A DigiHum approach to HistSocLing of Estonian

Peeter Tinits

25.07.2019

National Library of Latvia

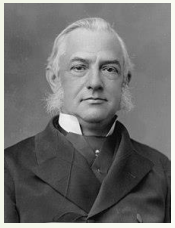
BSSD2019

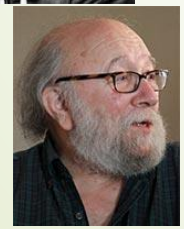
A case study in DH

- Project:
  - What mechanisms are responsible for spelling standardization at the end of 19th century Estonian?
- Data sources:
  - Anything relevant
- Timeline:
  - PhD dissertation

Some context

- Language change: intentional or not





Müller (1861), Saussure (1916), Lass (1999): „Intentional language change is peripheral to language“

Jesperson (1925), Ferguson (1989), Thomason (2007) „Intentional language change is quite central to language phenomenon“

Written grammars as part of language

- Language institutions and written norms have become normal for us.

Members of *Académie Française (2019)*

Languages without these grammars

- Sociolinguistic research has shown that this is not always the case
- Of ~7,000 languages, ~3,000 still are not written

Even much writing is without standards

Shakespere,
Shackspeare,
Shakespear,
Shakspere,
Shaxspere,
Shaxper,
Shakspeare,
Shackespeare,
Shackspere,
Shackespere

- E.g. English became standardised from ~18c onwards


A
DICTIONARY
OF THE
ENGLISH LANGUAGE:
IN WHICH
The WORDS are deduced from their ORIGINALS,
AND
ILLUSTRATED in their DIFFERENT SIGNIFICATIONS
BY
EXAMPLES from the beſt WRITERS.
TO WHICH ARE PREFIXED,
A HISTORY of the LANGUAGE,
AND
An ENGLISH GRAMMAR.
By SAMUEL JOHNSON, A.M.
In TWO VOLUMES.
VOL. I.
THE SECOND EDITION.
LONDON,
MDCCLV.

Samuel Johnson (1750): *Dictionary of English Language*

- *Shakespeare* standardized from 1860 onwards only

Standardization research

- Spelling practices vary by community.
- Mechanisms and developments may depend on particular circumstances.

- How did this take place for (some of) Estonian?

Estonian in the late 19c

- From 1800-1900:
  - Peasant population becomes independent
  - Quick developments in education
    - Literacy from 20% to 95%
  - Intense travel and movement
    - Urbanization, railroads
  - Social and cultural modernization
    - Cultural production and identity
    - Strive to become a nation

Language standardization

- One form instead of many (colour & color -> color)
- How can this happen?
  - Talk like your neighbour
  - Follow a book
  - Survive longer
- For mechanisms, we should follow the process

Finding the data

What is needed? What is available?

Texts

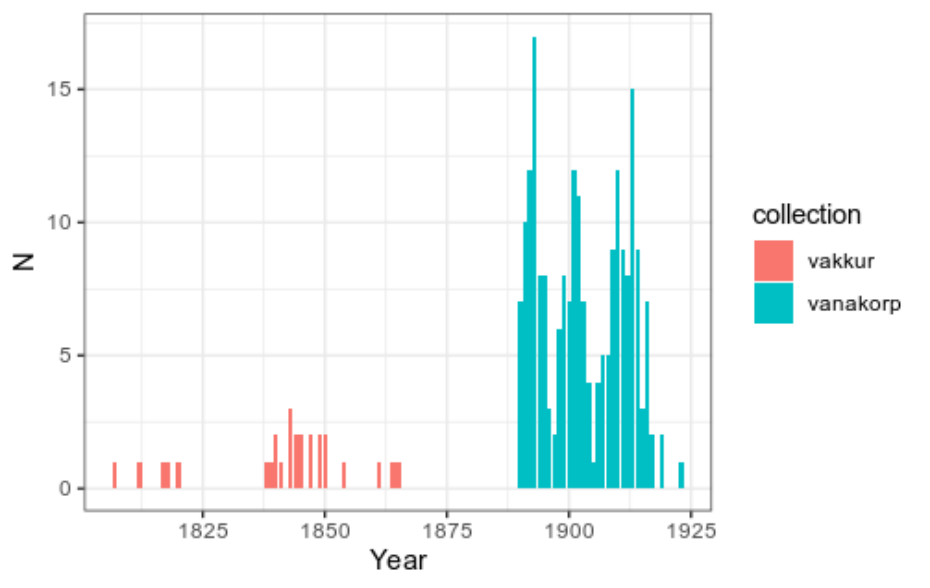
- To study writing, we need texts:
  - Ideally from a variety of people
    - Education, dialect, profession
- Sources:
  - Linguistic corpora
  - Digitized works

Linguistic corpora

- Linguistic corpora are balanced collections of texts to facilitate language research.
- For the period, there's two:
  - Written Estonian Corpus
    - some 300 book snippets in 1890-1930
  - Old Written Estonian Corpus
    - all texts 1500-1700, some 50 texts from later
- Small-ish and not much metadata

Linguistic corpora

- Texts have high quality (though spelling was not in focus)
- Relatively small though.

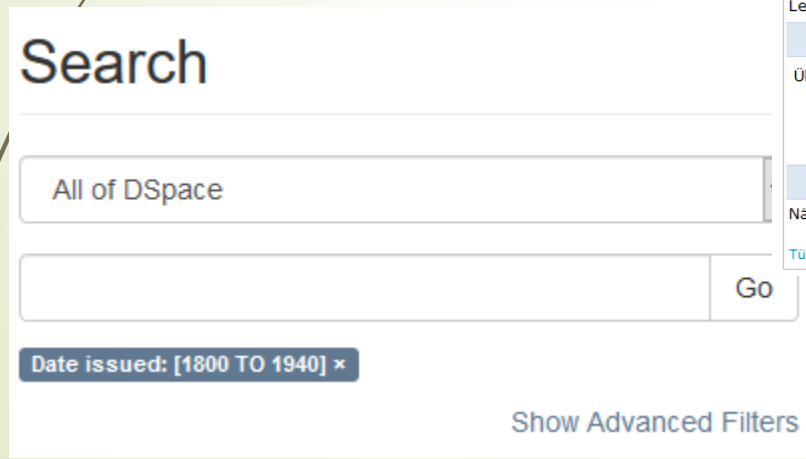


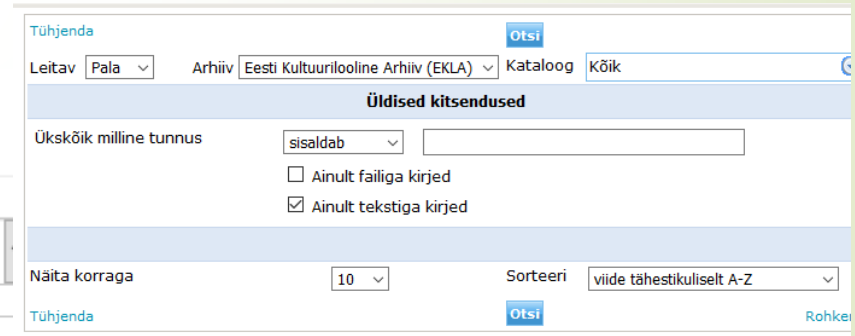
Digitized texts

- For the past 20-30 years, a huge number of texts have been digitized.
  - In various collections and formats
  - With varying editing practices

Examples of text collections

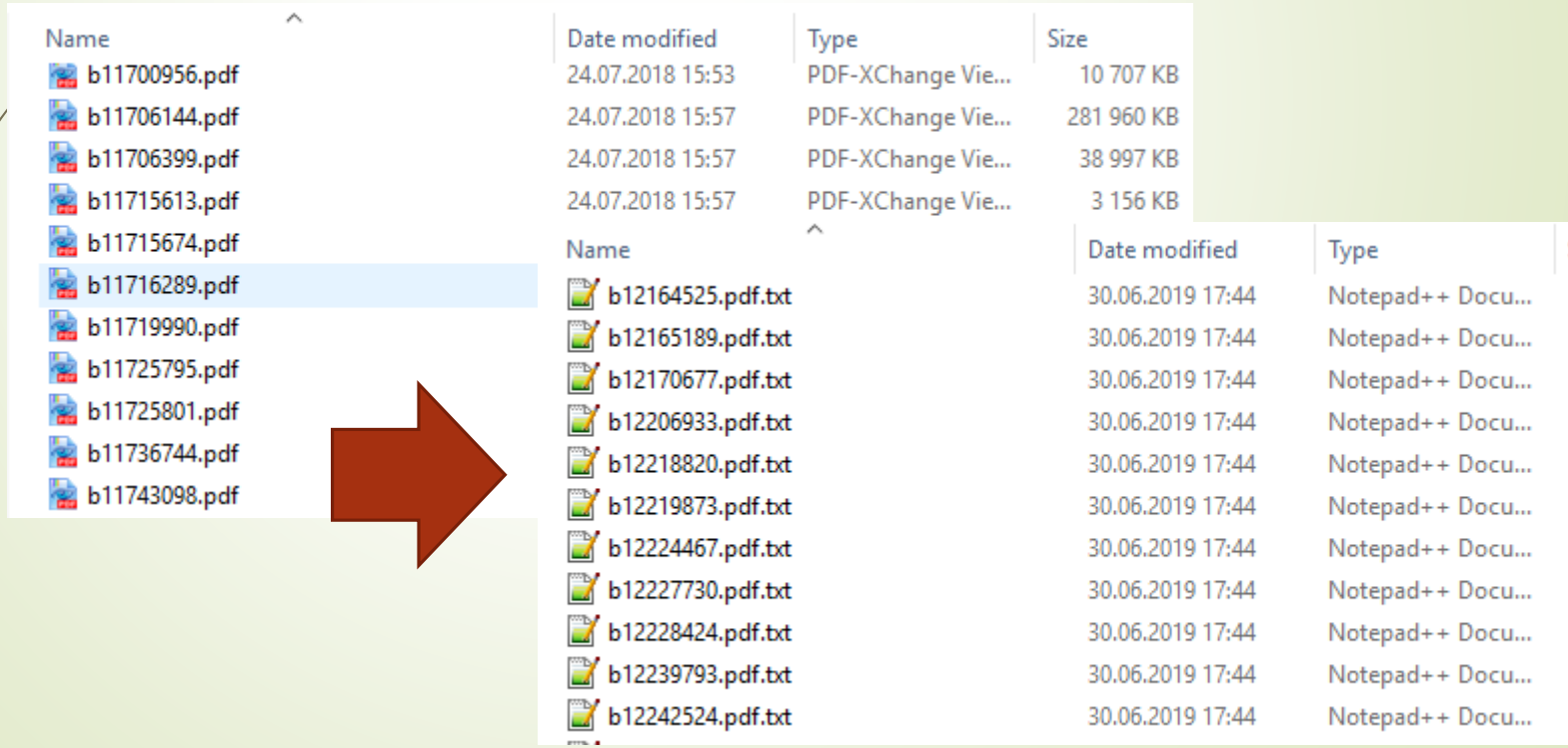
- A variety of interfaces and access points
  - Usually text mining is not an option they consider
  - But don't mind either





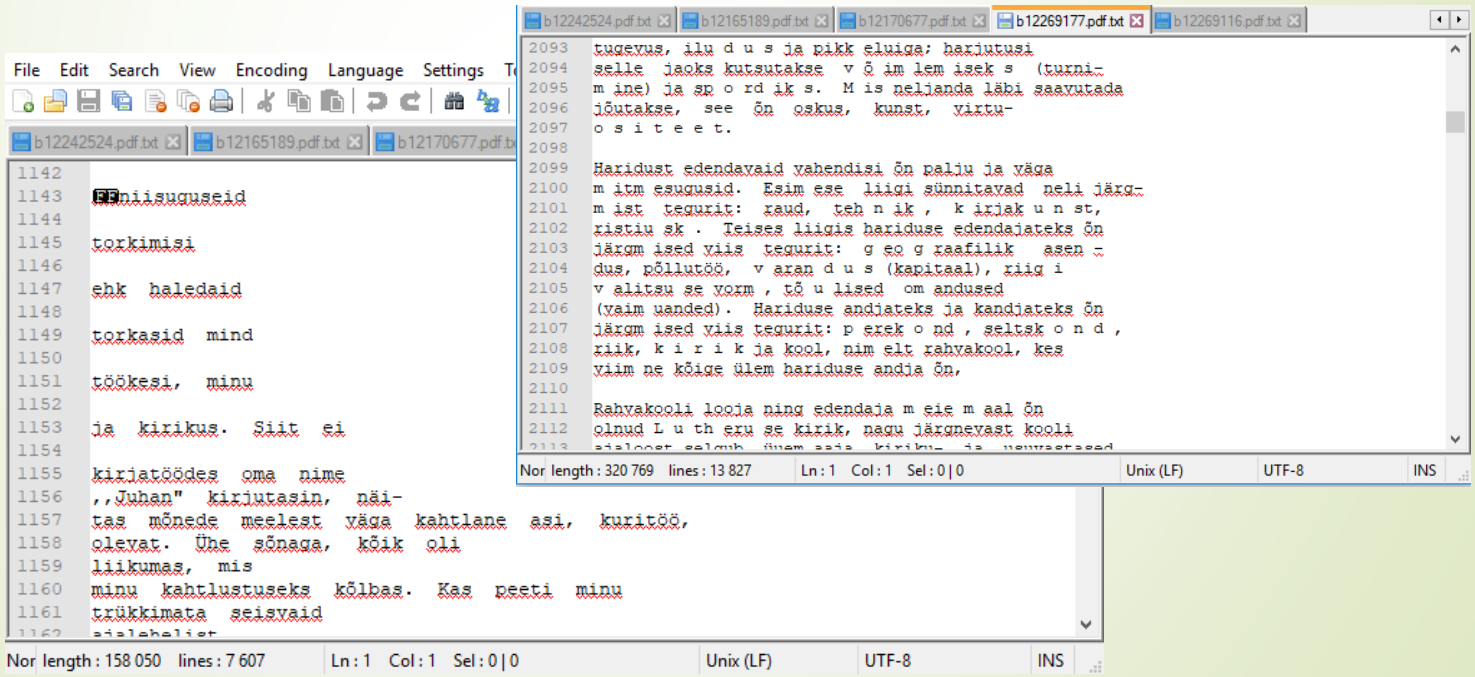
Getting the text

- Files of different format into text



Getting the text

- Finally, will have files that contain raw text
  - Quality will vary



Metadata

- Again in various formats
- Ought to be collected

```
pro0004___Bornhöhe, Eduard___"Tallinna narrid ja narrikesed"___1892___lk 3-12
pro0005___Bornhöhe, Eduard___"Willu wõitlused"___1890___lk 5-43
pro0006___Bornhöhe, Eduard___"Würst Gabriel"___1893___lk 229-240
pro0007___Eisen, M. J.___"Hiiu köster ja Saare kirikhärra"___1893___lk 43-52
pro0008___Eisen, M. J.___"Tartu saladused"___1891___lk 7-18
pro0009___[???]___"Hella"___1890___lk 3-12
pro0010___Hermann, K. A.___"Lapse mälestus"___1896___lk 1-8
pro0011___Hermann, K. A.___Rikka ja waese pulmad. Külajutukene.___1899___Lk. 29-37
pro0012___Hermann, K. A.___Uudisjutud Eesti rahwa elust.___1895___Lk. 37-45
pro0013___[???]___"Hugo ja Tekla". Jutustus Toolse (ranna) lossist ja ritterist, mis seal 15. aast
pro0014___Tüll, He                                  ema imelikud juhtumised wõera mere-saarte peal.___1891
pro0015___Jaanus,                                   olm juttu noorerahwa elust."___1893___lk 3-11
pro0016___Jakobson                                  899___lk 34-39
pro0017___Järw, J.                                  tustus Eesti minewikust."___1892___Lk. 127-136
```

Ein einfeltige weise zu Beten, fur einen guten freund Mart. Luther.

View/Open
r_iii_i_252i_10.pdf (17.68Mb)
Tekstituvastusega (1.431Mb)

Date
1535

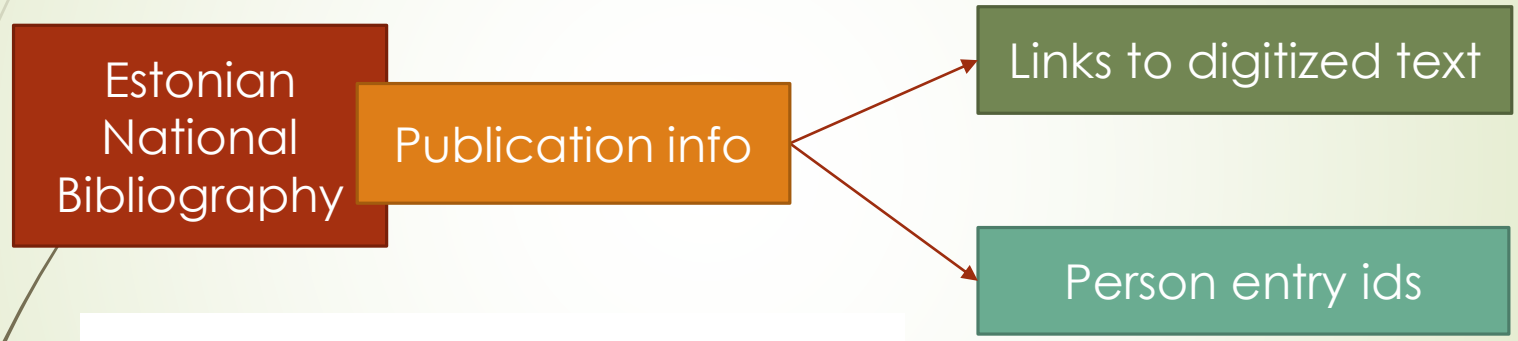
Author
Luther, Martin

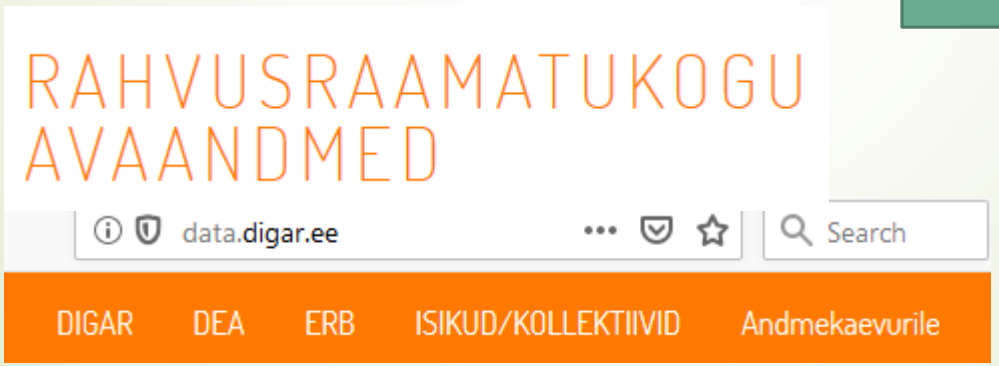
Metadata
Show full item record

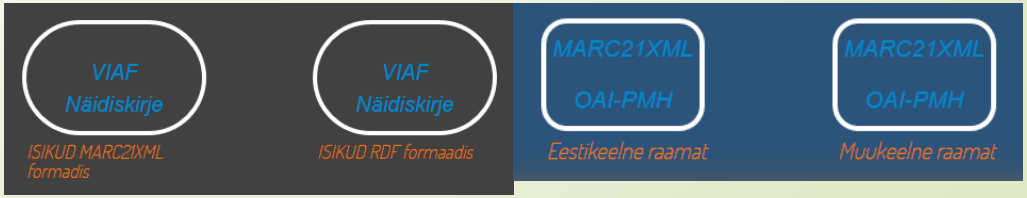
sordiaretaja
1.IX 1883 Harjumaal Kloostri valla Karilepa-Tõnul. Taluperemees Siim A., Liisa Vrager. Ae 1917 Anna Maria Volmer. Ants (1918, vt), Valve Jaagus (1920, vt), Ilmar (1927--92, vt). Vasalemma vallakool 1893--97, Paldiski algkool 1897--1900, Haapsalu linnakool 1902--03, Peterburi linnukasvatuskursus 1911--12, põllumajandusdoktor 1947. Sõjaväes Peterburis, vangistatud 1905--06, I maailmasõjas suurtükiväeametnik Tallinnas 1914--16, Eesti Sõjaväelaste Keskbüroo juhatuse liige, Harju maakonnanõukogu liige ja sekretär 1917--20, 1918 Eesti diviisi staabi mobilisatsiooniosakonna asjaajaja, Vabadussõjas ülemjuhataja staabi majandusülem, Eesti Sordiparanduse Seltsi Jõgeva sordikasvatuse osakonnajuhataja 1920--50, "Väikelooma-kasvataja" toimetaja 1919--21, ENSV TA korrespondentliige 1946, ÜN 1947—50 (II ks); ENSV teeneline teadlane 1945, NE preemia 1947, Stalini preemia 1948. Surnud 19.I 1950 Jõgeval, maetud Tartu Raadi kalmistule. EAT, EE, EBLI, EAT2, ENE1, ENE2, EABL, VjV, ENE2(14), ETeadBL – ISOTAMM 2

Metadata

- General indexes (e.g. Estonian National Bibliography)

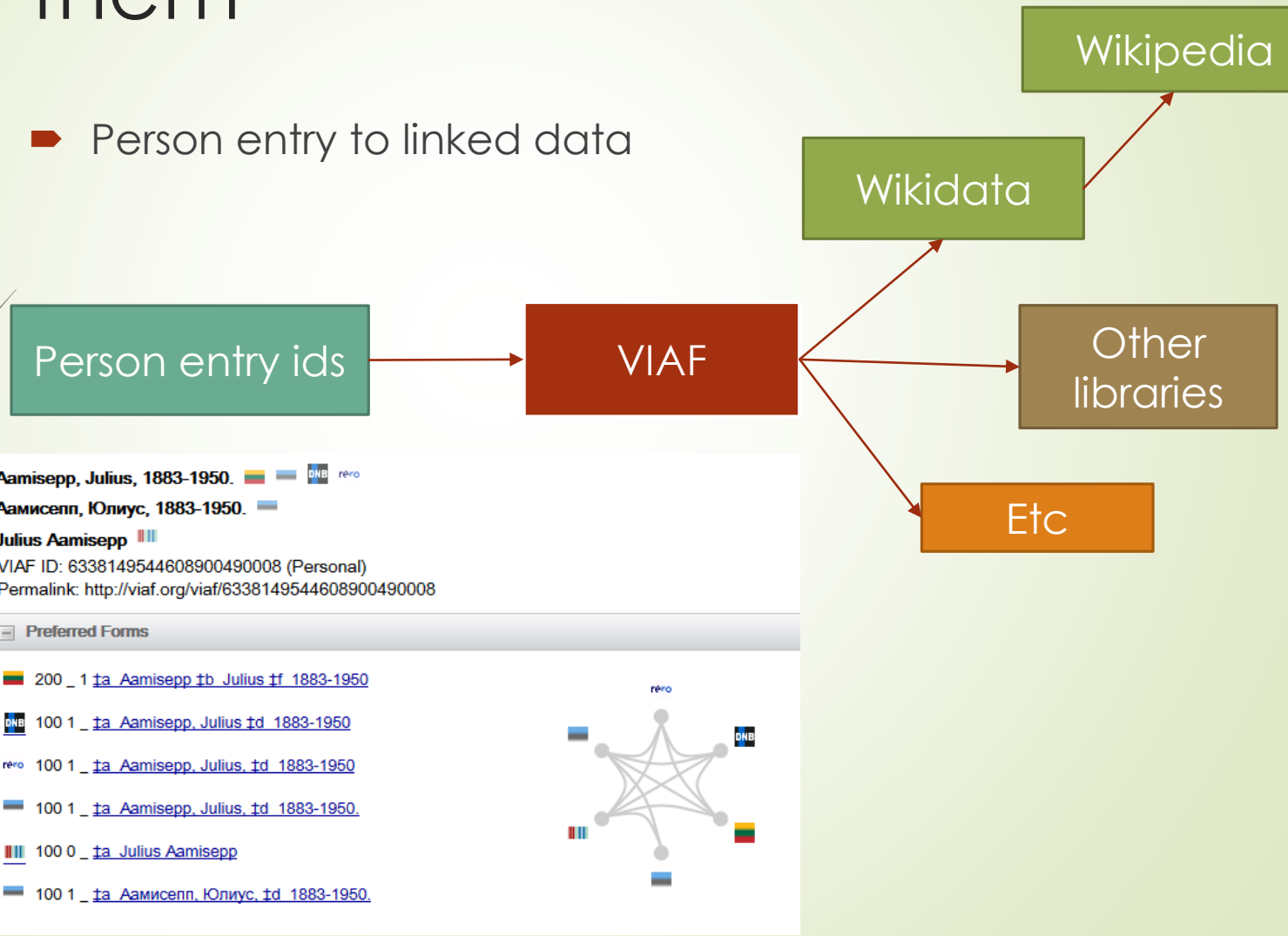




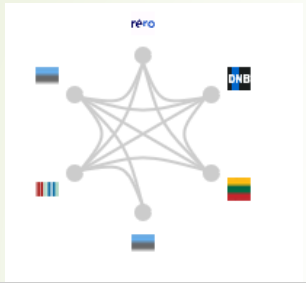


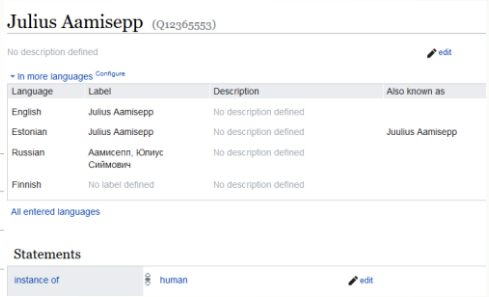
Collecting them all, linking them

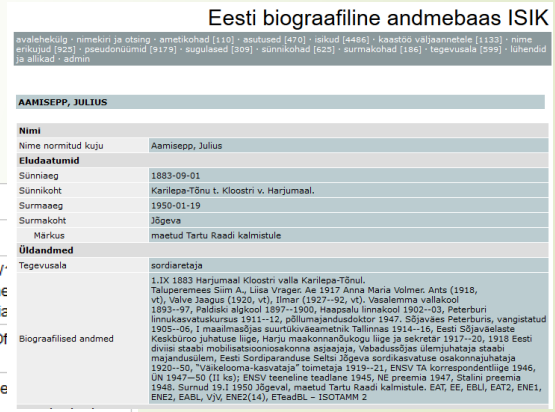
- Person entry to linked data



Aggregating the data





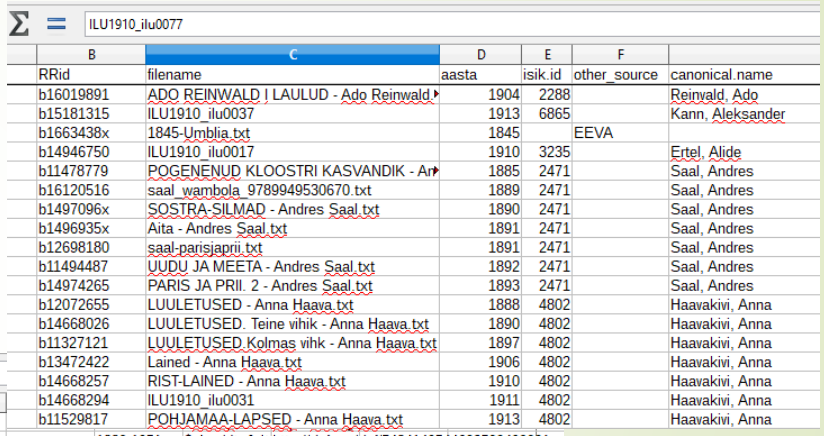


001		ERRR|a11124581 (VIAF cluster)
003		ERRR
008		990322|n|adnnnaabn || |a|
035		‡a (ERRR)a11124581
040		‡a ErRR ‡b est ‡c ErRR ‡d ErTUR
043		‡c ee
100	1	‡a Aamisepp, Julius, ‡d 1883-1950.
400	1	‡a Aamissepp, Julius.
500	1	‡a Аамисепп, Юлиус, ‡d 1883-1950.
670		‡a EE, 1. kd., 1985; EE, 14. kd., 2000.
680		‡i Eesti põllumajandusteadlane, sordiaretaja, põllumaj

100	1	‡a Aamisepp, Julius ‡d 1883-1950
375		‡a 1 ‡2 iso5218
500	1	‡0 (DE-101)1096201100 ‡0 (DE-588)1096201100 ‡0 http://d-nb.info/gnd/ Aamisepp, Ilmar ‡d 1927-1992 ‡4 bezf ‡4 http://d-nb.info/standards/eleme /gnd#familialRelationship ‡w r ‡i Beziehung familiaer ‡e Beziehung familia
548		‡a 1883-1950 ‡4 datl ‡4 http://d-nb.info/standards/elementset/gnd#dateOf ‡w r ‡i Lebensdaten
548		‡a 01.09.1883-19.01.1950 ‡4 datx ‡4 http://d-nb.info/standards/elementse /gnd#dateOfBirthAndDeath ‡w r ‡i Exakte Lebensdaten
550		‡0 (DE-101)04203194X ‡0 (DE-588)4203194-1 ‡0 http://d-nb.info/gnd/4203194-1 ‡a Agrarwissenschaftler ‡4 berc ‡4 http://d-nb.info/standards/elementset /gnd#professionOrOccupation ‡w r ‡i Charakteristischer Beruf
550		‡0 (DE-101)041779436 ‡0 (DE-588)4177943-5 ‡0 http://d-nb.info/gnd/4177943-5 ‡a Revolutionär ‡4 beru ‡4 http://d-nb.info/standards/elementset/gnd#professionOrOccupation ‡w r ‡i Beruf
551		‡a Karilepa ‡4 ortg ‡4 http://d-nb.info/standards/elementset/gnd#placeOfBirth ‡w r ‡i Geburtsort ‡9 v:Gem. Lääne-Harju

Manual linking

- Manual linking



ILU1910_ilu0077

	B	C	D	E	F	
	RRid	filename	aasta	isik.id	other_source	canonical.name
	b16019891	ADO REINWALD	LAULUD - Ado Reinwald	1904	2288		Reinvald, Ado
	b15181315	ILU1910_ilu0037		1913	6865		Kann, Aleksander
	b1663438x	1845-Umblia.txt		1845		EEVA	
	b14946750	ILU1910_ilu0017		1910	3235		Ertel, Alide
	b11478779	POGENENUD KLOOSTRI KASVANDIK - An		1885	2471		Saal, Andres
	b16120516	saal_wambola_9789949530670.txt		1889	2471		Saal, Andres
	b1497096x	SOSTRA-SILMAD - Andres Saal.txt		1890	2471		Saal, Andres
	b1496935x	Aita - Andres Saal.txt		1891	2471		Saal, Andres
	b12698180	saal-parisjaprii.txt		1891	2471		Saal, Andres
	b11494487	UUDU JA MEETA - Andres Saal.txt		1892	2471		Saal, Andres
	b14974265	PARIS JA PRII. 2 - Andres Saal.txt		1893	2471		Saal, Andres
	b12072655	LUULETUSED - Anna Haava.txt		1888	4802		Haavakivi, Anna
	b14668026	LUULETUSED. Teine vihik - Anna Haava.txt		1890	4802		Haavakivi, Anna
	b11327121	LUULETUSED.Kolmas vihk - Anna Haava.txt		1897	4802		Haavakivi, Anna
	b13472422	Lained - Anna Haava.txt		1906	4802		Haavakivi, Anna
	b14668257	RIST-LAINED - Anna Haava.txt		1910	4802		Haavakivi, Anna
	b14668294	ILU1910_ilu0031		1911	4802		Haavakivi, Anna
	b11529817	POHJAMAA-LAPSED - Anna Haava.txt		1913	4802		Haavakivi, Anna

B11643

	A	B	C	D	E	F	G	H	I	J	K	L				
11640	TRUE								Lorup, Johannes	Lorup, Johannes						
11641	TRUE								Loskit, Juhan	Loskit, Juhan				1889-1951	$aLoskit, Juh	http://viaf.org/viaf/54841495446025004900001
11642	TRUE								Loskit, Karl	Loskit, Karl				1884-1936	$aLoskit, Ka	http://viaf.org/viaf/184149542838700301920
11643			ISIK	1908					Lossius, Eduard		Lossius, Edu	1811-09-05	1870-03-29			
11644	TRUE		ISIK	1908					Lossius, Eduard	Lossius, Eduard				1811-1870	$aLossius, E	http://viaf.org/viaf/15540624
11645	TRUE								Losski, Nikolai	Losski, Nikolai				1870-1965	$aLosski, Ni	http://viaf.org/viaf/24696091
11646			VEPP	1549					Lossman, Helmar		Lossman, He	07.02.1903	21.08.1961			
11647			VEPP	4083					Lossman, Jüri		Lossman, Jü	04.02.1891	01.05.1984			
11648	TRUE		VEPP	4084					Lossmann, Arthur	Lossmann, Arthur				1877-1972	$aLossmann	http://viaf.org/viaf/304960182
11649			VEPP	4084					Lossmann, Artur		Lossmann, A	05.10.1877	01.09.1972			
11650			ISIK	1909					Lossmann, August~Postimees 1857-1907		Lossmann, A	1871-12-26				
11651	TRUE								Lossmann, J	Lossmann, J					$aLossmann, J.	
11652			ISIK	1910					Lossmann, Jaan		Lossmann, J	1866-00-00				
11653	TRUE	unsure	ISIK	1910					Lossmann, Jaan	Lossmann, Jaan					$aLossmann, Jaan	
11654	TRUE								Lossmann, Jüri	Lossmann, Jüri				1891-1984	$aLossmann	http://viaf.org/viaf/582814954460490049006
11655			VEPP	3238					Lossmann, Peeter		Lossmann, P	24.03.1919	08.10.2013			
11656			ISIK	1911					Lotarjov, Igor Vassiljevitš~Eesti kirjanike le		Lotarjov, Igor	1887-05-16	1941-12-20			

Harmonization

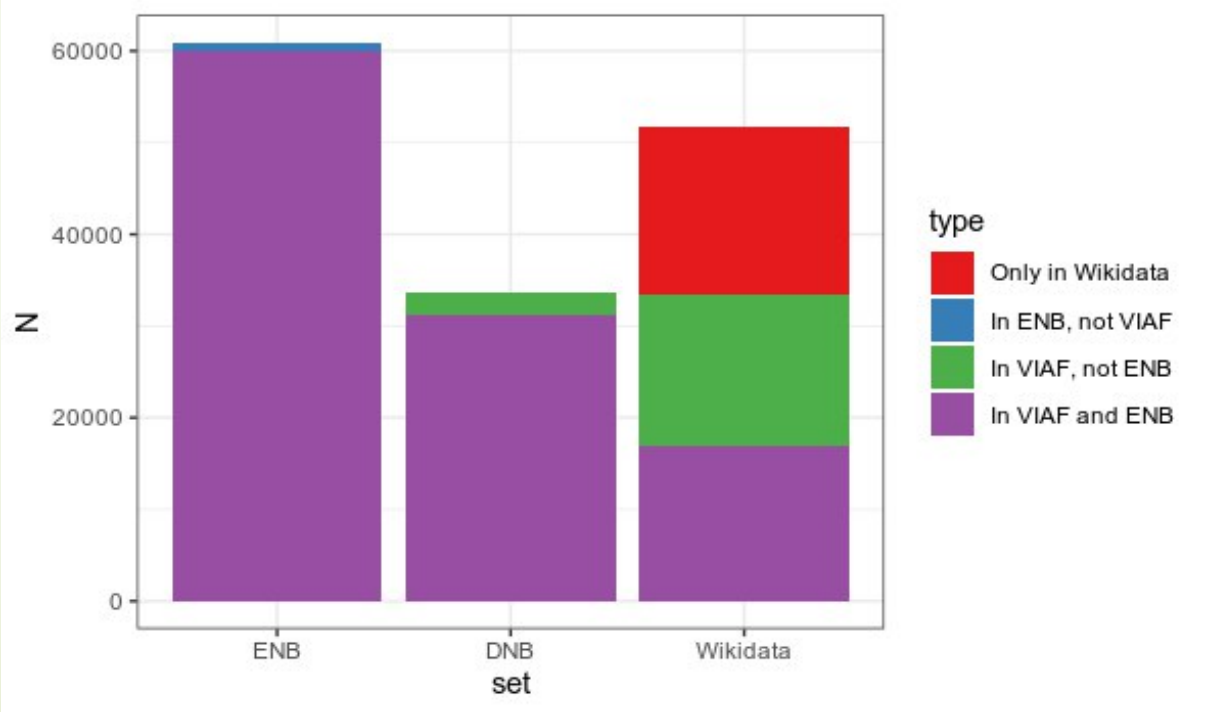
- For archival puproses, accuracy is key.
- For analysis, data harmonization is often needed.

```
183   works[kirjastus=="w.bormi pär.",kir
184   works[kirjastus=="f.w.borm",kirjastu
185   works[str_detect(kirjastus,"noor ee
186   works[str_detect(kirjastus,"mutsu"),
187   works[str_detect(kirjastus,"ploompu
188   works[str_detect(kirjastus,"postimeh
189   works[str_detect(kirjastus,"postimee
190   works[str_detect(kirjastus,"h\\.laak
191   works[str_detect(kirjastus,"eesti ki
192   works[kirjastus=="eesti kirjandusese
193   works[kirjastus=="hermann",kirjastus
194   works[kirjastus=="hermann'i raamatuk
195   works[kirjastus=="pealadu hermanni rmtkpl.",kirjastus:="k.a.hermann"]
196   works[kirjastus=="hermanni raamatukauplus",kirjastus:="k.a.hermann"]
197   works[kirjastus=="hermann'i rmtkpl.",kirjastus:="k.a.hermann"]
198   works[kirjastus=="pealadu hermanni kaupluses",kirjastus:="k.a.hermann"]
199   works[kirjastus=="leoke'se kirjastus",kirjastus:="h.leoke"]
200   works[kirjastus=="leoke",kirjastus:="h.leoke"]
201   works[kirjastus=="leoke'se raamatuäri antikvariaat",kirjastus:="h.leoke"]
```

```
29    #works <- works#[!is.na(koht),.N,by=koht]
30    works <- works[,koht:=str_replace_all(koht,"s$","")]
31    works <- works[,koht:=str_replace_all(koht,"l$","")]
32    #works <- works[!duplicated(works)]
33    #places[places!=""&places!="S.l."&places!="S.l.,"&places!="S. l."&places!="S. l.,"]
34    works <- works[koht!=""&koht!="S.l."&koht!="S.l.,"&koht!="S. l."&koht!="S. l.,"]
35    works[str_detect(koht,"Paide"),koht:="Paide"]
36    works[str_detect(koht,"Weissenstein"),koht:="Paide"]
37    works[str_detect(koht,"Вейсенштейн"),koht:="Paide"]
38    works[str_detect(koht,"Вейссенштейн"),koht:="Paide"]
39    works[str_detect(koht,"Haapsalu"),koht:="Haapsalu"]
40    works[str_detect(koht,"Hapsa"),koht:="Haapsalu"]
41    works[str_detect(koht,"Гапсаль"),koht:="Haapsalu"]
42
43    works[str_detect(koht,"Keila"),koht:="Keila"]
44    works[str_detect(koht,"Rakvere"),koht:="Rakvere"]
45    works[str_detect(koht,"Wesenberg"),koht:="Rakvere"]
46    works[str_detect(koht,"Везенберг"),koht:="Rakvere"]
47
```

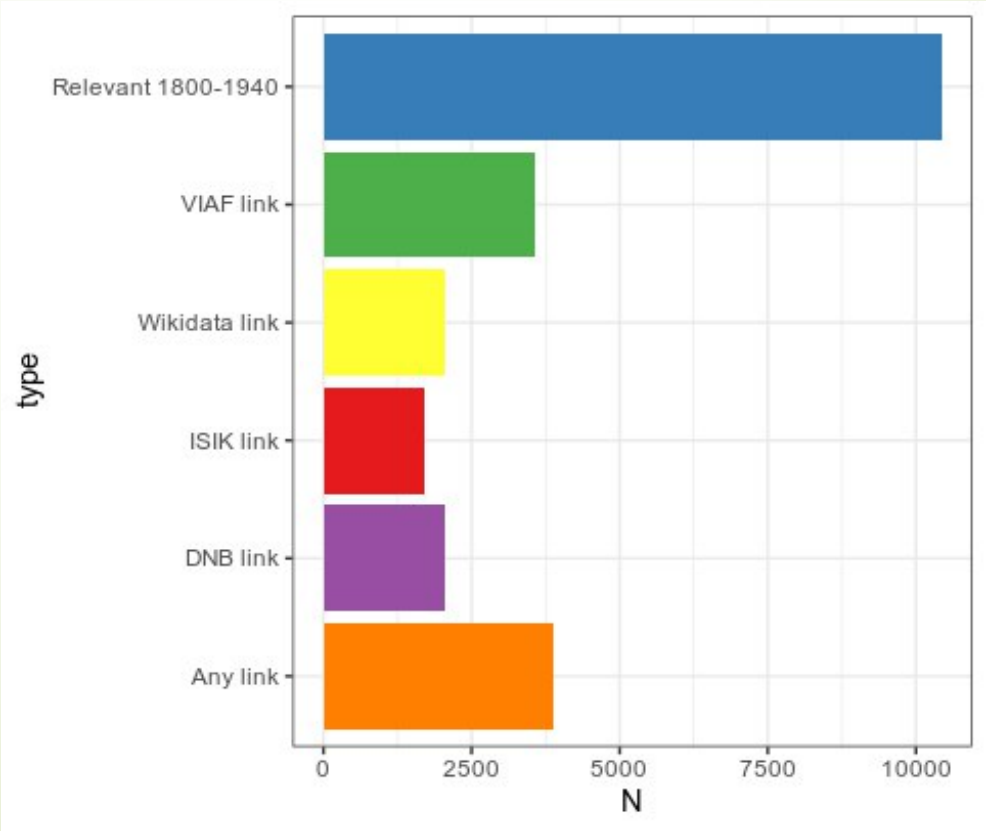
Results of linking

- Linkability of archives



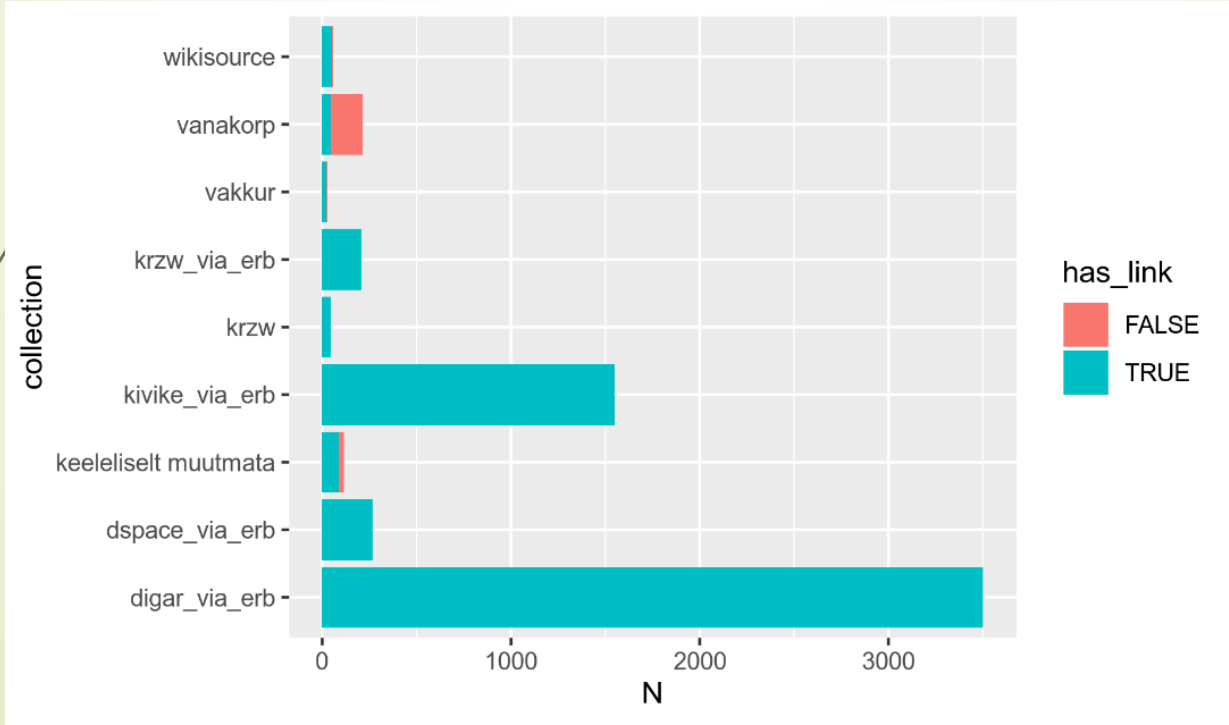
Results of linking

- Linked author metainformation



Results of text collection

- 5132 texts total, of 4186 unique publications



Processing the corpus

- Raw texts can be processed
- Lemmatize, POS-tag etc (used Python EstNLTK)

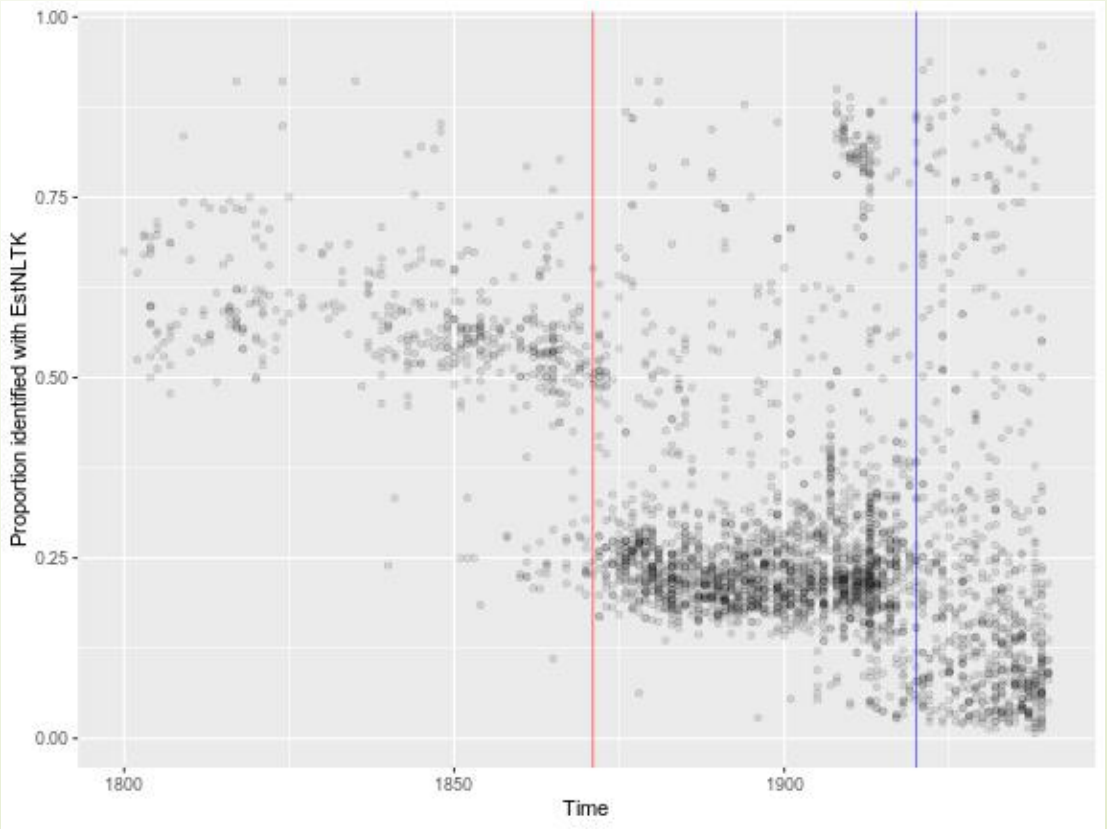
```
from estnltk import Text
text = Text('mõeldud')
text.tag_analysis()
```

```
[('Usjas', 'A', 'omadussõna algvõrre'),
 ('kaslane', 'S', 'nimisõna'),
 ('ründas', 'V', 'tegusõna'),
 ('künklikul', 'A', 'omadussõna algvõrre'),
 ('maastikul', 'S', 'nimisõna'),
 ('tünjat', 'A', 'omadussõna algvõrre'),
 ('Tallinnfilmi', 'H', 'pärisnimi'),
 ('režissööri', 'S', 'nimisõna')]
```

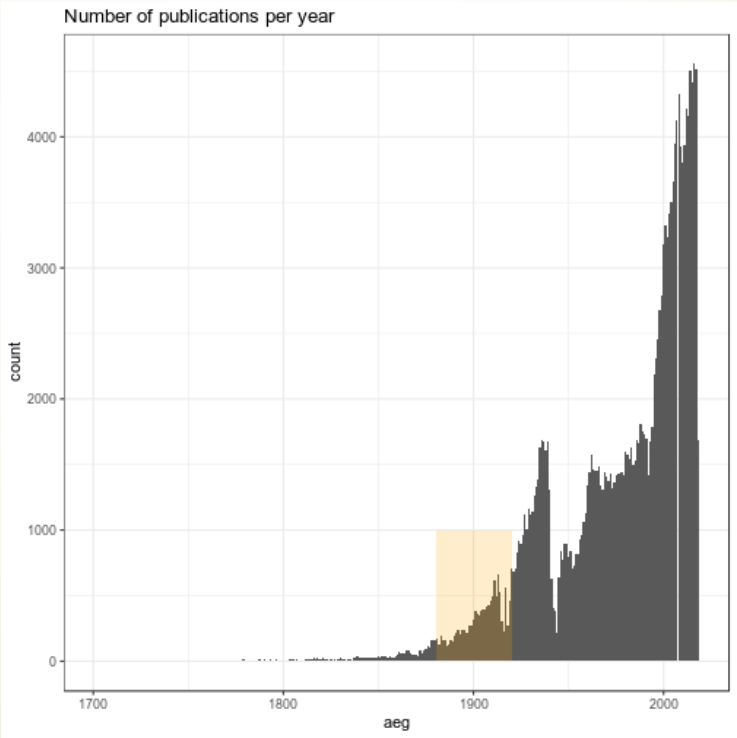

```
{'paragraphs': [{'end': 7, 'start': 0}],
 'sentences': [{'end': 7, 'start': 0}],
 'text': 'mõeldud',
 'words': [{'analysis': [{'clitic': '',
     'ending': '0',
     'form': '',
     'lemma': 'mõeldud',
     'partofspeech': 'A',
     'root': 'mõel=dud',
     'root_tokens': ['mõeldud']},
    {'clitic': '',
     'ending': '0',
     'form': 'sg n',
     'lemma': 'mõeldud',
     'partofspeech': 'A',
     'root': 'mõel=dud',
     'root_tokens': ['mõeldud']},
    {'clitic': '',
     'ending': 'd',
     'form': 'pl n',
     'lemma': 'mõeldud',
```

Results of the corpus

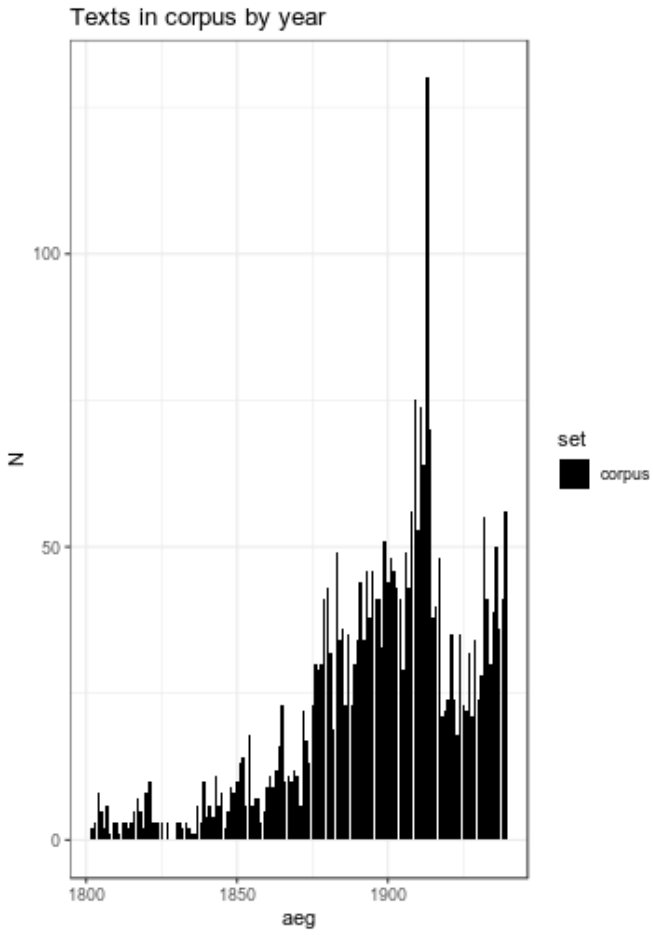
- NLP success rate over time

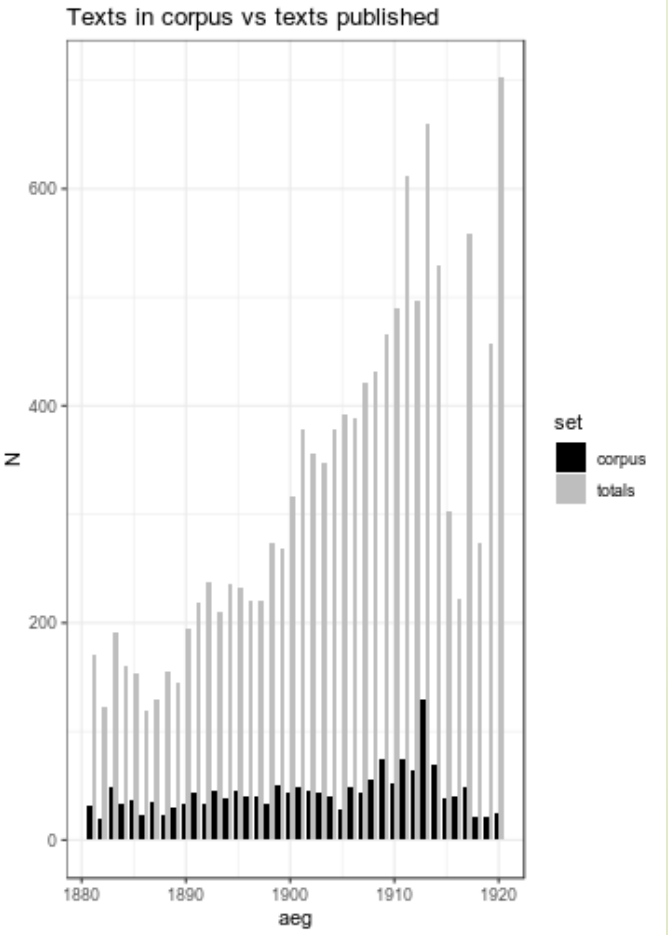


The published texts



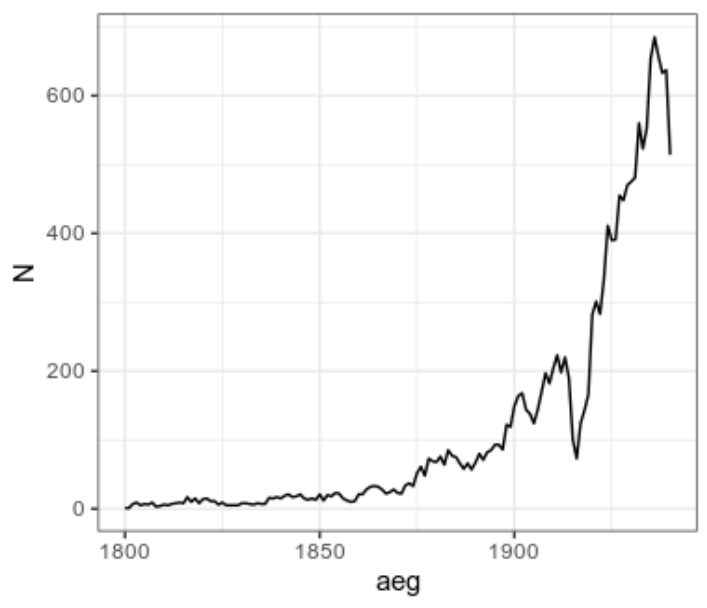
Representativeness

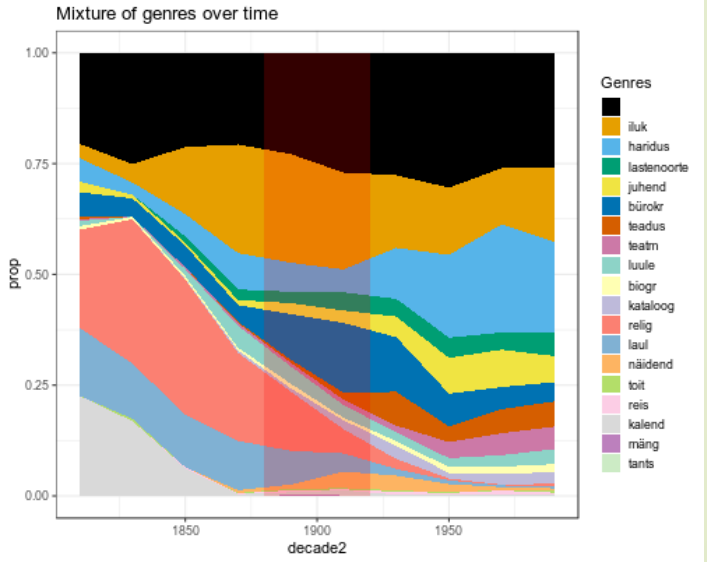


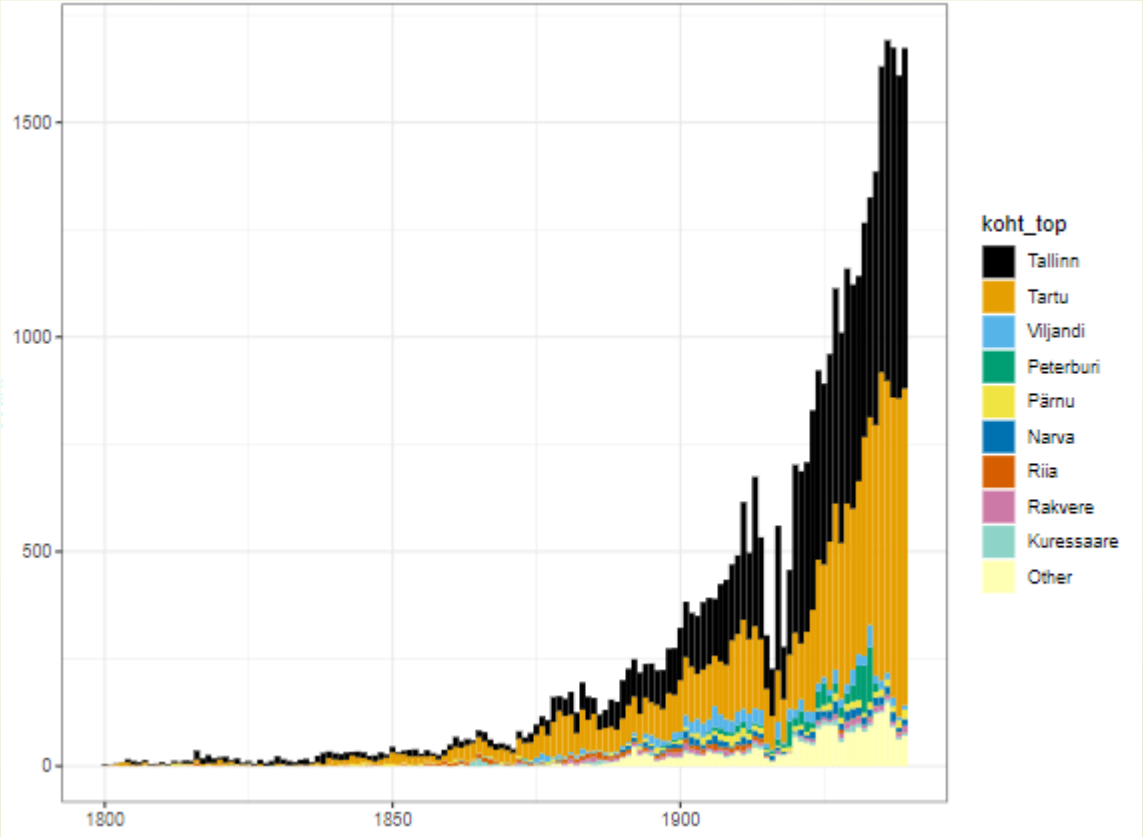


Publishing industry

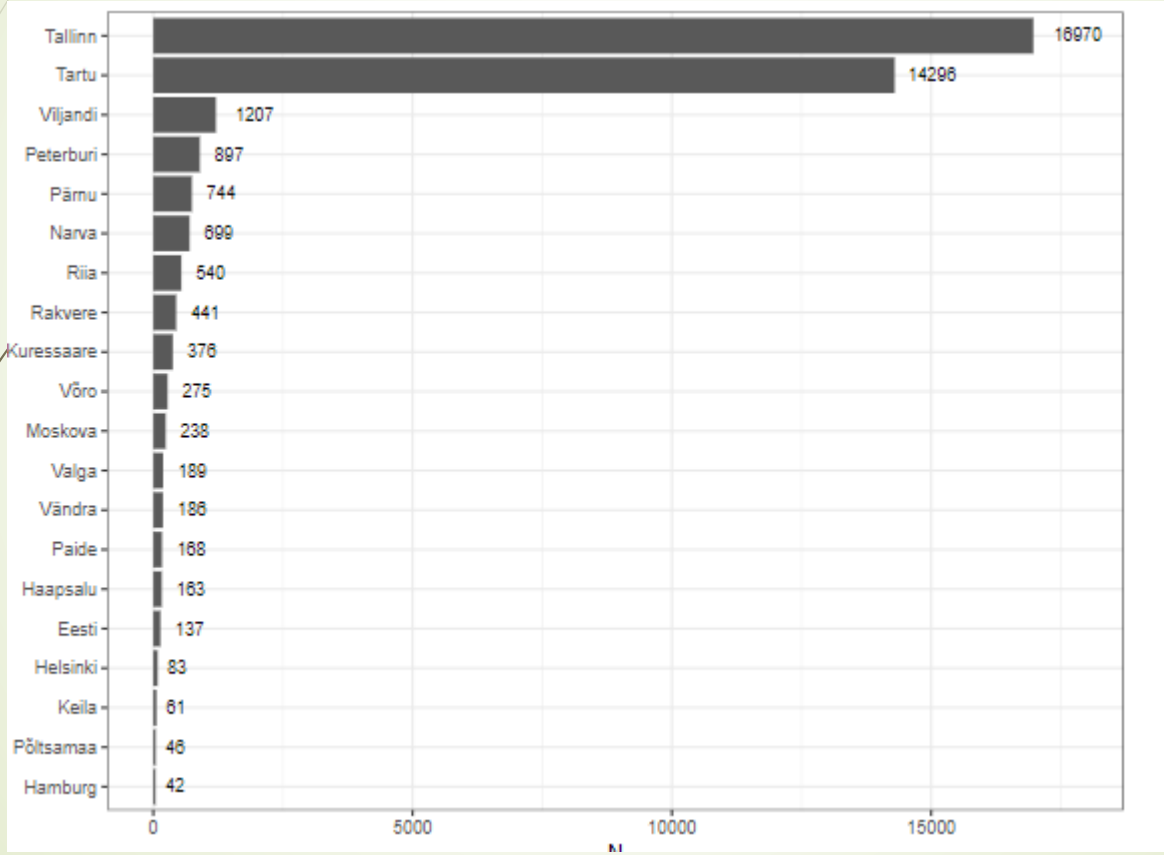
- More authors and open genres





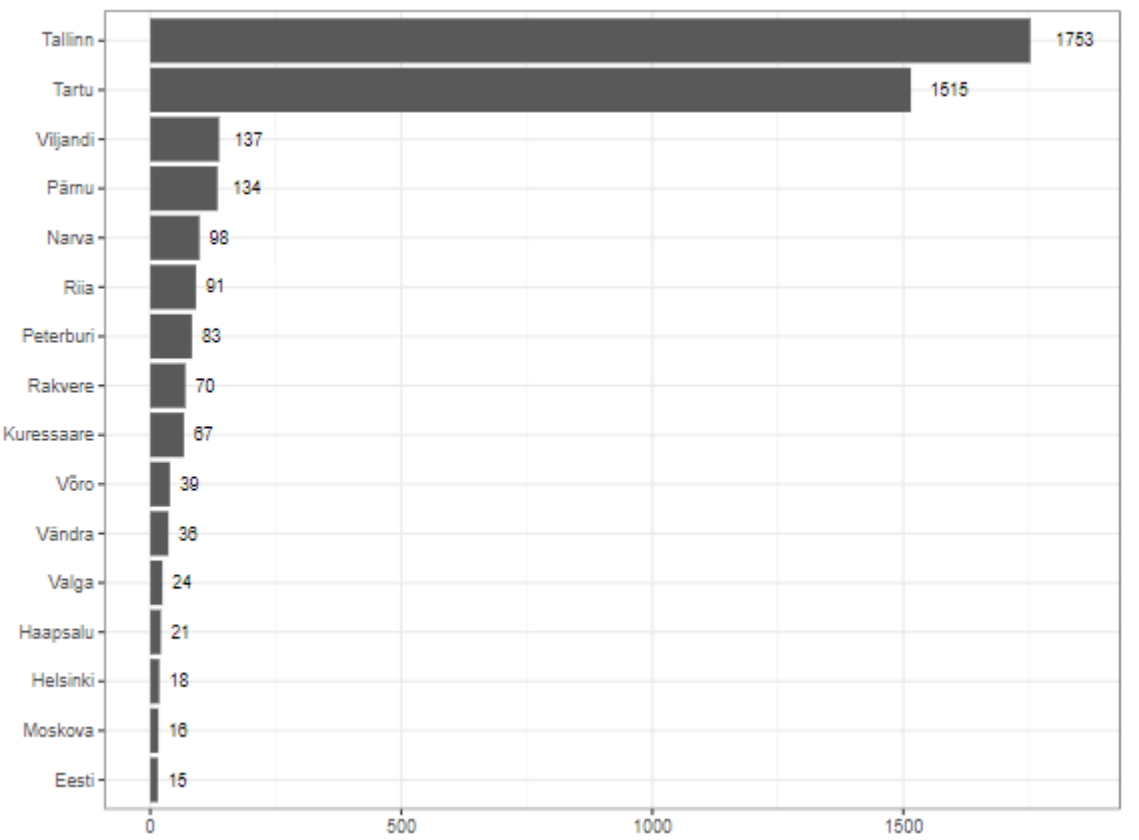

1500
1000
500
0
1800
1850
1900
koht_top
Tallinn
Tartu
Viljandi
Peterburi
Pärnu
Narva
Riia
Rakvere
Kuressaare
Other

A few cities dominate



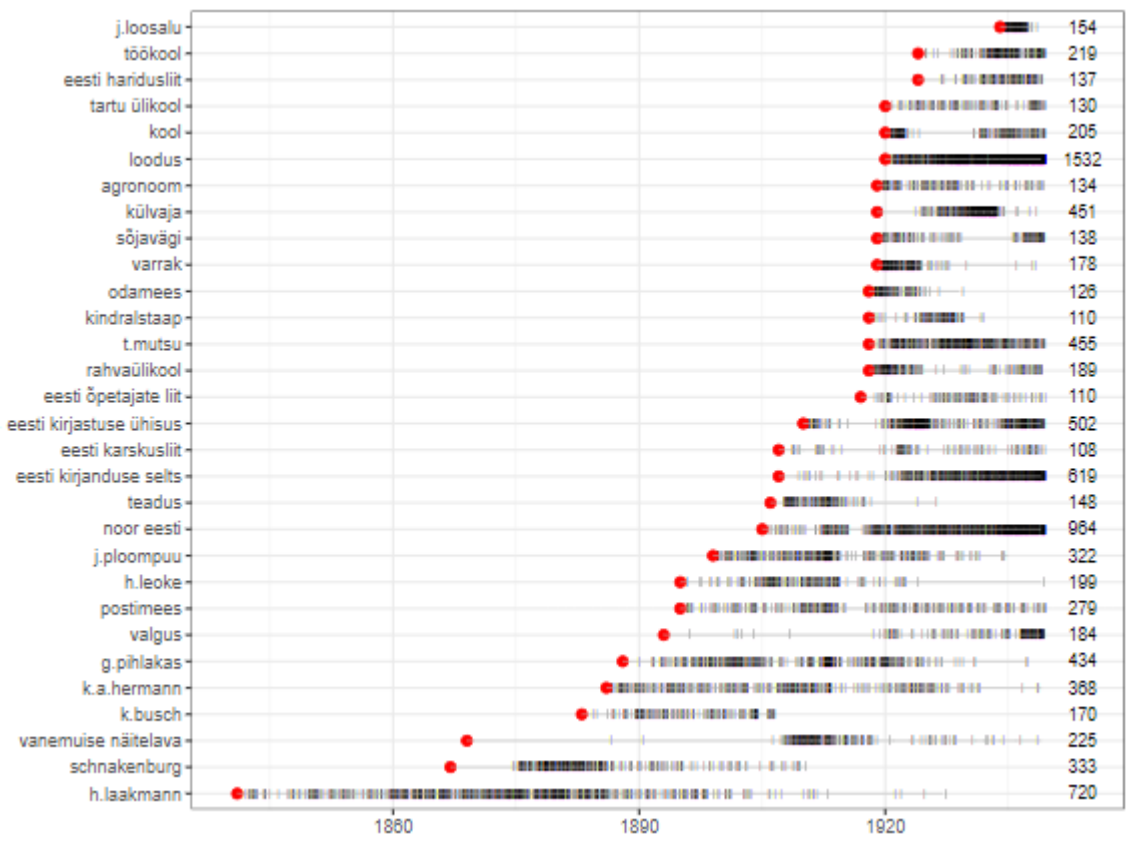
In corpus

- s


Tallinn
Tartu
Viljandi
Pärnu
Narva
Riia
Peterburi
Rakvere
Kuressaare
Võro
Vändra
Valga
Haapsalu
Helsinki
Moskova
Eesti
1753
1515
137
134
98
91
83
70
67
39
36
24
21
18
16
15
0
500
1000
1500

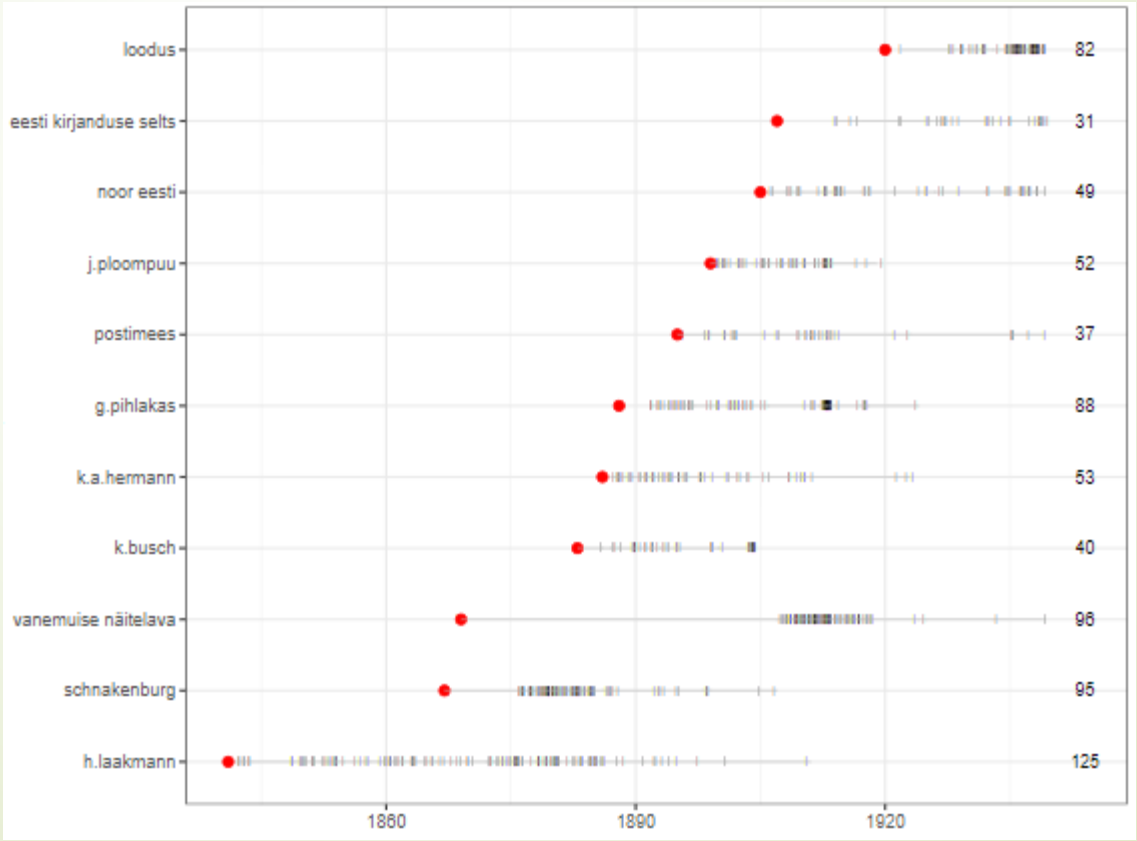
Publishers

- Top publishers (harmonized)



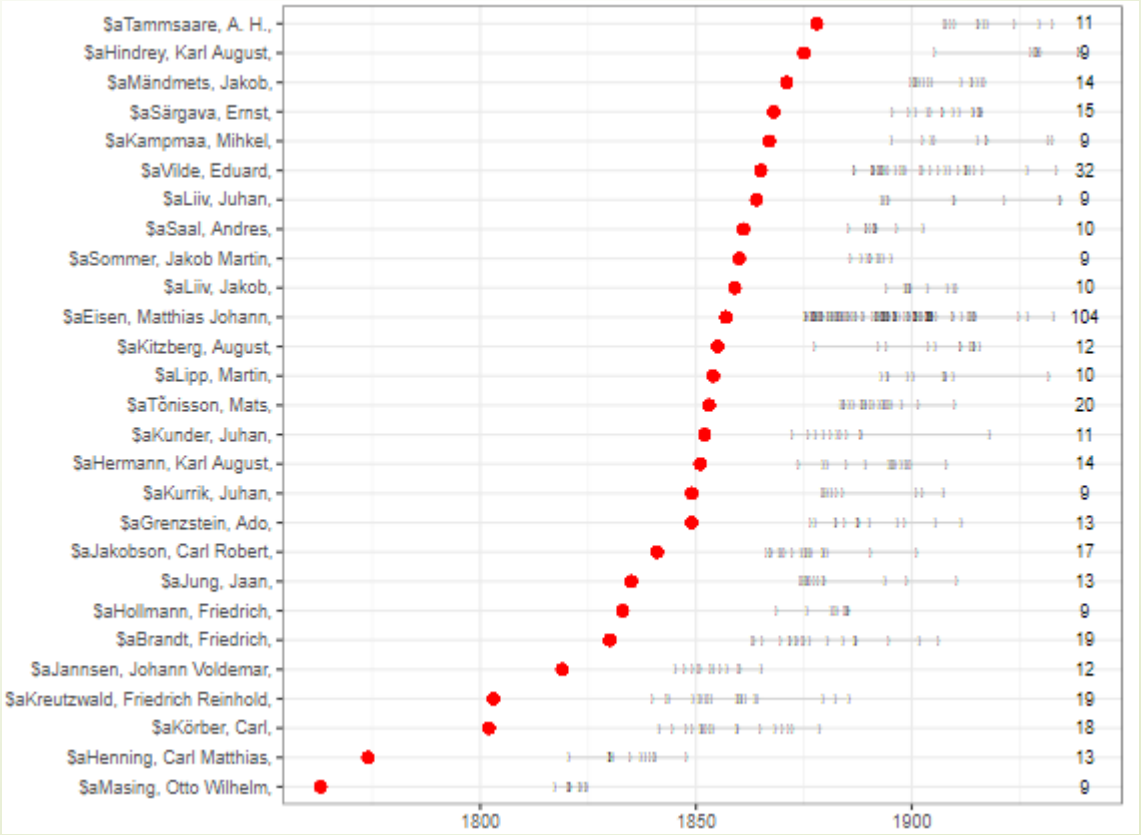
Publishers

- Top publishers (harmonized) within corpus



Authors

- Top authors (harmonized) within corpus



Other types of data

- There is a lot of relevant scholarship.
- Just thinking about it in terms of „data“ is new.

- Needs creativity in finding and utilizing the data

Biographic data

- Assemble life histories
  - from various sources

WIKIDATA

VIKIPEEDIA
Vaba entsüklopeedia

EESTI RAHVUS-RAAMATUKOGU

DEUTSCHE NATIONAL BIBLIOTHEK

Eesti biograafiline andmebaas ISIK

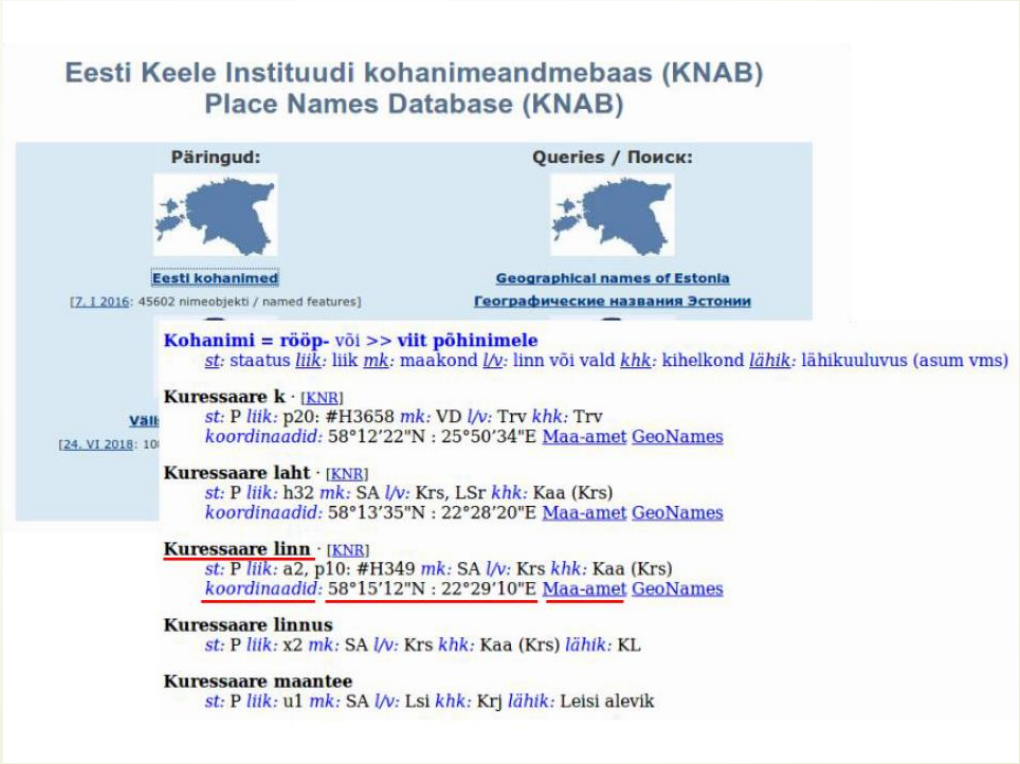
avalehekülg · nimekiri ja otsing · ametikohad [108] · asutused [468] · isikud [4485] · kaastöö väljaannetele [1131] · nime erikujud [925] · pseudonüümid [9179] · sugulased [309] · sünnikohad [624] · surmakohad [186] · tegevusala [599] · lühendid ja allikad · admin

SAAL, ANDRES

Nimi	
Nime normitud kuju	Saal, Andres
Allikad	Postimees 1857-1907, 235-236; EAE-40, lk. 200, EE (2000), XIV, lk. 447.
Eludaatumid	
Sünniaeg	1861-05-21
Sünnikoht	Selja k. Tori v. Pärnumaa
Surmaaeg	1931-06-23
Surmakoht	Los Angeles
Üldandmed	
Tegevusala	kirjanik, õpetaja
Biograafilised andmed	SAAL, Andres. 21.V 1861 Tori khk Tori v Laasil, Taluom (ms puusepp) Jaak S. (1817--1902), Ano Lindebaum (ms toatüdr, 1818--92). Ae Emilie Rosalie Moks (1871--1954). Rosa Regina Boiley (1902--30, maalikunstn), Leo Henry Wladimir (1904--65, ins). Selja v-k, Tori khkk, vallak.õp eksam 1880, TU vabakuulaja 1886--89, Cronenbergi fototehnikak Baieris, Viini paljundustehnikainst, Selja v-k õp 1880--84, Oleviku toim.l 1884--90 ja tsinkograaf 1893--97, Trt Jaani kir.k õp 1884--90, toiduainete kaupmees Trt-s 1890--93, fotograaf Frankfurdis 1897--98, topograaf Jaaval 1898--1920, Los Angeleses 1920--31, jäi pimedaks 1928. Kirjanik. Srn 23.VI 1931 Hollywoodis, tuhastati (urn 1932 Trt Maarja kir-s, 1944 kirj.muus-s). PmA, EBL, EE, EBLI, ENE1, EKBL, EKrL, ENE2, EKoBL, ENE2(14)

Person	Aamisepp, Julius
Geschlecht	männlich
Quelle	Pressearchiv des Herder-Instituts Marburg Wikipedia (Stand: 25.09.2018): https://en.wikipedia.org/wiki/Julius_Aamisepp
Zeit	Lebensdaten: 1883-1950
Land	Estland (XA-EE)
Geografischer Bezug	Geburtsort: Karilepa (Gem. Lääne-Harju) Sterbeort: Jõgeva
Beruf(e)	Agrarwissenschaftler Revolutionär

Placename locations



Eesti Keele Instituudi kohanimeandmebaas (KNAB)
Place Names Database (KNAB)

Päringud:

Eesti kohanimed

[7. I 2016: 45602 nimeobjekti / named features]

Queries / Поиск:

Geographical names of Estonia
Географические названия Эстонии

Kohanimi = rööp- või **>> viit põhinimele**
st: staatus *liik:* liik *mk:* maakond *l/v:* linn või vald *khk:* kihelkond *lähik:* lähikuuluvus (asum vms)

Kuressaare k · [KNR]
st: P *liik:* p20: #H3658 *mk:* VD *l/v:* Trv *khk:* Trv
koordinaadid: 58°12'22"N : 25°50'34"E Maa-amet GeoNames

Kuressaare laht · [KNR]
st: P *liik:* h32 *mk:* SA *l/v:* Krs, LSr *khk:* Kaa (Krs)
koordinaadid: 58°13'35"N : 22°28'20"E Maa-amet GeoNames

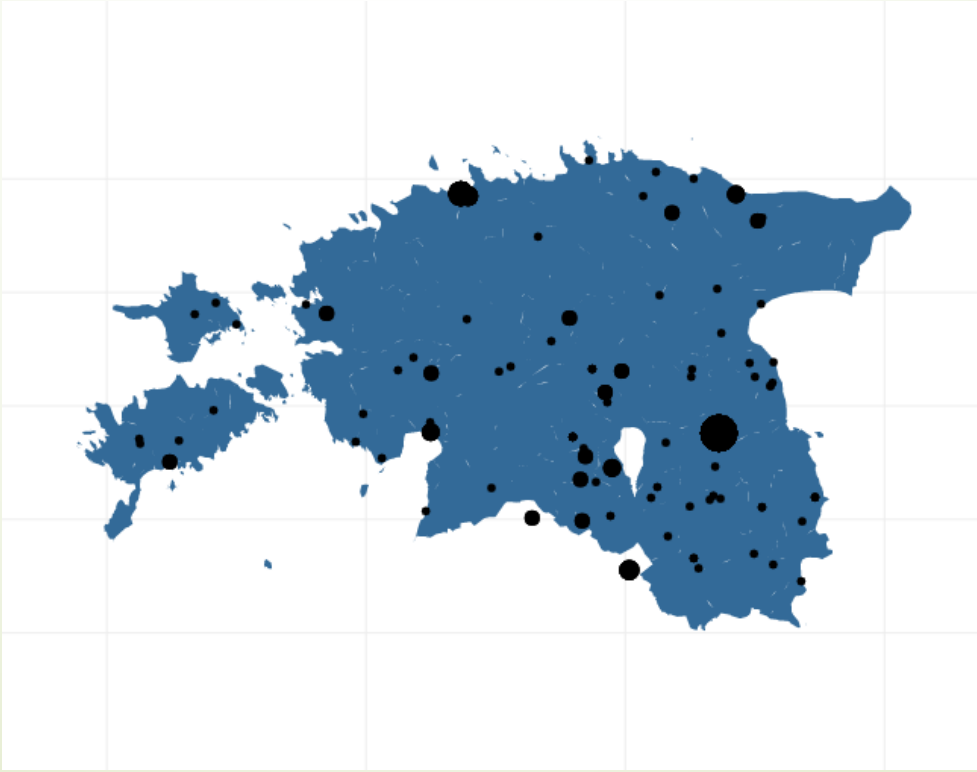
Kuressaare linn · [KNR]
st: P *liik:* a2, p10: #H349 *mk:* SA *l/v:* Krs *khk:* Kaa (Krs)
koordinaadid: 58°15'12"N : 22°29'10"E Maa-amet GeoNames

Kuressaare linnus
st: P *liik:* x2 *mk:* SA *l/v:* Krs *khk:* Kaa (Krs) *lähik:* KL

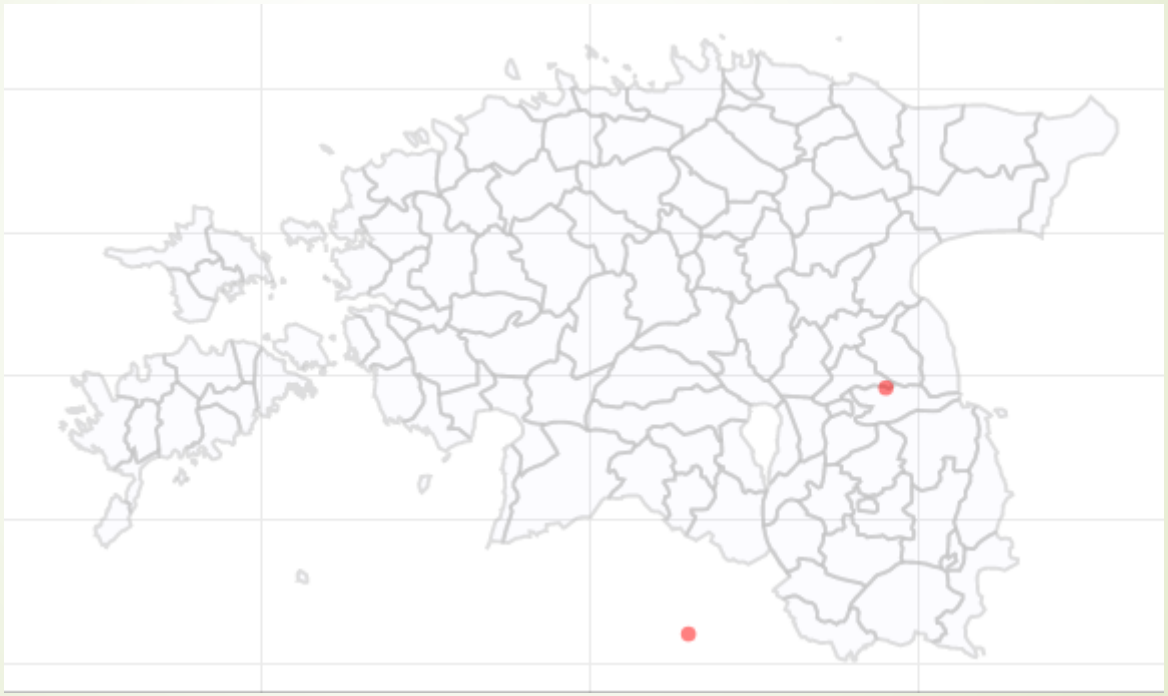
Kuressaare maantee
st: P *liik:* u1 *mk:* SA *l/v:* Lsi *khk:* Krj *lähik:* Leisi alevik

Birthplaces of writers

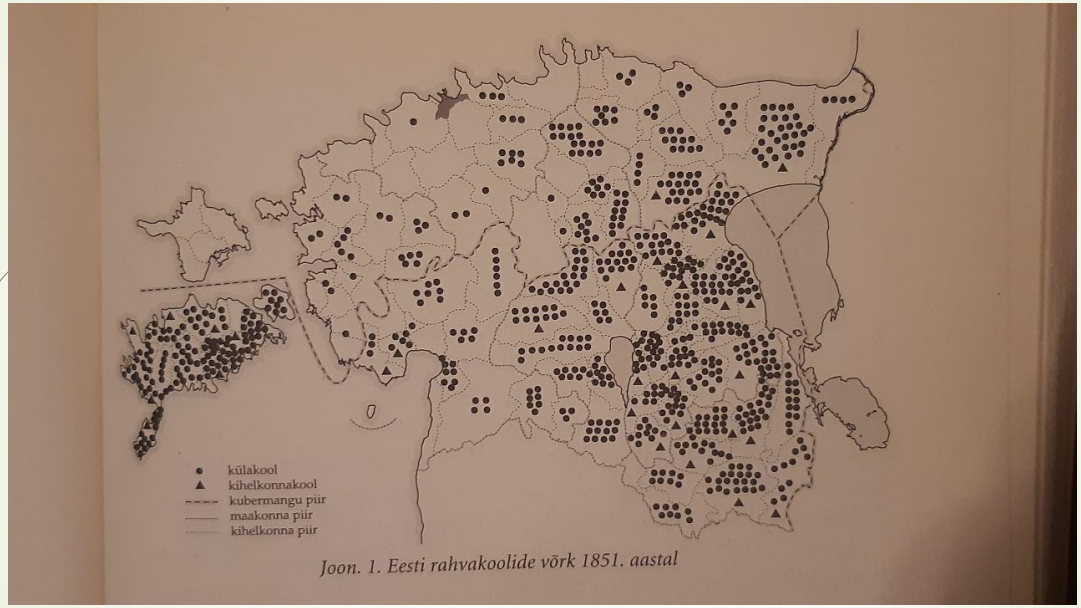
- Some birthplaces of writers in corpus



Birthplaces by decade

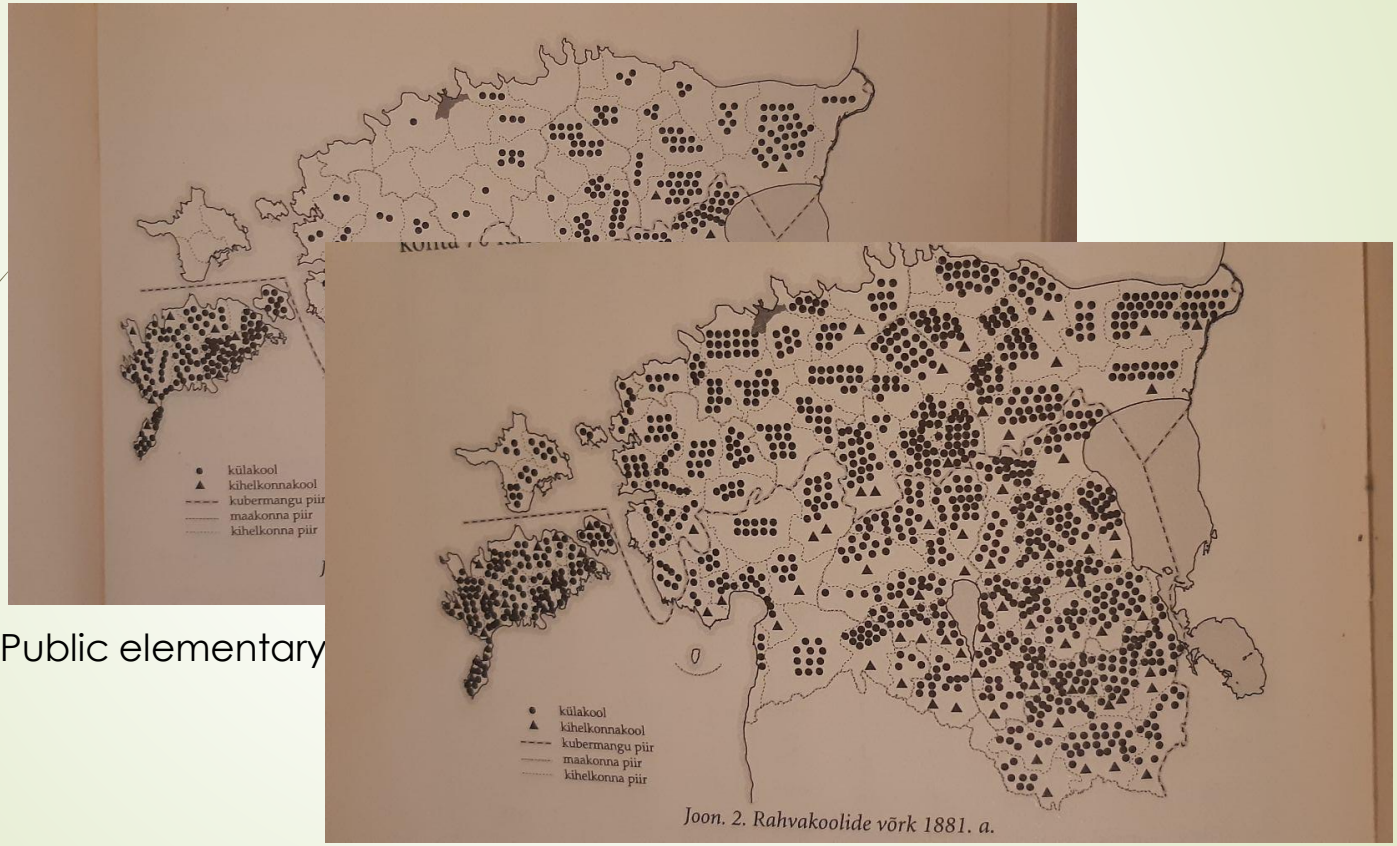


Education by decade



Public elementary school network in 1851

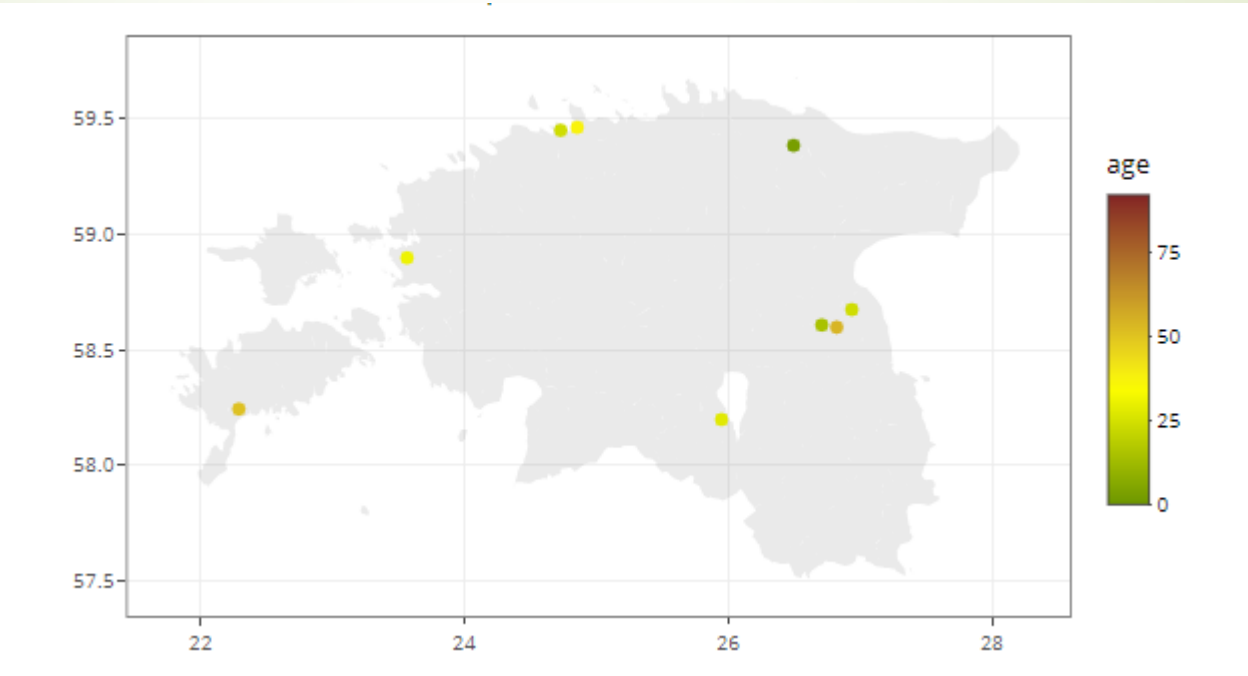
Education by decade



Public elementary

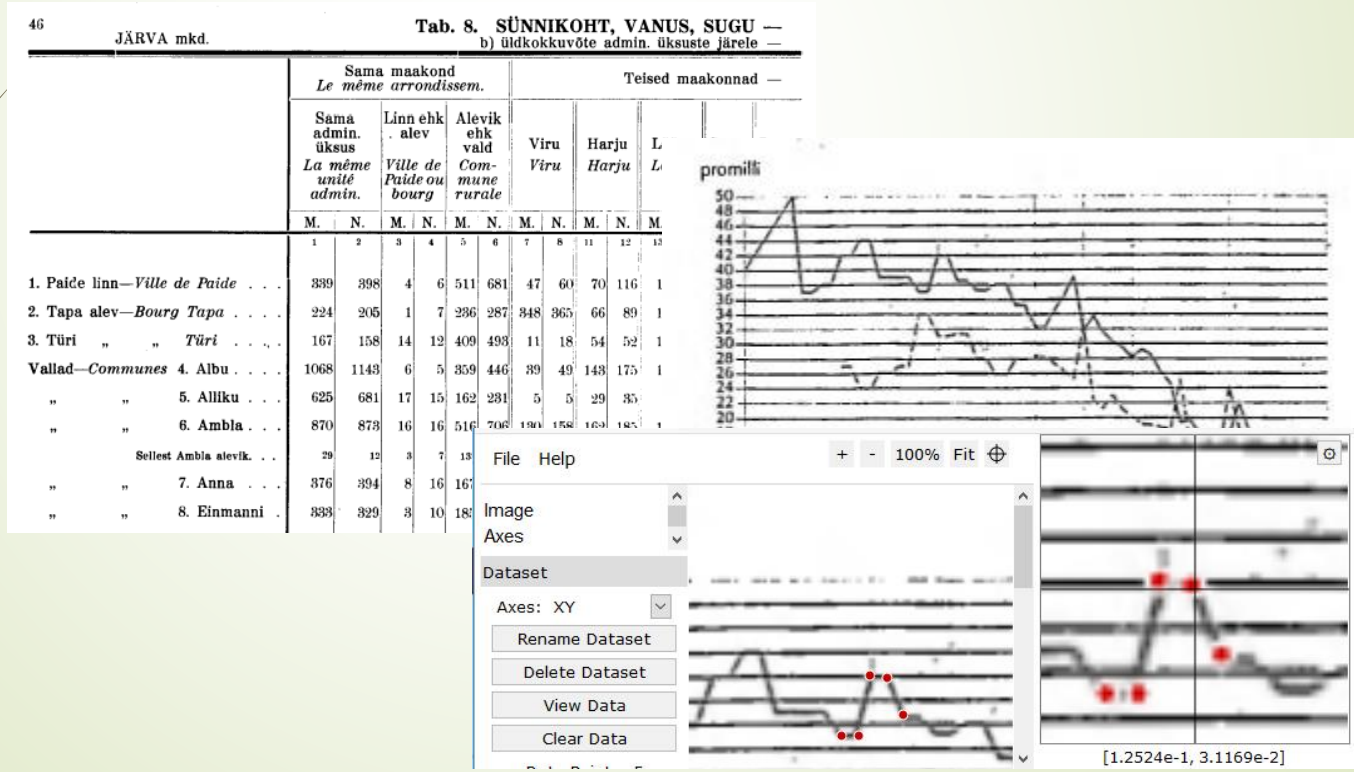
Public elementary school network in 1881

Life histories of writers



Demographic data

- Censuses and analyses

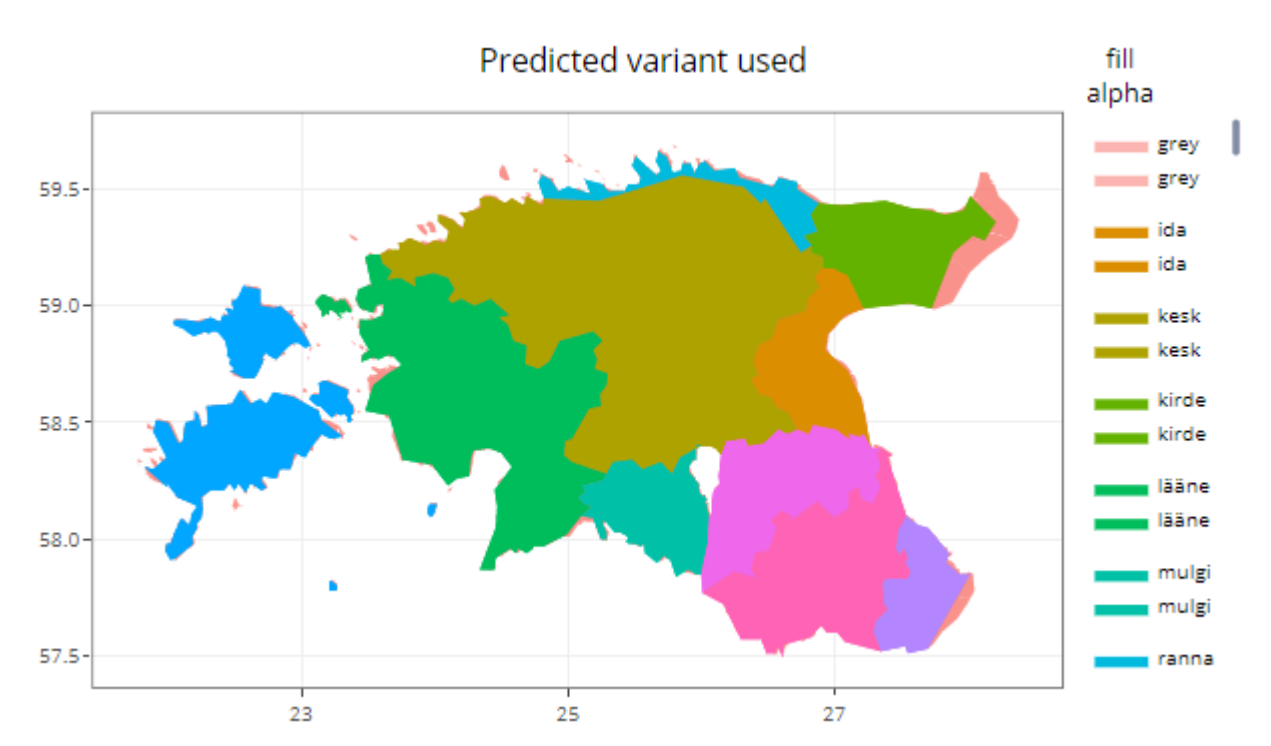


46 JÄRVA mkd.

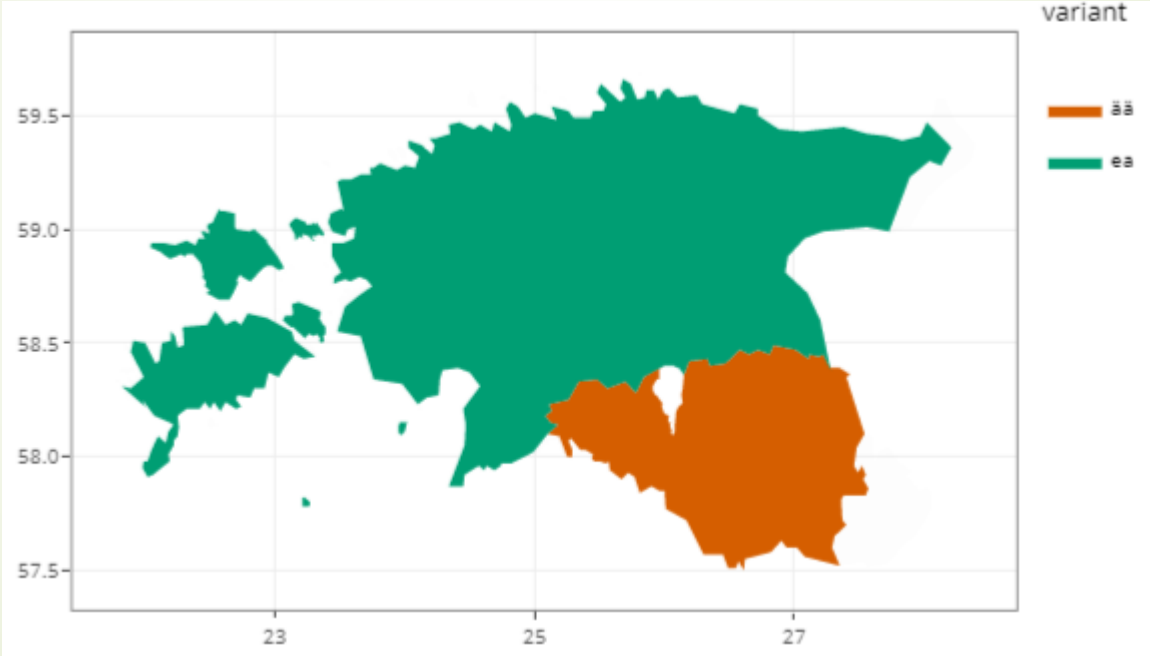
Tab. 8. SÜNNIKOHT, VANUS, SUGU — b) üldkokkuvõte admin. üksuste järele —

	Sama maakond *Le même arrondissem.*						Teised maakonnad —				
	Sama admin. üksus *La même unité admin.*		Linn ehk alev *Ville de Paide ou bourg*		Alevik ehk vald *Commune rurale*		Viru *Viru*		Harju *Harju*		L *L*
	M.	N.	M.	N.	M.	N.	M.	N.	M.	N.	M.
	1	2	3	4	5	6	7	8	11	12	13
1. Paide linn—*Ville de Paide*	339	398	4	6	511	681	47	60	70	116	1
2. Tapa alev—*Bourg Tapa*	224	205	1	7	236	287	348	365	66	89	1
3. Türi „ „ *Türi*	167	158	14	12	409	493	11	18	54	52	1
Vallad—*Communes* 4. Albu	1068	1143	6	5	359	446	39	49	143	175	1
„ „ 5. Alliku	625	681	17	15	162	231	5	5	29	35	
„ „ 6. Ambla	870	873	16	16	516	706	190	158	163	185	1
Sellest Ambla alevik.	29	12	3	7	13						
„ „ 7. Anna	376	394	8	16	167						
„ „ 8. Einmanni	333	329	3	10	183						

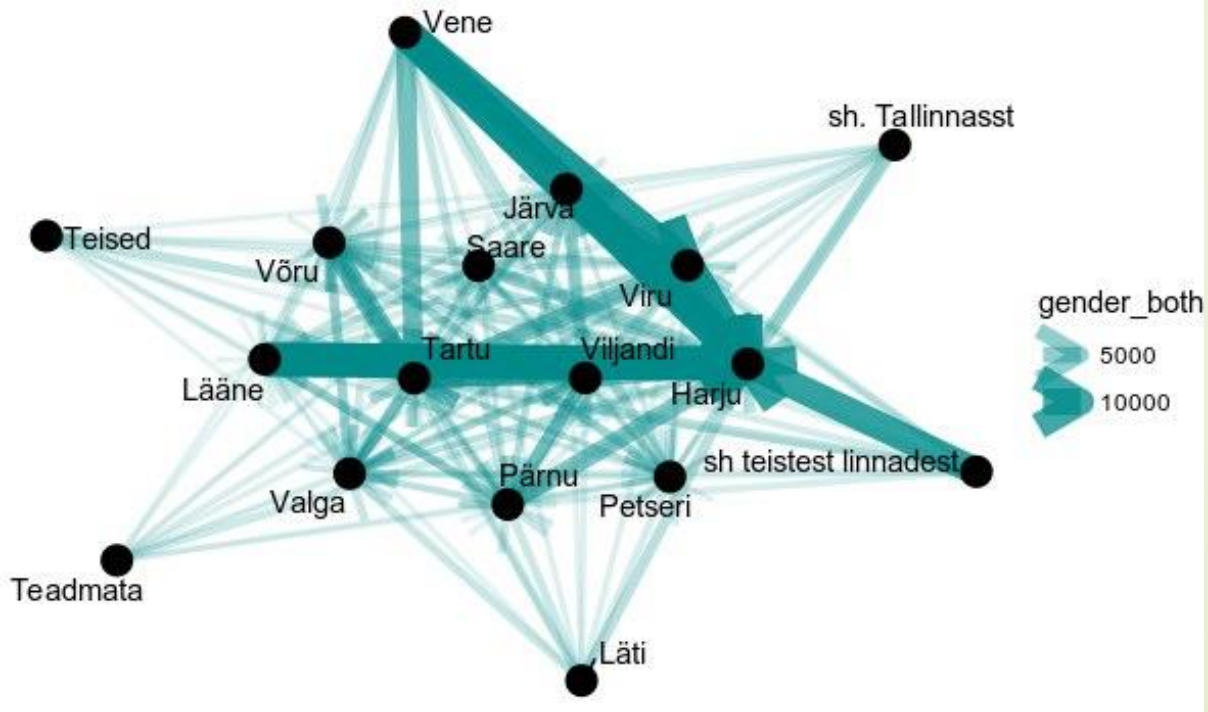
Dialect geography



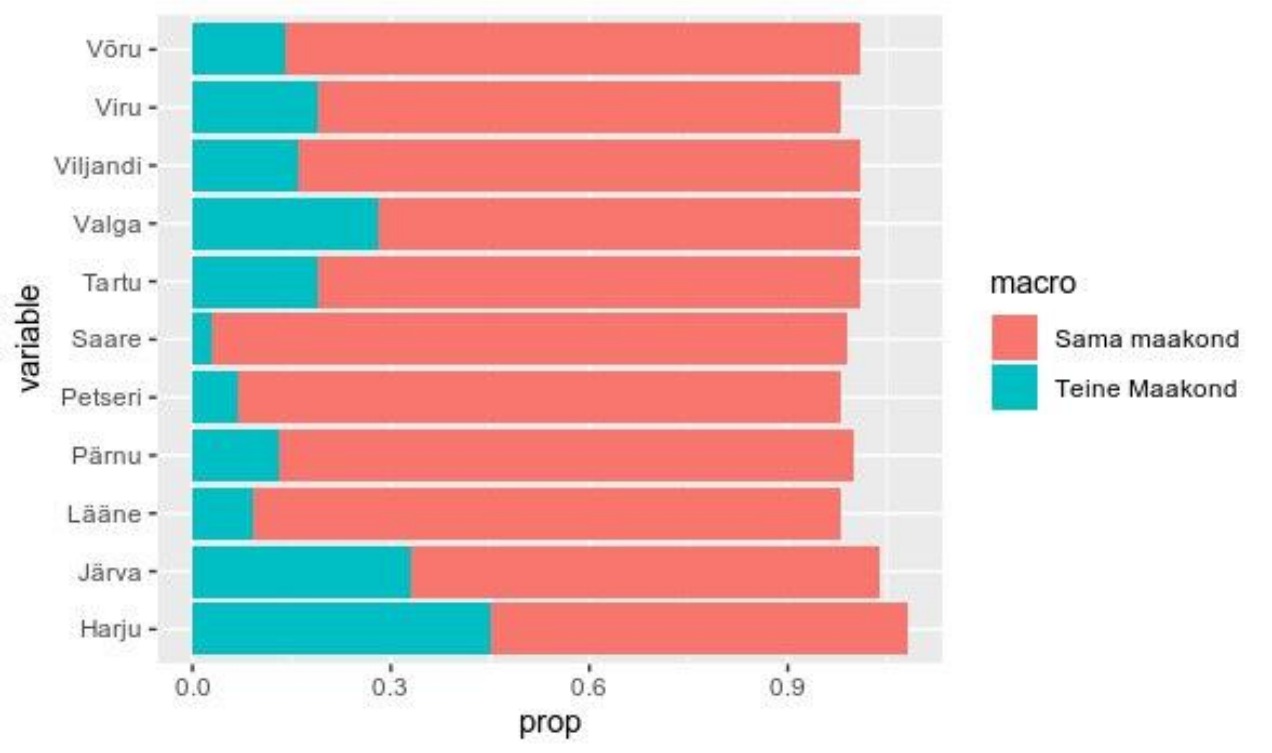
Dialect predicted values



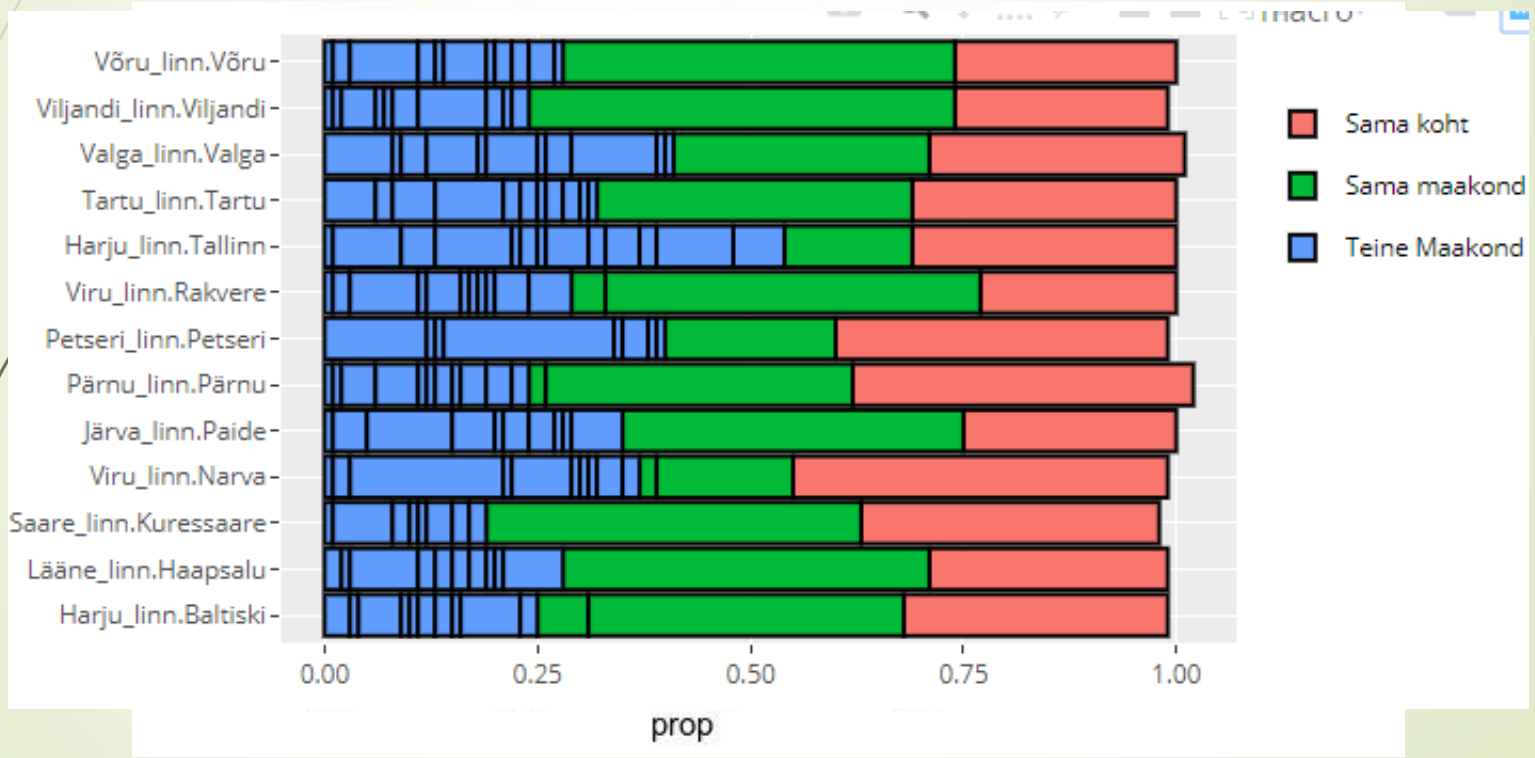
Dialect compositions in cities



Migration and dialects



Migration and dialects



Getting to research

Operationalizing the question

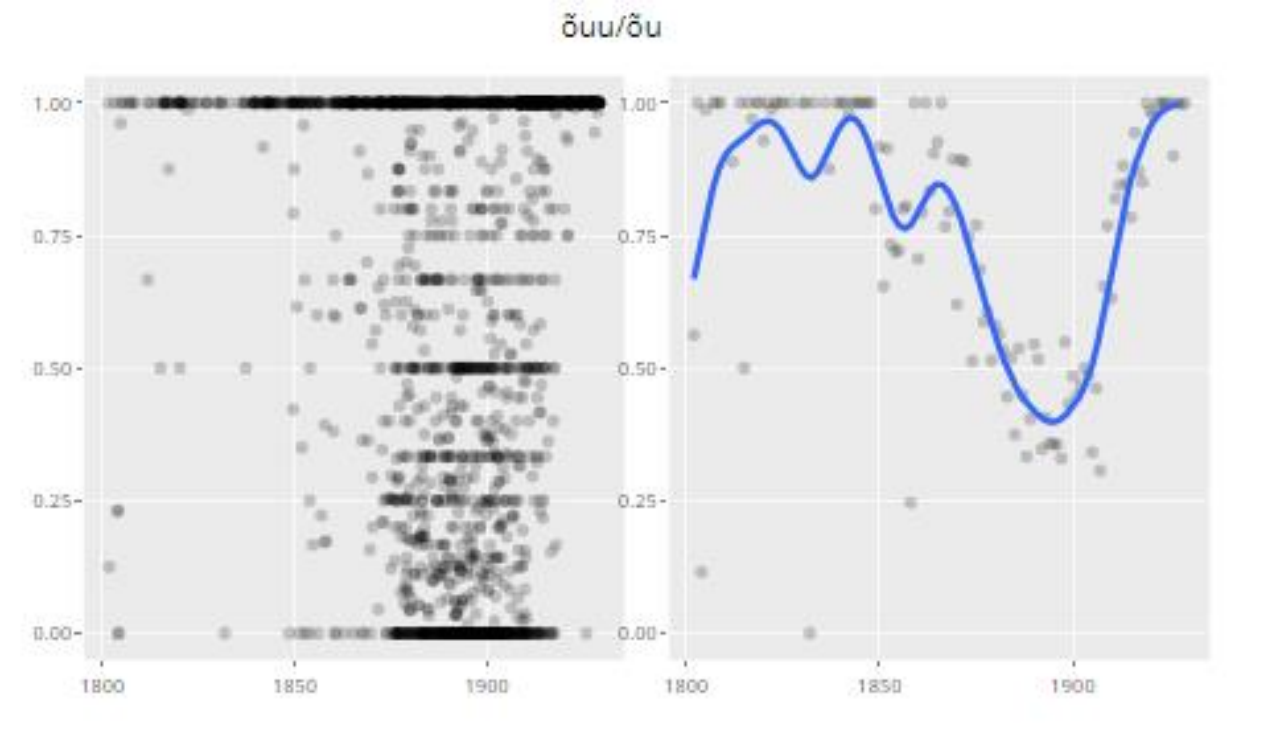
Making the study

- Specify the measured variables
- Measure in suitable texts
- Combine with metadata
- Check the mechanisms

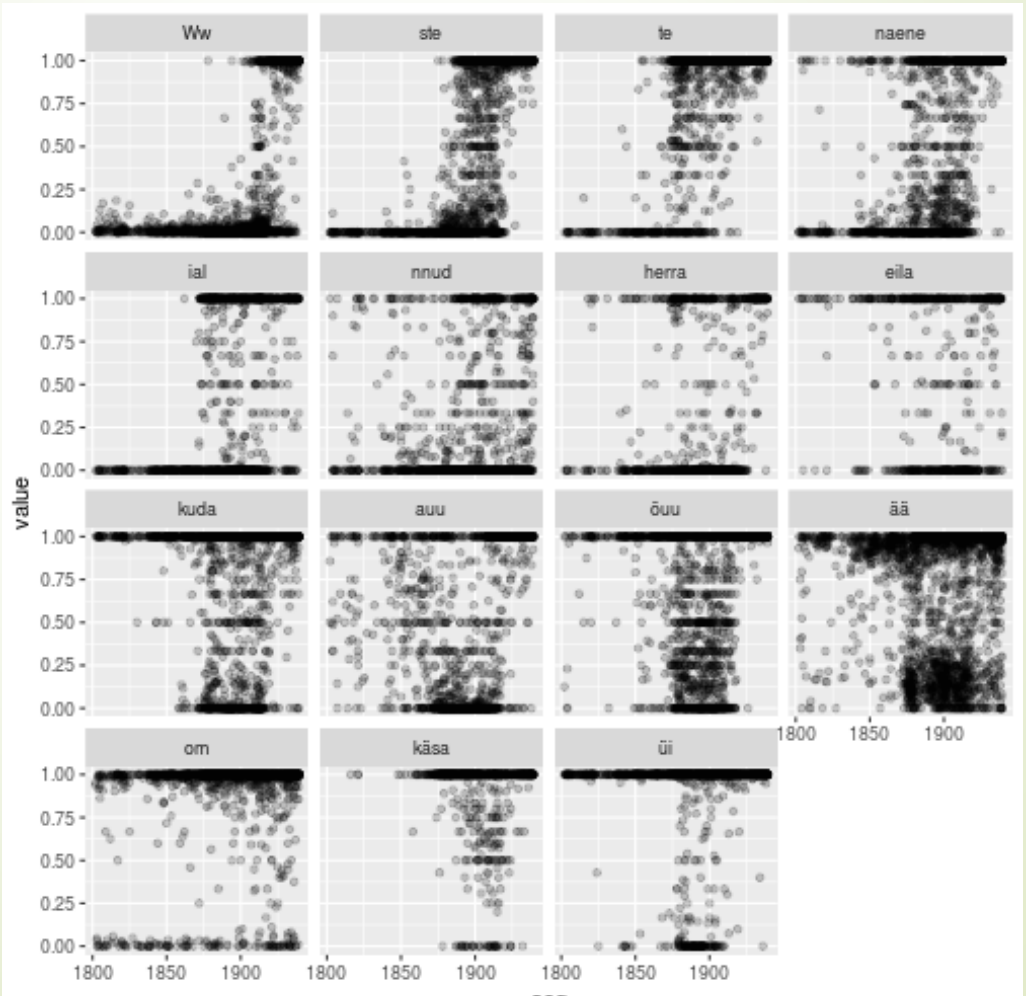
Variants and variation

- A set of variables:
  - w / v – wabariik, vabariik
  - ää / ea – hää, hea; sääl, seal
  - üi / üü – nüid, nüüd
  - herra / härra
  - naine / naene
  - om / on
- Fairly common in frequency
- Have interesting variation in 1880-1920

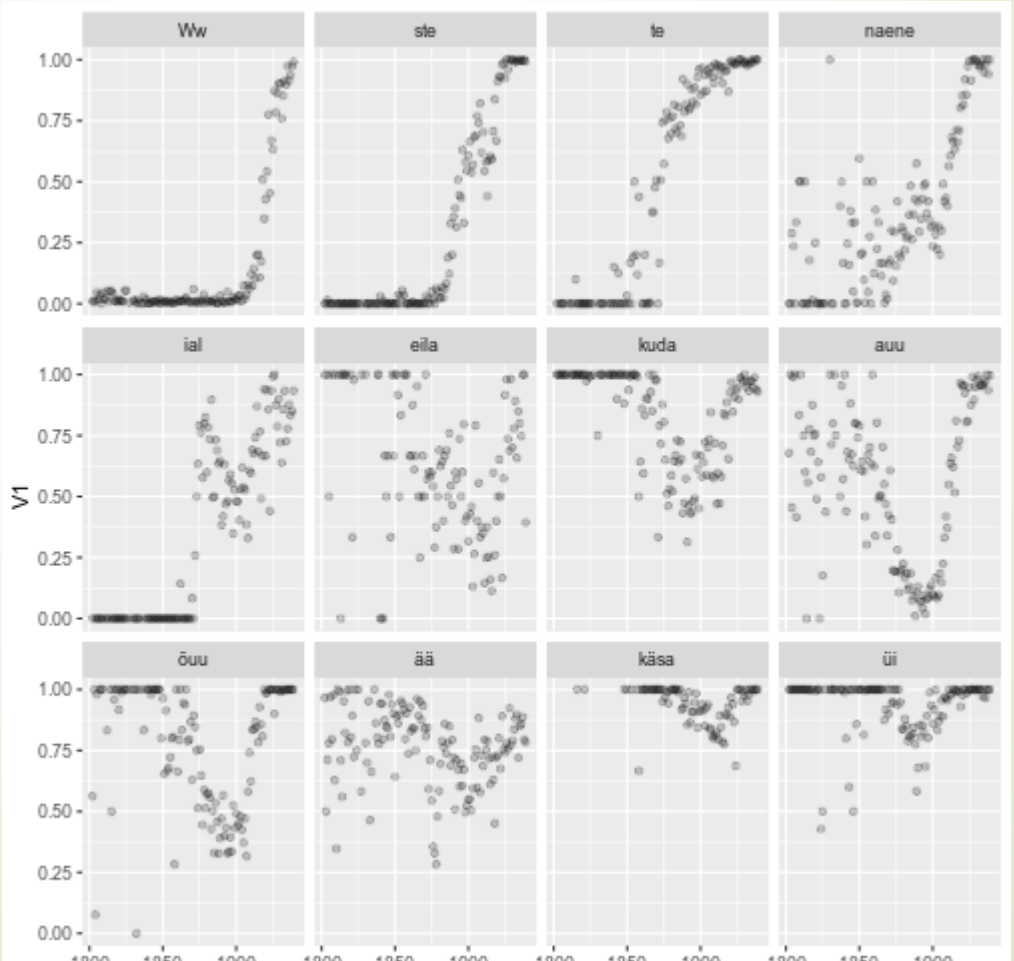
Variants and variation



Study on particular variants



Average trends



Language prescription

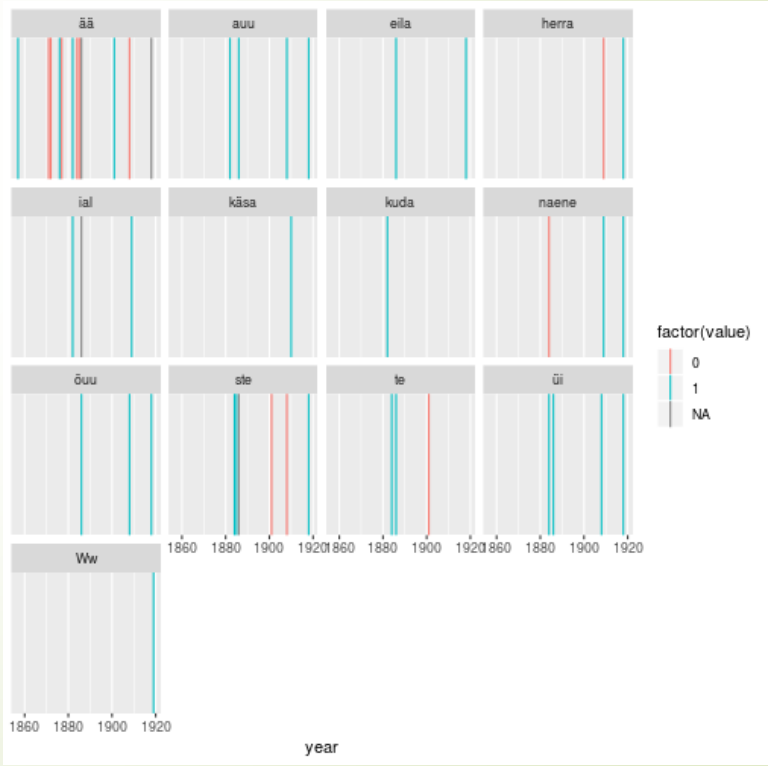
- Historian's overviews

- käänata sõna *juus* (=juuksekarv) järgmiselt: *juukse, juuksed, juuste*;
- käänata sõna *kodu* ainsuse omastavas *kodu* ja sisseütlevas *koju*;
- kirjutada *üü*, mitte *üi* sõnades nagu *nüüd, hüüdma, tüütama, süütama*;
- kirjutada *kaine, naine, laine*, mitte *kaene, naene, laene*, aga *naeris, paene* (sõnast *paas*), *vaene*; samuti *aed, aednik*, mitte *aid, aidnik*;
- kirjutada *kullassepp* (kahe *s*-iga), mitte *kullasepp*;
- kirjutada *herra, veevel*, mitte *härra, väävel*, aga *närvid*, mitte *nervid*;
- kirjutada *väike, kõik, lõikama*, mitte *veike, keik, leikama*;
- kirjutada *kähar*, mitte *kähär*;

Overview of proposed norms in 1909 (Raag 2008)

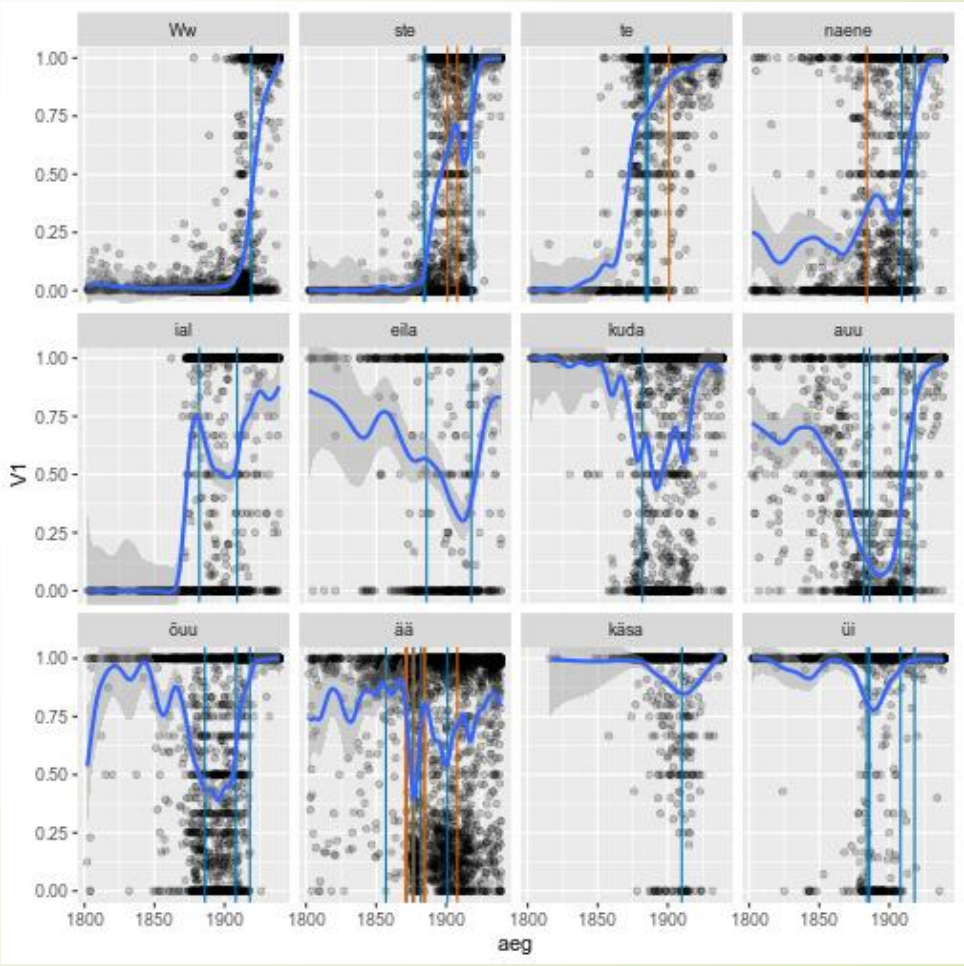
Language prescription

- Prescriptive events on the timeline
- Blue (suggested norm towards 1)
- Red (suggested norm towards 1)

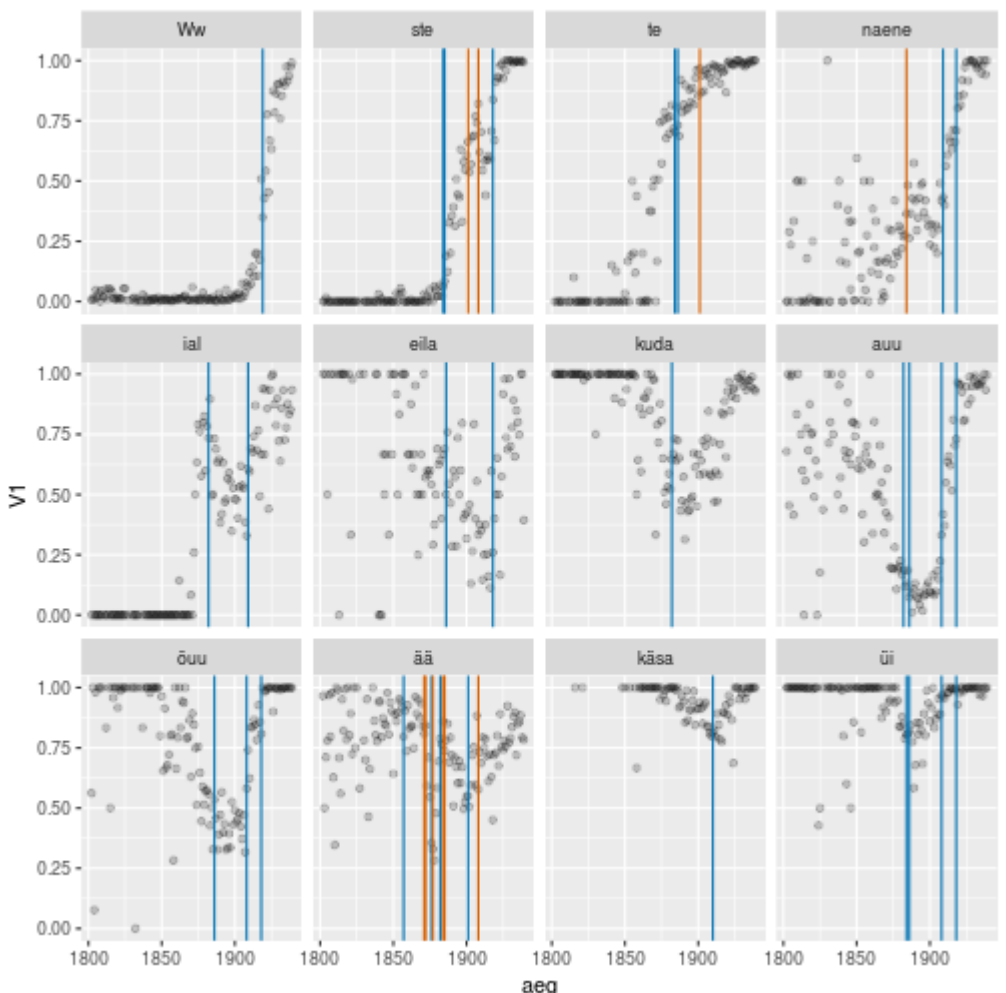


Language prescription

- Prescriptive events align very well with trends in usage.
- => at least some degree of prescriptive influence.



Language prescription

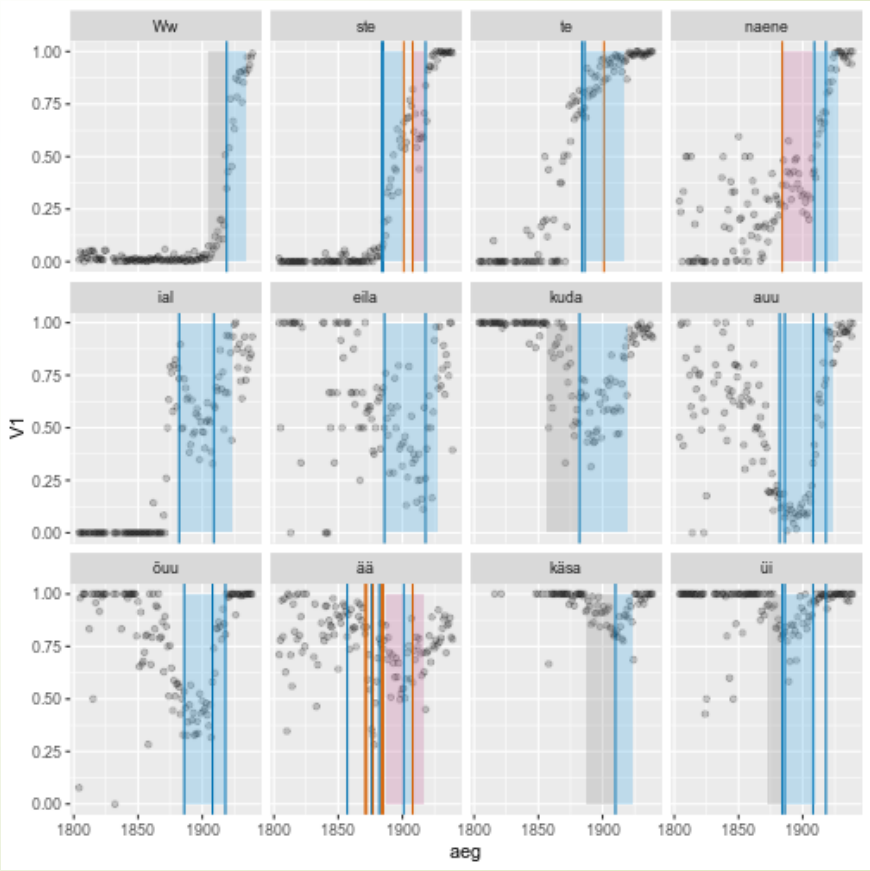


Mechanisms of standardization

- Influences on language usage
  - Home dialect
  - Everyday language use
  - Education
  - Prescription
  - Etc
- What made a suggestion successful?
- Generational change vs individual change?

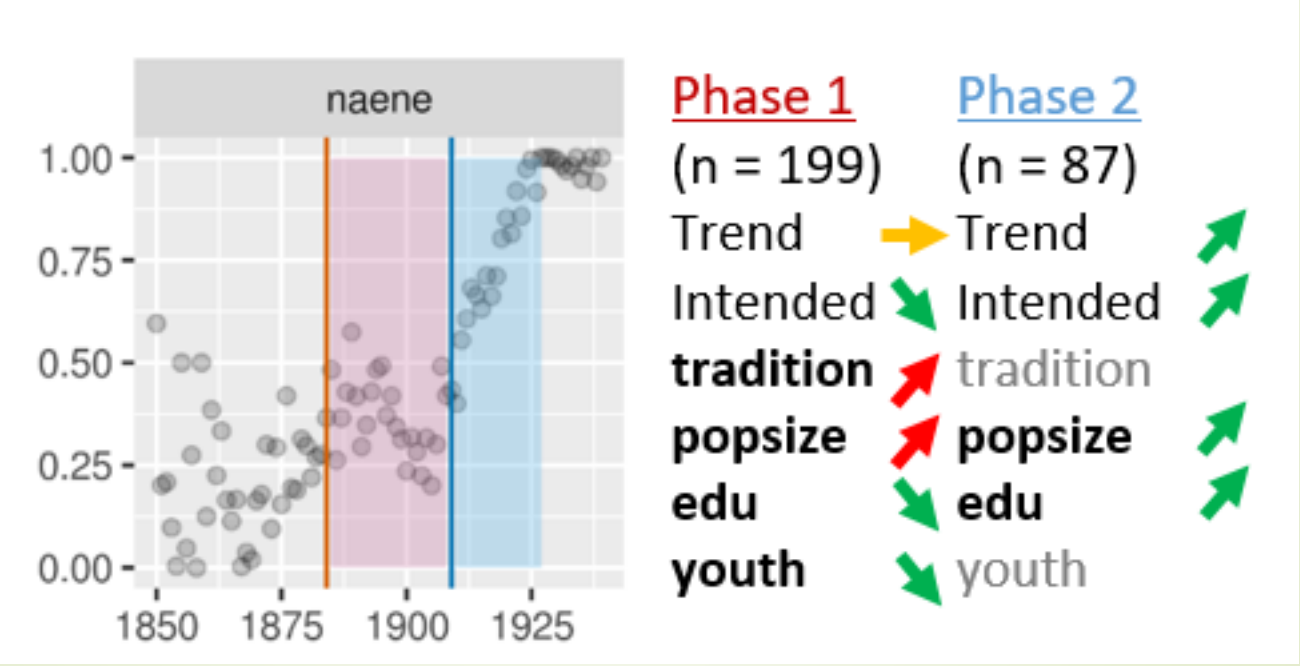
Phases of change

- Looking at the aftermath of prescriptive events
- Blue (suggested norm towards 1)
- Red (suggested norm towards 1)



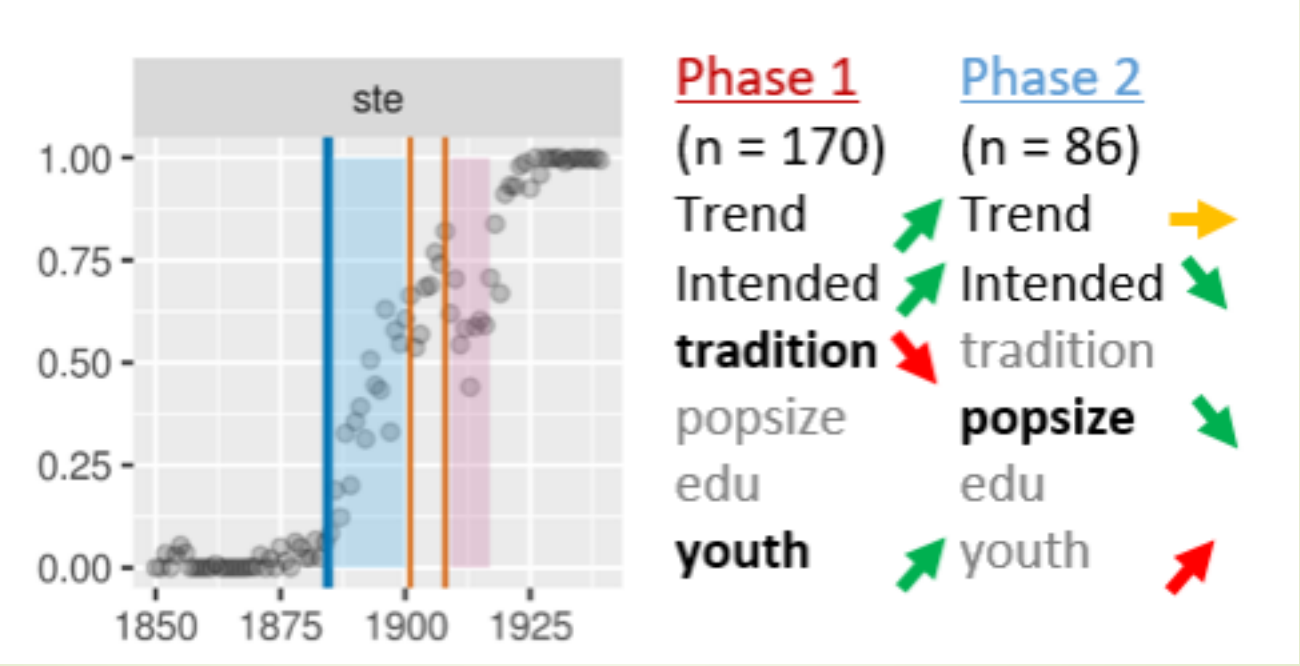
Simple tendencies in data

- Significant predictors (logistic regression)
  - Model does not include dialect info



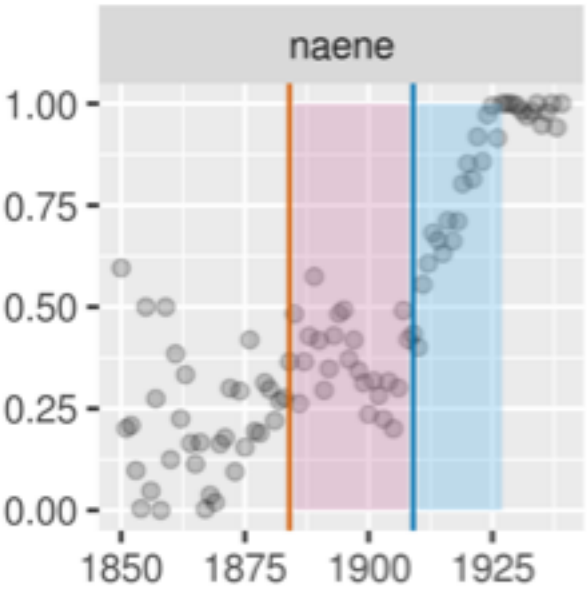
Simple tendencies in data

- Significant predictors (logistic regression)
  - Model does not include dialect info



Simple tendencies in data

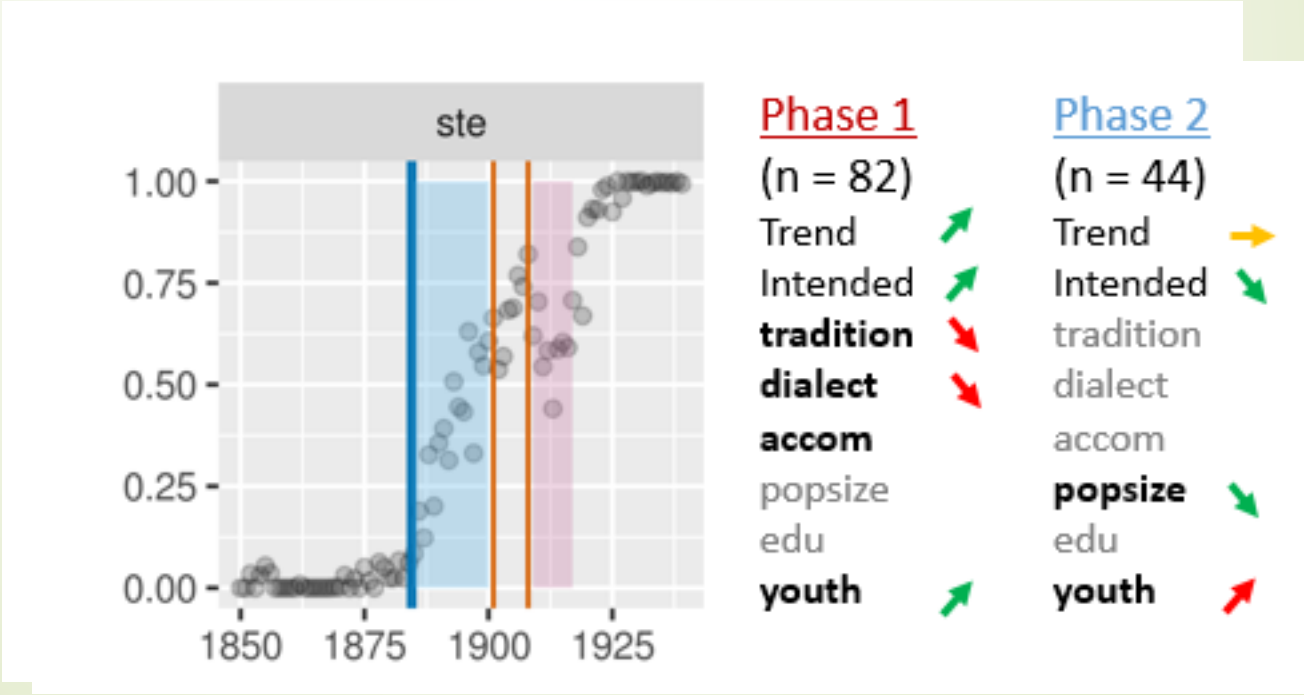
- Significant predictors (logistic regression)
  - Model includes dialect info

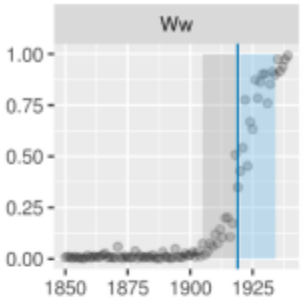


Phase 1 (n = 107)		Phase 2 (n = 35)	
Trend	→	Trend	↗
Intended	↘	Intended	↗
tradition	↗	tradition	
dialect	↗	dialect	
accom	↘	accom	
popsize		**popsize**	↗
edu		**edu**	↗
youth	↘	youth	

Simple tendencies in data

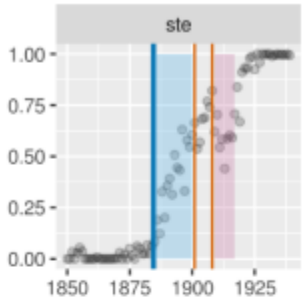
- Significant predictors (logistic regression)
  - Model includes dialect info





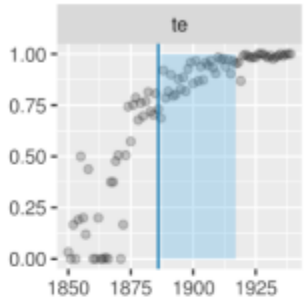
Phase 1 (n = 91)
Trend
Intended
tradition
popsize
edu
youth

Phase 2 (n = 56)
Trend
Intended
tradition
popsize
edu
youth

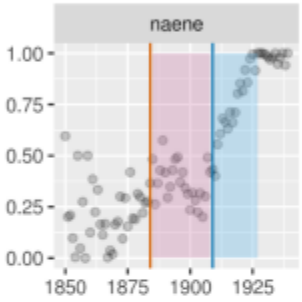


Phase 1 (n = 170)
Trend ↗
Intended ↗
tradition ↘
popsize
edu
youth ↗

Phase 2 (n = 86)
Trend →
Intended ↘
tradition
popsize ↘
edu
youth ↗

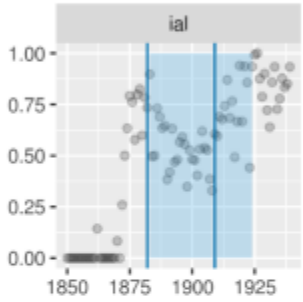


Phase 1 (n = 265)
Trend ↗
Intended ↗
tradition ↘
popsize
edu
youth ↗



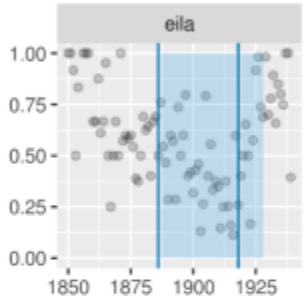
Phase 1 (n = 199)
Trend →
Intended ↘
tradition ↗
popsize ↗
edu ↘
youth ↘

Phase 2 (n = 87)
Trend ↗
Intended ↗
tradition
popsize ↗
edu ↗
youth

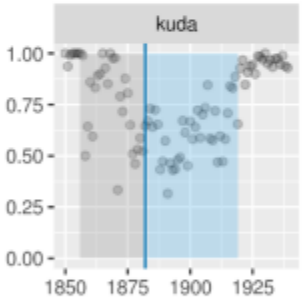


Phase 1 (n = 210)
Trend →
Intended ↗
tradition
popsize ↘
edu
youth

Phase 2 (n = 71)
Trend ↗
Intended ↗
tradition
popsize
edu
youth

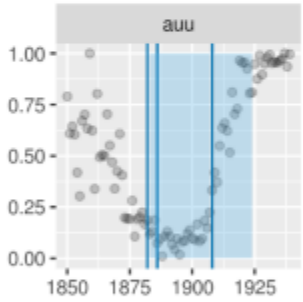


Phase 1 (n = 101)
Trend
Intended
tradition
popsize
edu
youth



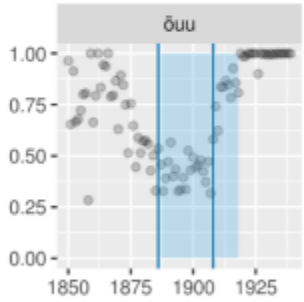
Phase 1 (n = 117)
Trend
Intended
tradition
popsize
edu
youth

Phase 2 (n = 323)
Trend ↗
Intended ↗
tradition
popsize ↘
edu
youth



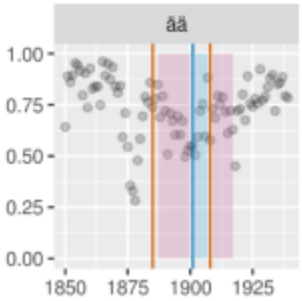
Phase 1 (n = 206)
Trend →
Intended ↗
tradition
popsize
edu
youth ↘

Phase 2 (n = 71)
Trend ↗
Intended ↗
tradition ↘
popsize ↗
edu ↗
youth



Phase 1 (n = 200)
Trend →
Intended ↗
tradition
popsize ↗
edu
youth ↘

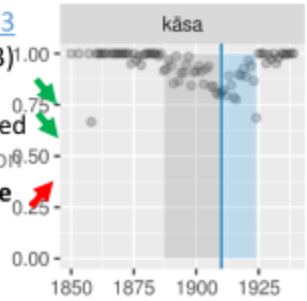
Phase 2 (n = 77)
Trend ↗
Intended ↗
tradition
popsize ↗
edu
youth



Phase 1 (n = 158)
Trend ↘
Intended ↘
tradition ↗
popsize
edu
youth ↘

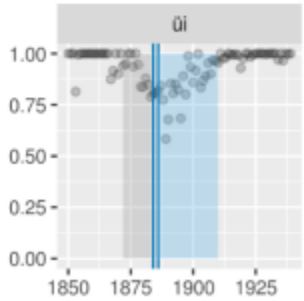
Phase 2 (n = 57)
Trend ↗
Intended ↘
tradition ↘
popsize ↗
edu
youth ↘

Phase 3 (n = 93)
Trend ↗
Intended ↗
tradition
popsize ↘
edu
youth



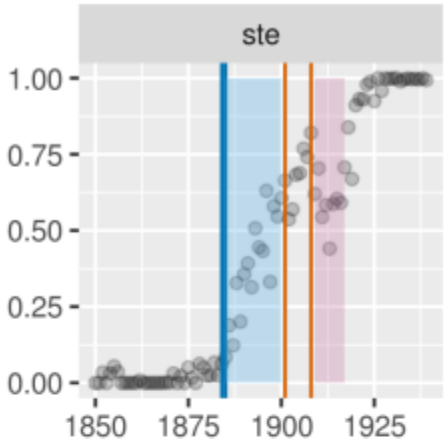
Phase 1 (n = 166)
Trend
Intended
tradition
popsize
edu
youth

Phase 2 (n = 59)
Trend
Intended
tradition
popsize
edu
youth

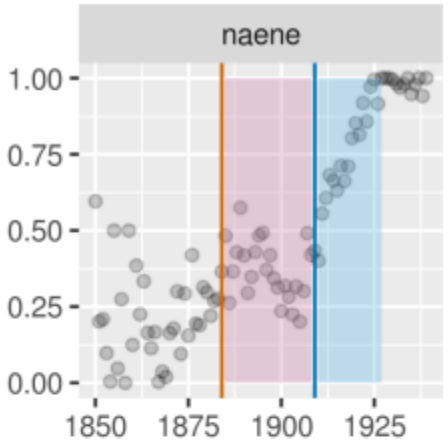


Phase 1 (n = 110)
Trend ↘
Intended
tradition
popsize
edu
youth

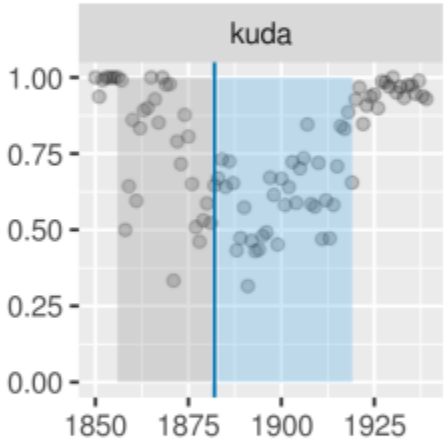
Phase 2 (n = 261)
Trend ↗
Intended ↗
tradition
popsize ↘
edu
youth



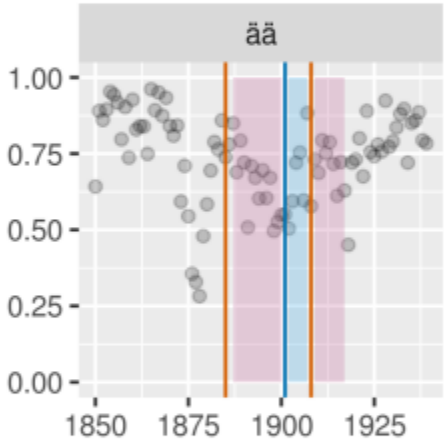
Phase 1 (n = 82)	Phase 2 (n = 44)
Trend ↗	Trend →
Intended ↗	Intended ↘
tradition ↘	tradition
dialect ↘	dialect
accom	accom
popsize	**popsize ↘**
edu	edu
youth ↗	**youth ↗**



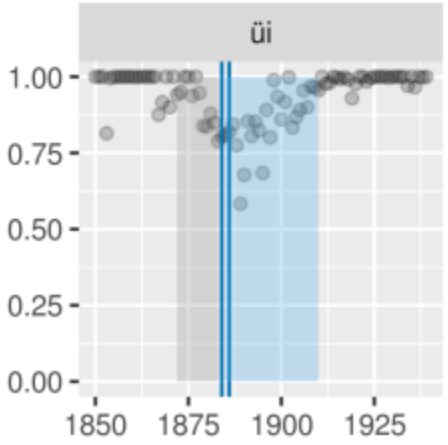
Phase 1 (n = 107)	Phase 2 (n = 35)
Trend →	Trend ↗
Intended ↘	Intended ↗
tradition ↗	tradition
dialect ↗	dialect
accom ↘	accom
popsize	**popsize ↗**
edu	**edu ↗**
youth ↘	youth



Phase 1 (n = 59)	Phase 2 (n = 161)
Trend ↘	Trend ↗
Intended	Intended ↗
tradition	tradition
dialect	dialect
accom	accom
popsize	**popsize ↘**
edu	edu
youth	youth



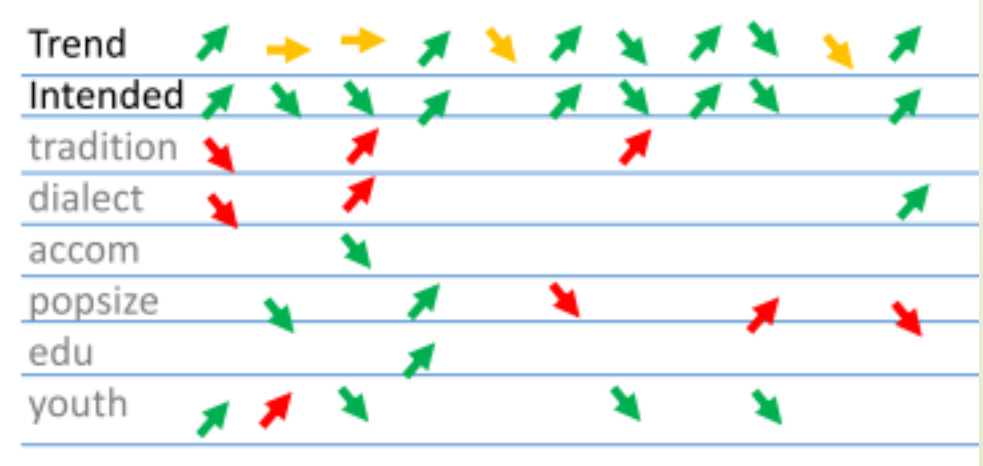
Phase 1 (n = 77)	Phase 2 (n = 28)	Phase 3 (n = 45)
Trend ↘	Trend ↗	Trend ↘
Intended ↘	Intended ↗	Intended ↘
tradition ↗	tradition	tradition
dialect	dialect	dialect
accom	accom	accom
popsize	popsize	**popsize ↗**
edu	edu	edu
youth ↘	youth	**youth ↘**



Phase 1 (n = 49)	Phase 2 (n = 128)
Trend ↘	Trend ↗
Intended	Intended ↗
tradition	tradition
dialect	**dialect ↗**
accom	accom
popsize	**popsize ↘**
edu	edu
youth	youth

A basic overview

- Significant factors across all phases
  - Top (no dialect info), bottom (with dialect info)



Study conclusions

- Main hypothesis (late 19c Estonian):
  - Spontaneous changes align with an influx of new authors
  - Prescriptive events line up clearly with trends
  - Some generational influence apparent: successful changes are led by young and large cities.
- Study generally:
  - Data is everywhere and has not been collected
  - Already „first look“ at it can give insights
  - Thinking in terms of data also adds to theory

General discussion

- 80% of DH research is finding the data and getting it ready
  - Manual + technological work
  - Good-enough solutions for current tasks
- Open collections can have a variety of uses
  - Difficult to predict
- Also by-products of research can be useful
  - Databases & processing
  - Visualizations & understanding

Some advice/thoughts

- Start & stay open, document everything!
- Think how your current work fits to the whole project
  - Work with manageable chunks and fit them in right away
- Don't worry about lack of skills beforehand, but consult specialists!
  - Usually your problems are not new and have solutions.
  - However working with „data" is new to humanities so you may need to step outside your institute/department.

Interactive plot links

- Birthplaces of authors, publishers etc 1800-1940 by decade
- Geolocated life histories from biographical data over time
- Sources of immigrants by county in 1922
- One linguistic variable measured within books

Thank you!

peeter.tinits@gmail.com

@yrgsupp

TALLINN UNIVERSITY

UNIVERSITY OF TARTU
1632

EESTI RAHVUS-RAAMATUKOGU