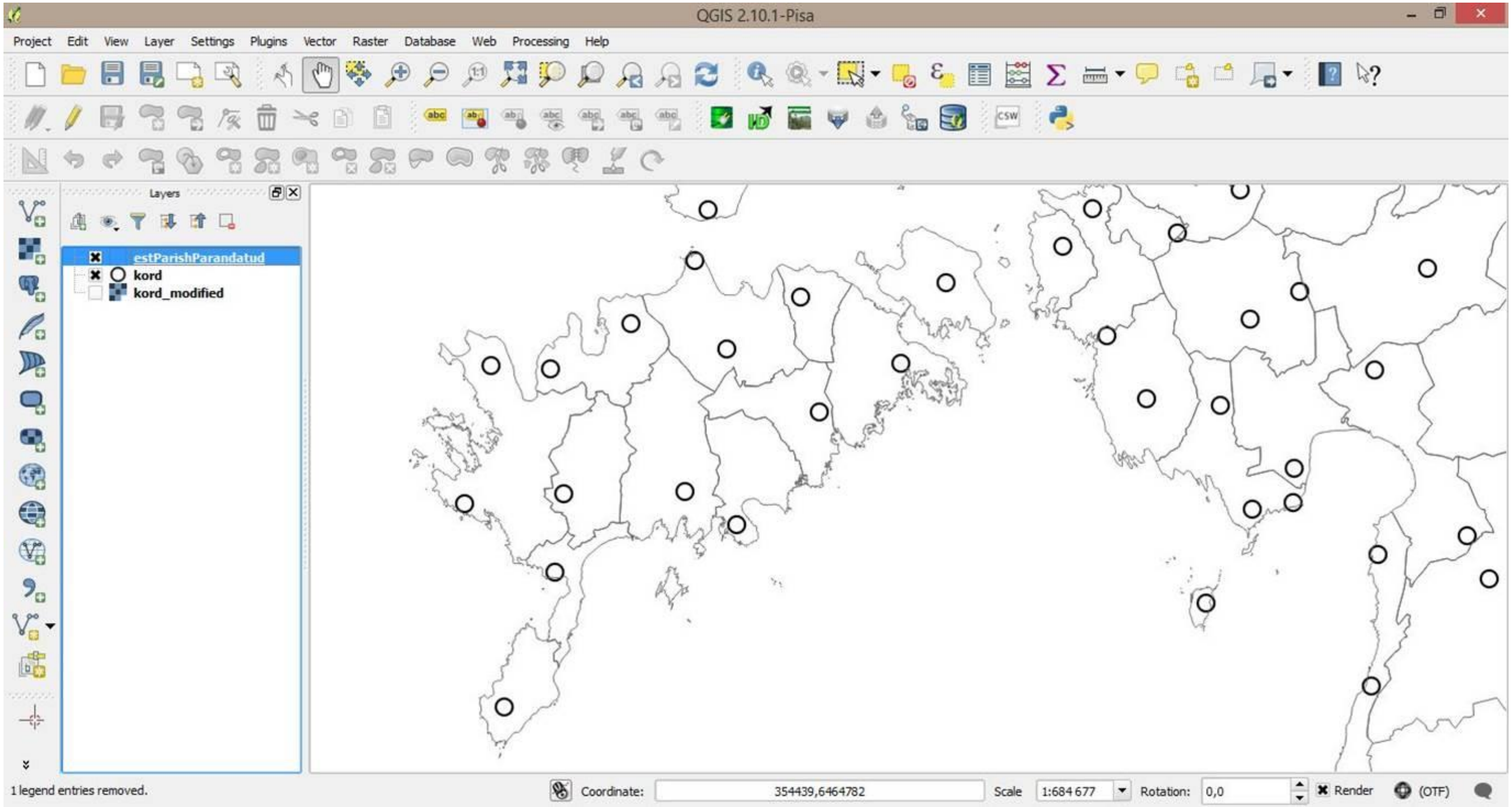
Attribute table - kord :: Features total: 827, filtered: 8...

asustusyks = ε | Update All | Update Selected

	valmis_t_4	valmis_t_5	valmis_t_6	valmis_t_7	v
218	NULL	NULL	0.000000	kord	NULL
33	NULL	NULL	0.000000	kõrd	NULL
36	NULL	NULL	0.000000	kõrd	NULL
41	Vold.	Palm	1905.000000	kõrd	NULL
61	NULL	NULL	0.000000	kõrd	NULL
106	NULL	NULL	0.000000	kõrd	T. s. A
110	NULL	NULL	0.000000	kõrd	NULL
115	NULL	NULL	0.000000	kõrd	NULL
171	NULL	NULL	0.000000	kõrd	NULL
190	NULL	NULL	0.000000	kõrd	NULL
195	NULL	NULL	0.000000	kõrd	NULL
198	Karl; A.	Ast; Ilves	0.000000	kõrd	NULL
201	NULL	NULL	0.000000	kõrd	NULL
273	NULL	NULL	0.000000	kõrd	NULL

1 feature(s) selected on layer kord.

Coordinate: 348251,6440559 Scale 1:479 929 Rotation: 0,0 USER:100002



☒ estParishParandatud
☒ ○ kord
□ kord_modified

1 legend entries removed.

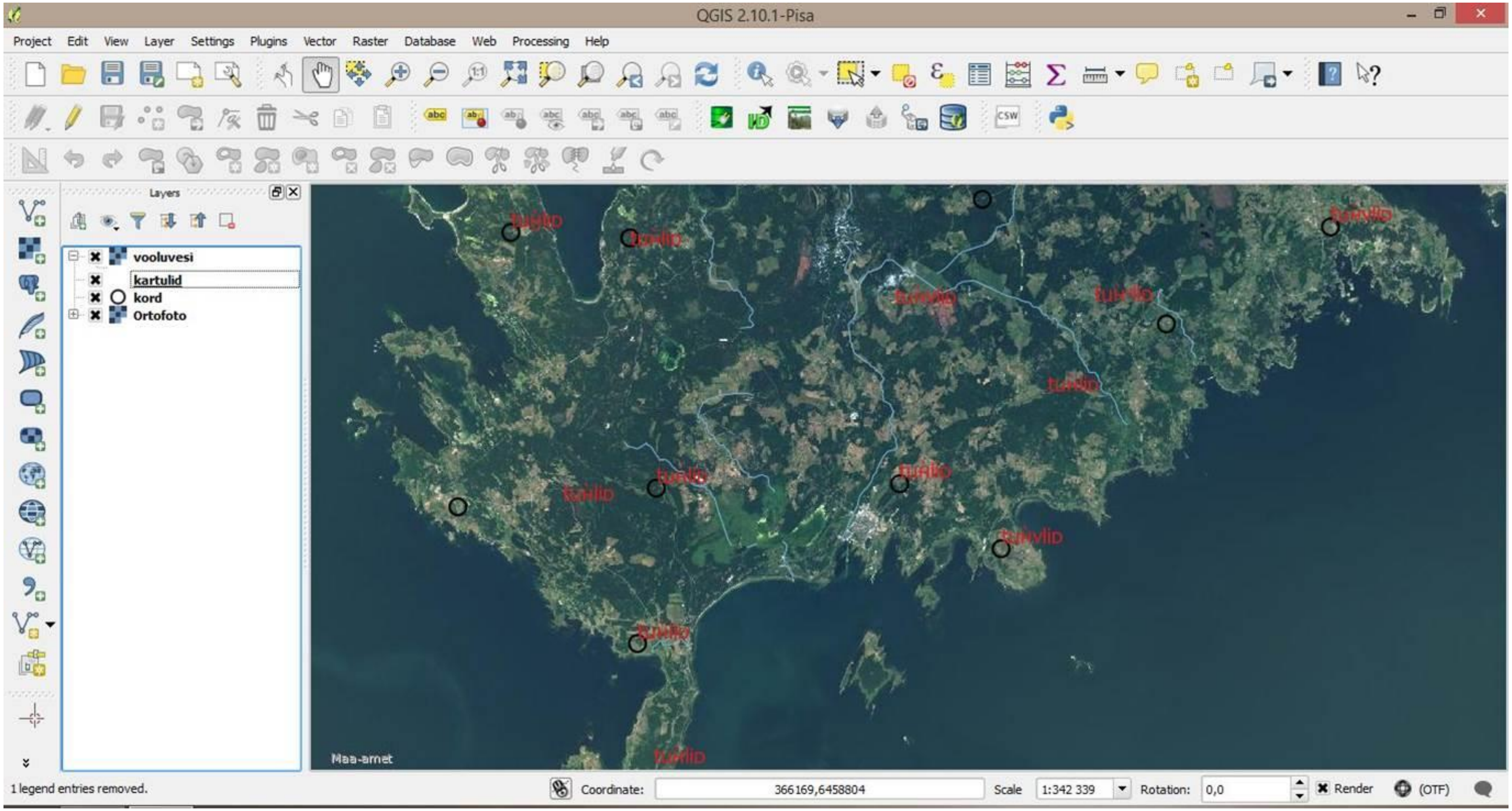
Coordinate: 354439,6464782

Scale 1:684 677

Rotation: 0,0

☒ Render

(OTF)



☒ vooluvesi
☒ kartulid
☒ ○ kord
☒ Ortofoto

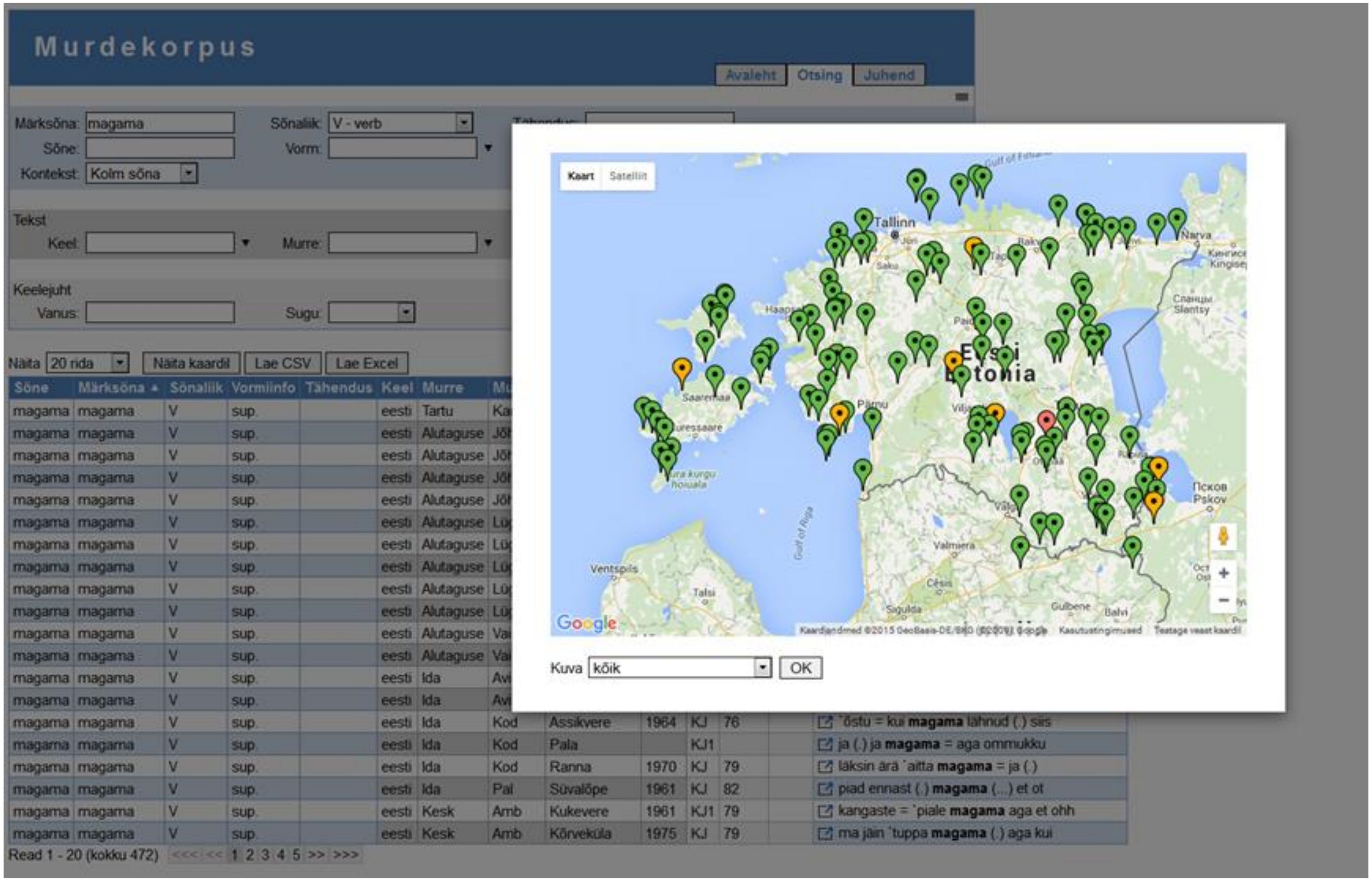
1 legend entries removed.

Coordinate: 366169,6458804 Scale 1:342 339 Rotation: 0,0 ☒ Render (OTF)

Corpus of Estonian Dialects

- www.murre.ut.ee/mkweb
- CED consists of dialect recordings (mainly from 1960-1970s), phonetically transcribed texts, dialect texts in simplified transcription, morphologically annotated texts, syntactically parsed texts, and a database with metadata about all the recordings and texts
- Older, non-mobile, less educated speakers
- 10 traditional dialect areas
- Morphologically annotated texts in XML format, which enables automated data extraction

CED user interface



Murdekorpus

Märksõna: magama
Sõnaliik: V - verb
Tähendus:
Sõne:
Vorm:
Kontekst: Kolm sõna

Tekst
Keel:
Murre:

Keelejuht
Vanus:
Sugu:

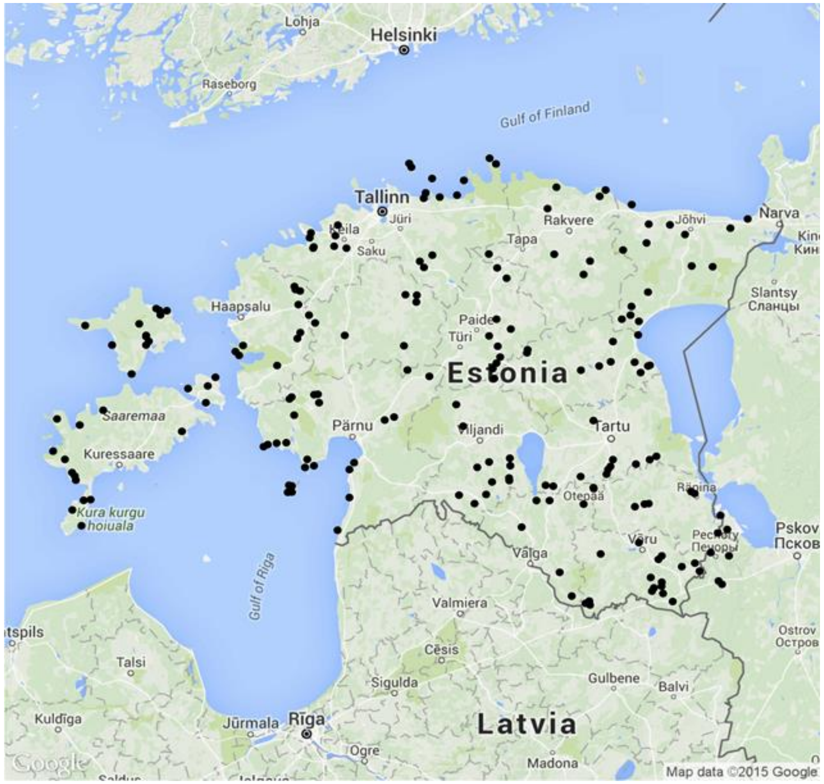
Näita 20 rida

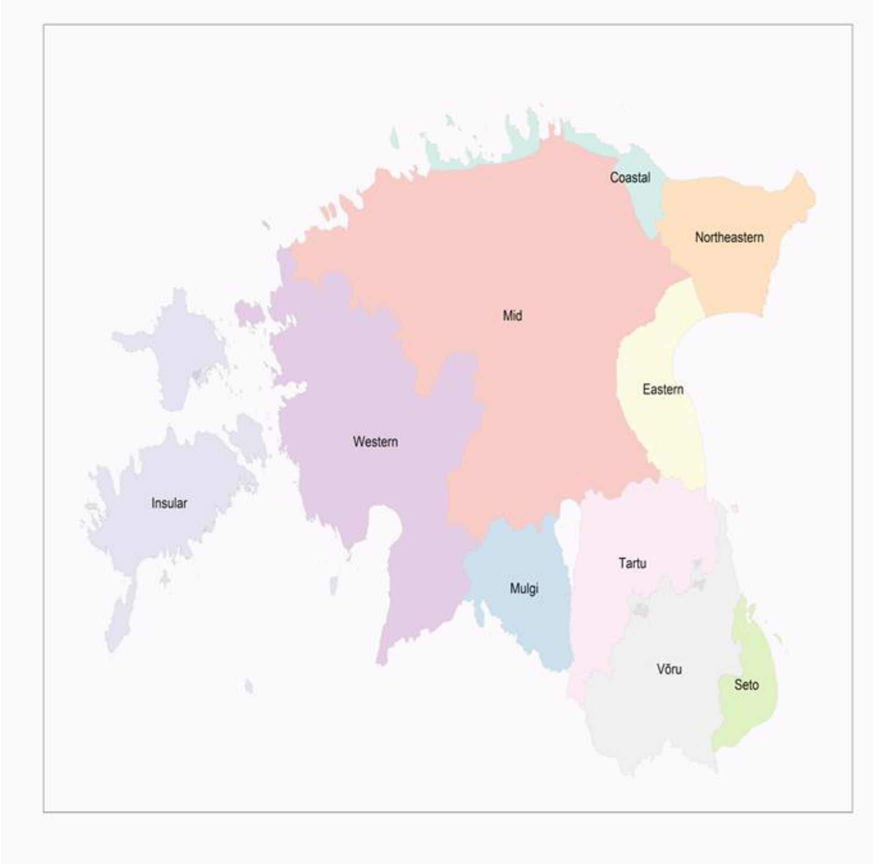
Sõne	Märksõna ▲	Sõnaliik	Vorminfo	Tähendus	Keel	Murre	Mu						
magama	magama	V	sup.		eesti	Tartu	Ka						
magama	magama	V	sup.		eesti	Alutaguse	Jõ						
magama	magama	V	sup.		eesti	Alutaguse	Jõ						
magama	magama	V	sup.		eesti	Alutaguse	Jõ						
magama	magama	V	sup.		eesti	Alutaguse	Jõ						
magama	magama	V	sup.		eesti	Alutaguse	Lü						
magama	magama	V	sup.		eesti	Alutaguse	Lü						
magama	magama	V	sup.		eesti	Alutaguse	Lü						
magama	magama	V	sup.		eesti	Alutaguse	Lü						
magama	magama	V	sup.		eesti	Alutaguse	Lü						
magama	magama	V	sup.		eesti	Alutaguse	Va						
magama	magama	V	sup.		eesti	Alutaguse	Va						
magama	magama	V	sup.		eesti	Ida	Av						
magama	magama	V	sup.		eesti	Ida	Av						
magama	magama	V	sup.		eesti	Ida	Kod	Assikvere	1964	KJ	76		´õstu = kui **magama** lähnud (.) siis
magama	magama	V	sup.		eesti	Ida	Kod	Pala		KJ1			ja (.) ja **magama** = aga ommukku
magama	magama	V	sup.		eesti	Ida	Kod	Ranna	1970	KJ	79		läksin ärä ´aitta **magama** = ja (.)
magama	magama	V	sup.		eesti	Ida	Pal	Süvalõpe	1961	KJ	82		piad ennast (.) **magama** (...) et ot
magama	magama	V	sup.		eesti	Kesk	Amb	Kukevere	1961	KJ1	79		kangaste = ´piale **magama** aga et ohh
magama	magama	V	sup.		eesti	Kesk	Amb	Kõrveküla	1975	KJ	79		ma jäin ´tuppa **magama** (.) aga kui

Read 1 - 20 (kokku 472) <<< << 1 2 3 4 5 >> >>>

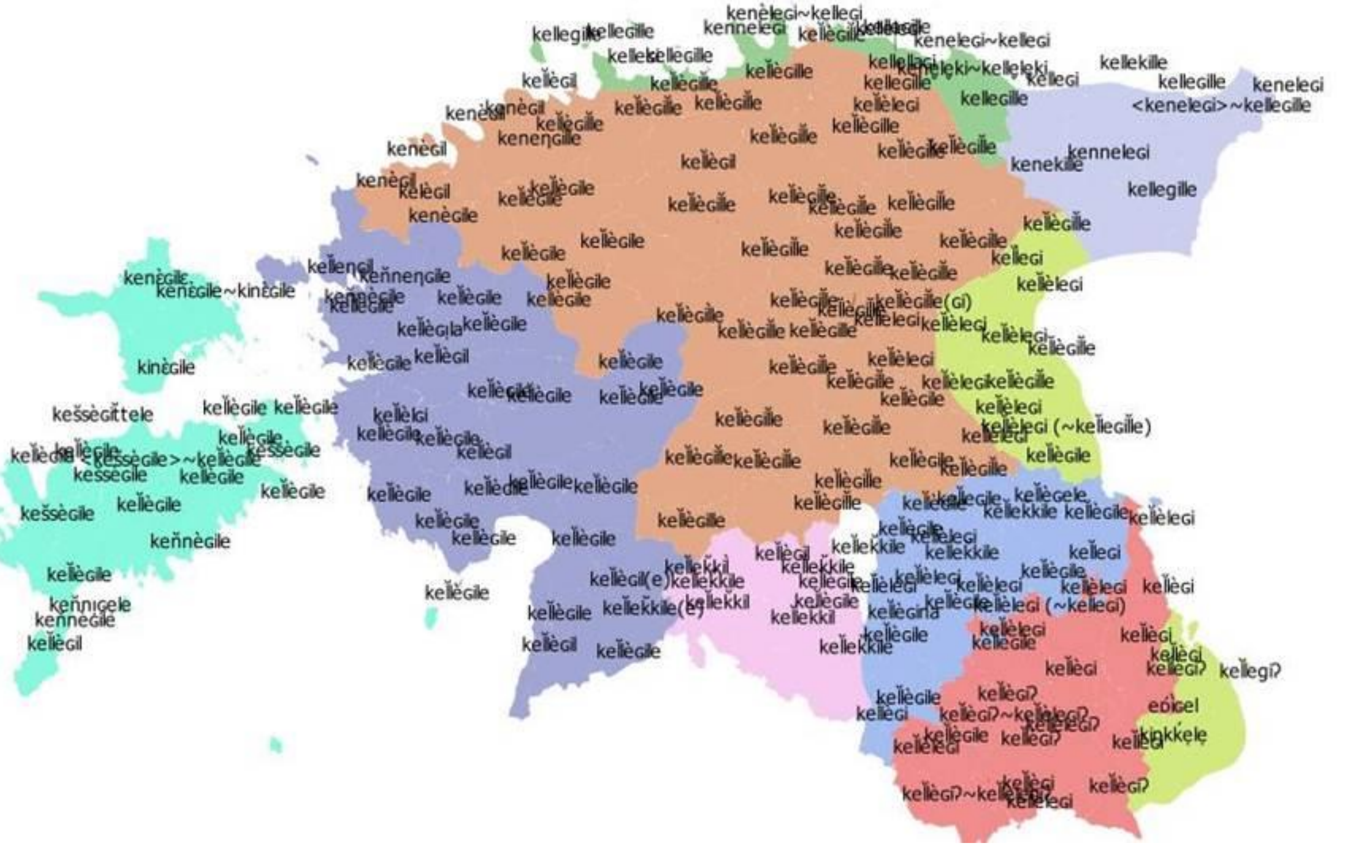
Kuva kõik

CED map data


Lohja
Helsinki
Raseborg
Gulf of Finland
Tallinn
Jüri
Keila
Saku
Tapa
Rakvere
Jõhvi
Narva
Kin
Кин
Slantsy
Сланцы
Haapsalu
Paide
Türi
Estonia
Saaremaa
Kuressaare
Kura kurgu hoiuala
Pärnu
Viljandi
Tartu
Otepää
Räpina
Võru
Pechory
Печоры
Pskov
Псков
Valga
Valmiera
Gulf of Riga
tspils
Talsi
Cēsis
Sigulda
Gulbene
Balvi
Ostrov
Остров
Kuldīga
Jūrmala
Rīga
Ogre
Latvia
Madona
Google
Map data ©2015 Google


Coastal
Northeastern
Mid
Eastern
Western
Insular
Tartu
Mulgi
Võru
Seto

Combining data sources



Applications and conclusions

- CED data: frequencies and typicalities
- Atlas data: categorical
- Digitalized geographical information enables to compare this data
- Aggregating the data
- Linguistic distances based on both data sources
- Comparison of old data and new research findings
- Measuring linguistic and geographic distances

References

CED = Corpus of Estonian Dialects. <*http://www.murre.ut.ee/estonian-dialect-corpus/*>

Google Maps (2014) Google.

Kahle, David & Wickham, Hadley (2013) ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.3.

<*http://CRAN.R-project.org/package=ggmap*>

EKI (2014) = Eesti Keele Instituudi kohanimeandmebaasi kihelkonnapiiride andmestik [Map of the Place Name Database created by the Institute of Estonian Language].

References

Maa-amet Geoportal. <*http://geoportaal.maaamet.ee/eng/*>

QGIS Development Team (2015). QGIS Geographic Information System (*version* 2.10.1). Open Source Geospatial Foundation.

Saareste, Andrus (1938). Eesti murdeatlas. I vihik. Tartu: Eesti Kirjanduse Selts.

Saareste, Andrus (1941). Eesti murdeatlas. II vihik. Tartu: Teaduslik Kirjandus.

Saareste, Andrus (1955). Petit Atlas des parlers estoniens. Väike eesti murdeatlas. – Skrifter utgivna av Kungl. Gustav Adolfs Akademien 28. Uppsala: Almqvist & Wiksell.