A DigiHum approach to HistSocLing of Estonian

Peeter Tinits
25.07.2019
National Library of Latvia
BSSD2019

A case study in DH

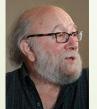
- Project:
 - What mechanisms are responsible for spelling standardization at the end of 19th century Estonian?
- Data sources:
 - Anything relevant
- Timeline:
 - PhD dissertation

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



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

Eesti Kirjakeele Korpus: 1890ndad

Statistika ja bibliograafia

- Korpuse koostisosad valdkondade kaupa
- Algallikad paberkandjatel failide kaupa.

Tekstid

Märgendamata tekst, iga lause eraldi real

- Ajakirjandustekstid (zip-fail 586 Kb)
- Ilukirjandustekstid (zip-fail 438 Kb)

Vana kirjakeele korpus

tekstidest

Avaleht Otsing Juhend Tekstid Väljaanded

Canb'ent Paulitudes

Sair De Bistrecter und Santigsaire Cellusier Nichtandung
saire Cellusier Nichtandung
Wieber E Jed/
Demonster

Commander

Schopman Collins on De testienen.

De Earlier von Alleidere Angeleinen.

M. HENRIC O. Cellus Vieller von Seiner

Schopman Schopman Seiner

Corda Sphin Seiner Jeden Vielleren

Schopman Seiner Jeden Vielleren

Corda Sphin Seiner Jeden Vielleren

Co

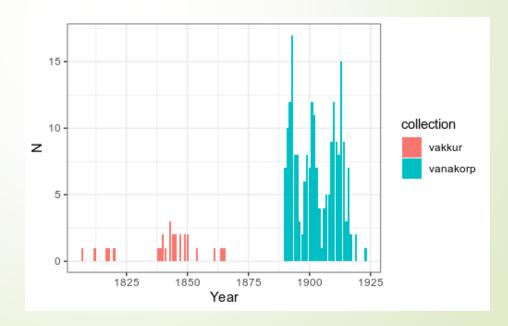
Lessi vana kirjakeele korpus sisalada 15. kuni 19. Sajandi tekset Vanemad tekstid on morfoloogiliselt märgendatud, st neist tekstides saab infot otsida sõnade tänapäevases kirjaviisis algvormide ning vormiinfo järgi.

• 15. ja 16. sajandist on korpusesse lisatud kõik teadaolevad ja

- 15. ja 16. sajandist on korpusesse lisatud koik teadaolevad ja säilinud eestikeelsed tekstid (v.a nimeloendid), nii käsikirjad kui ka trükised.
- 17. sajandist on korpusesse lisatud enamik säilinud trükitekste.
 18. ja 19. sajandist on lisatud valik trükitekste. Märgendatud on os
- 19. sajandi II poolest on lisatud vallakohtute protokollid, mis on automaatsett märgendatud; vt täpsemalt M.-L. Pilvik, K. Muischnek, G. Jaanimäe, L. Lindström, K. Lust, S. Orasmaa, T. Türna "Moistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine"

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







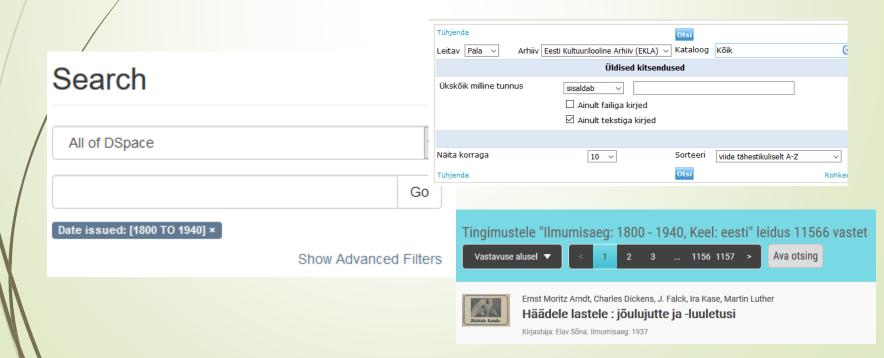






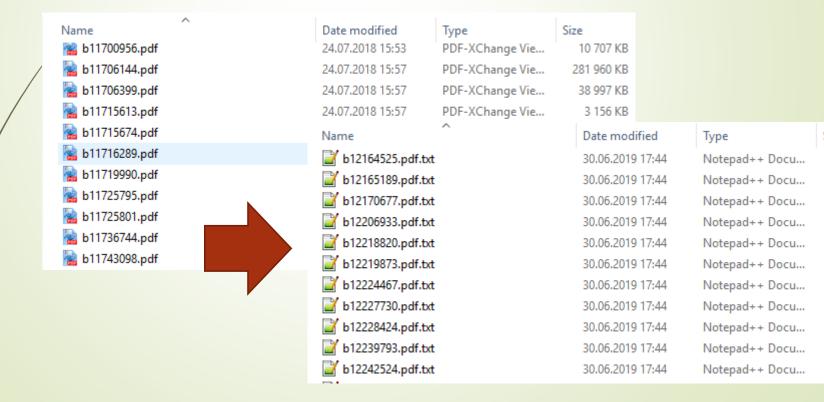
Examples of text collections

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



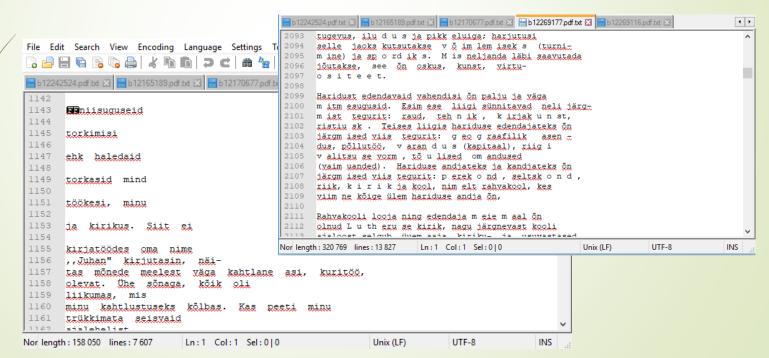
Getting the text

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

```
Bornhöhe, Eduard "Tallinna narrid ja narrikesed" 1892 lk 3-12
                         "Willu wõitlused" 1890 lk 5-43
pro0005
         Bornhöhe, Eduard
                           "Würst Gabriel" 1893 1k 229-240
pro0006
         Bornhöhe, Eduard
         Eisen, M. J. "Hiiu köster ja Saare kirikhärra" 1893 lk 43-52
pro0007
pro0008
         Eisen, M. J. "Tartu saladused" 1891 lk 7-18
pro0009
        [???] "Hella" 1890 lk 3-12
         Hermann, K. A. "Lapse mälestus" 1896 lk 1-8
pro0010
         Hermann, K. A. Rikka ja waese pulmad. Külajutukene. 1899 Lk. 29-37
         Hermann, K. A. Uudisjutud Eesti rahwa elust. 1895 Lk. 37-45
pro0012
         [???] "Hugo ja Tekla". Jutustus Toolse (ranna) lossist ja ritterist, mis seal 15. aast
pro0013
pro0014
         Tüll, He
                                         ema imelikud juhtumised wõera mere-saarte peal. 1891
                 Ein einfeltige weise zu
pro0015
         Jaanus.
                                         olm juttu noorerahwa elust." 1893 lk 3-11
         Jakobson Beten, fur einen guten
pro0016
                                         399 lk 34-39
pro0017 Järw, J. freund Mart. Luther.
                                         tustus Eesti minewikust." 1892 Lk. 127-136
```



View/Open

↑ r_iii_i_252i_10.pdf
(17.68Mb)

↑ Tekstituvastusega

Date 1535

Author Luther, Martin

Metadata Show full item record

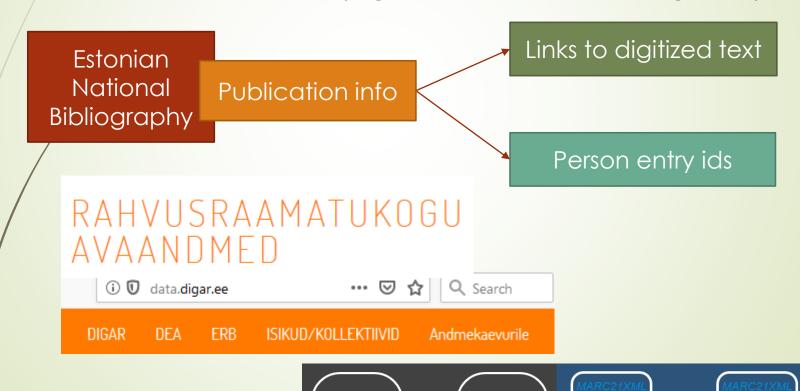
sordiaretaja

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

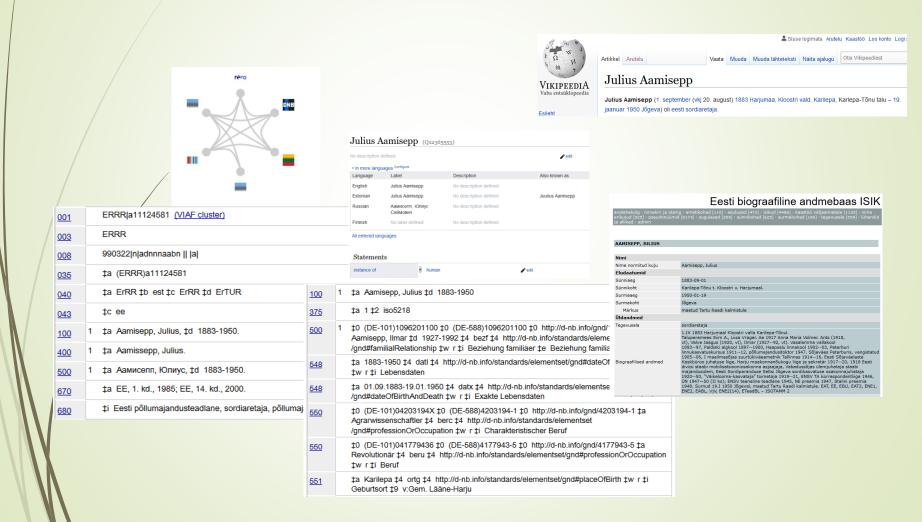
Metadata

General indexes (e.g. Estonian National Bibliography)

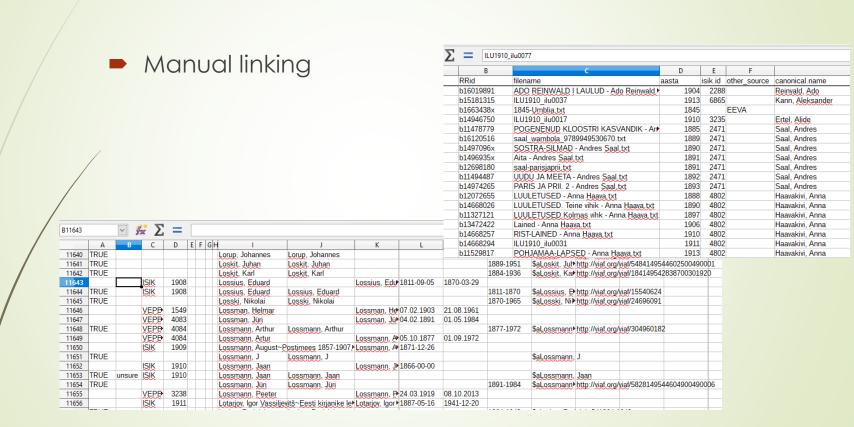


Collecting them all, linking them Wikipedia Person entry to linked data Wikidata Other VIAF Person entry ids libraries Aamisepp, Julius, 1883-1950. = Revo Аамисепп, Юлиус, 1883-1950. Etc Julius Aamisepp IIII VIAF ID: 6338149544608900490008 (Personal) Permalink: http://viaf.org/viaf/6338149544608900490008 Preferred Forms 200 _ 1 <u>†a Aamisepp †b Julius †f 1883-1950</u> NB 100 1 _ <u>‡a Aamisepp, Julius ‡d 1883-1950</u> réro 100 1 _ ta Aamisepp, Julius, td 1883-1950 100 1 _ <u>†a Aamisepp, Julius, †d 1883-1950.</u> III 100 0 _ <u>†a Julius Aamisepp</u> 100 1 _ <u>‡а Аамисепп, Юлиус, ‡d 1883-1950.</u>

Aggregating the data



Manual linking



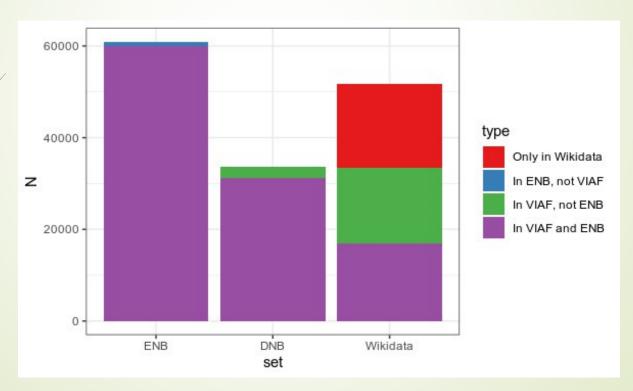
Harmonization

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

```
#works <- works#[!is.na(koht),.N,by=koht]
                                             works <- works[,koht:=str replace all(koht, "s$", "")]
                                             works <- works[,koht:=str_replace_all(koht,"l$","")]</pre>
works[kirjastus=="w.bormi par.",kir; 34
                                             works <- works[koht!=""&koht!="S.l."&koht!="S.l.,"&koht!="S. l."&koht!="S. l.,"]
works[kirjastus=="f.w.borm",kirjastu35
                                             works[str detect(koht,"Paide"),koht:="Paide"]
works[str_detect(kirjastus, "noor ee! 36
                                             works[str detect(koht,"Weissenstein"),koht:="Paide"]
works[str_detect(kirjastus,"mutsu") 37
                                             works[str_detect(koht,"Вейсенштейн"),koht:="Paide"]
works[str detect(kirjastus,"ploomput 38
                                             works[str_detect(koht, "Вейссенштейн"), koht:="Paide"]
                                             works[str_detect(koht, "Haapsalu"),koht:="Haapsalu"]
works[str_detect(kirjastus, "postimel 39
                                             works[str_detect(koht, "Hapsa"),koht:="Haapsalu"]
works[str_detect(kirjastus,"postimee 40
                                             works[str detect(koht,"Гапсаль"),koht:="Haapsalu"]
works[str_detect(kirjastus,"h\\.laal 42
works[str_detect(kirjastus, "eesti ki 43
                                             works[str_detect(koht, "Keila"), koht:="Keila"]
works[kirjastus=="eesti kirjanduses@44"
                                             works[str_detect(koht, "Rakvere"),koht:="Rakvere"]
works[kirjastus=="hermann",kirjastus 45
                                             works[str_detect(koht,"Wesenberg"),koht:="Rakvere"]
works[kirjastus=="hermann'i raamatul 46"
                                             works[str detect(koht,"Везенберг"),koht:="Rakvere"]
works[kirjastus=="pealadu hermanni rhtkpi. ,kirjastus== k.a.nermann j
works[kirjastus=="hermanni raamatukauplus",kirjastus:="k.a.hermann"]
works[kirjastus=="hermann'i rmtkpl.",kirjastus:="k.a.hermann"]
works[kirjastus=="pealadu hermanni kaupluses",kirjastus:="k.a.hermann"]
works[kirjastus=="leoke'se kirjastus",kirjastus:="h.leoke"]
works[kirjastus=="leoke",kirjastus:="h.leoke"]
works[kirjastus=="leoke'se raamatuäri antikvariaat",kirjastus:="h.leoke"]
```

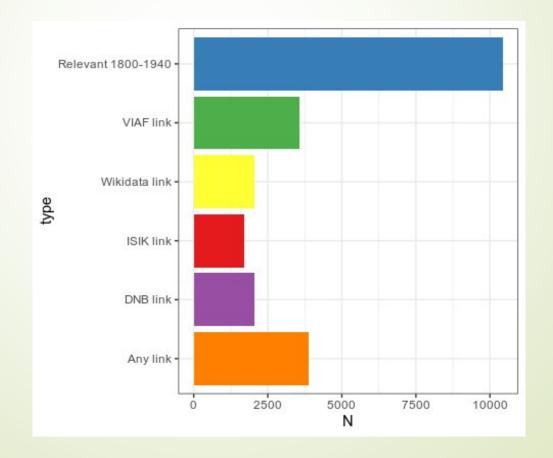
Results of linking

Linkability of archives



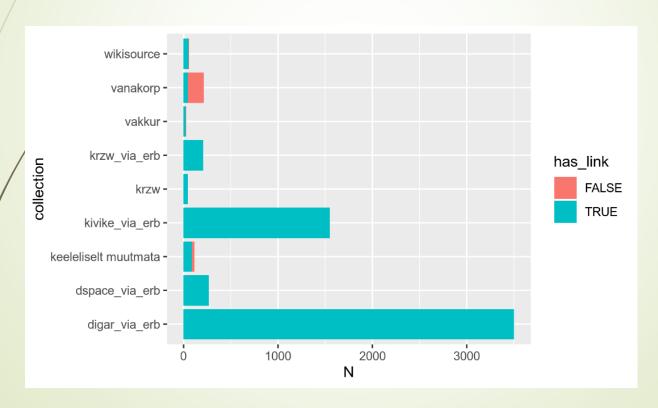
Results of linking

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)

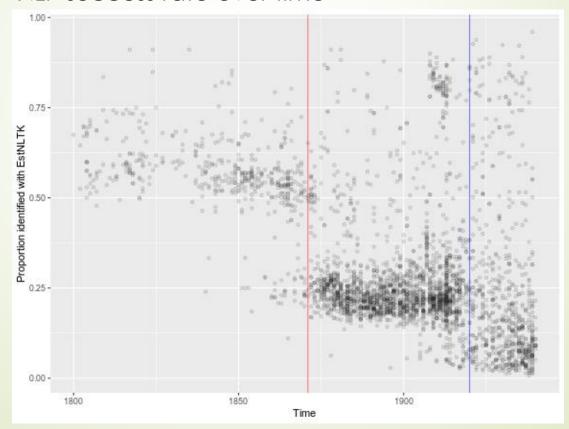
```
from estnltk import Text
text = Text('moeldud')
text.tag_analysis()
```

```
[('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')]
```

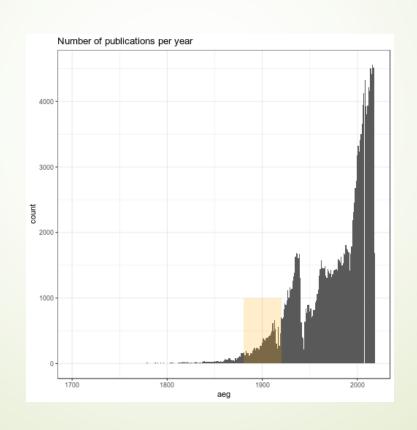
```
{'paragraphs': [{'end': 7, 'start': 0}],
 'sentences': [{'end': 7, 'start': 0}],
 'text': 'mõeldud',
 'words': [{'analysis': [{'clitic': '',
     'ending': '0',
     'form': '',
     'lemma': 'mõeldud',
     'partofspeech': 'A',
     'root': 'mõel=dud',
     'root_tokens': ['moeldud']},
    {'clitic':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

Results of the corpus

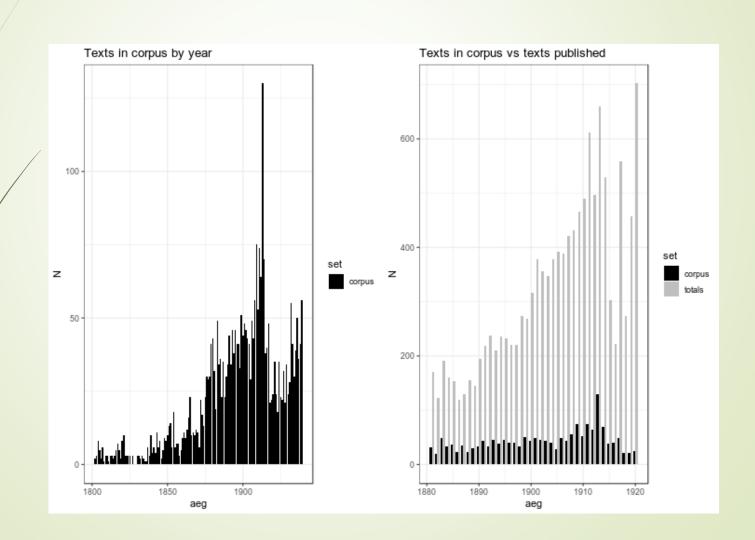
NLP success rate over time



The published texts

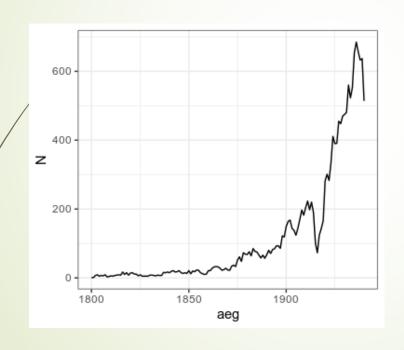


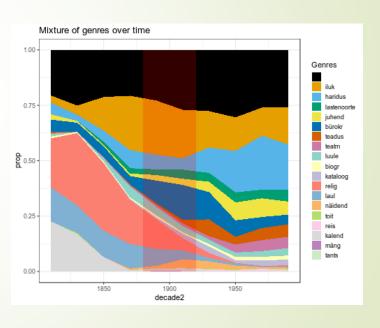
Representativeness

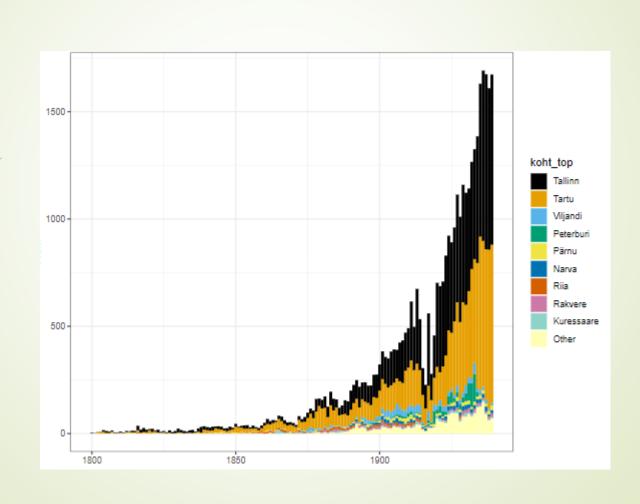


Publishing industry

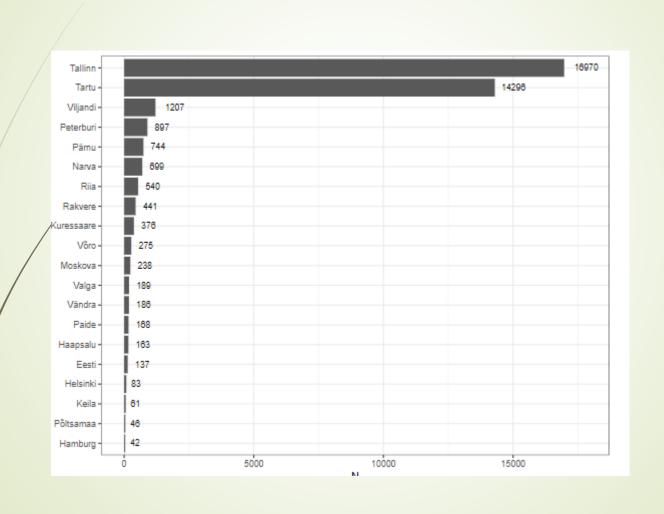
More authors and open genres



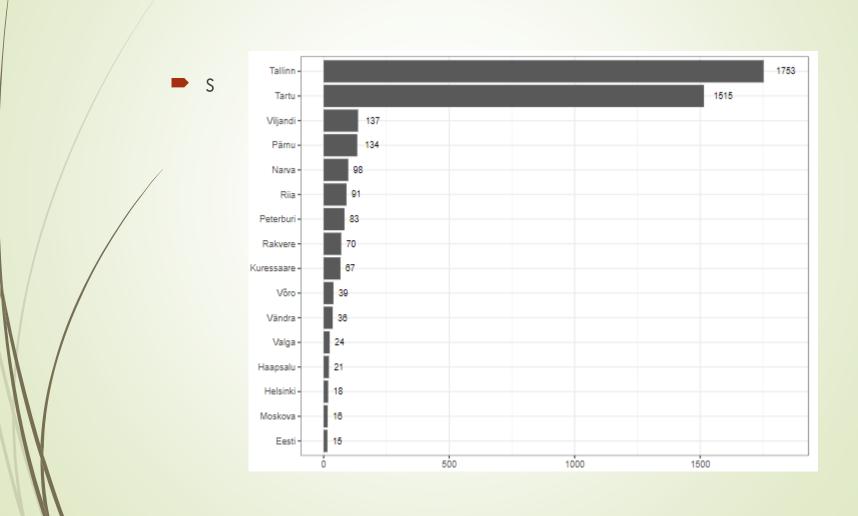




A few cities dominate

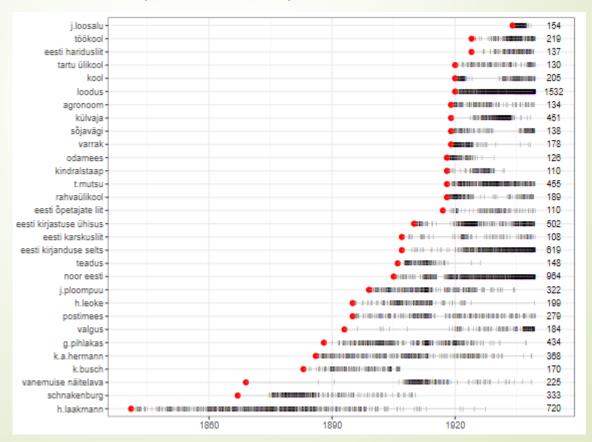


In corpus



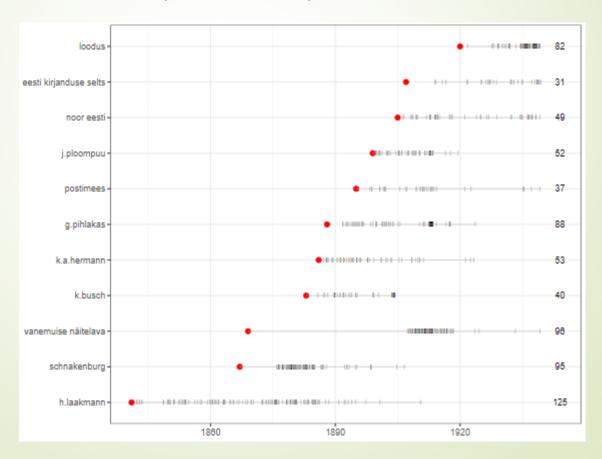
Publishers

Top publishers (harmonized)



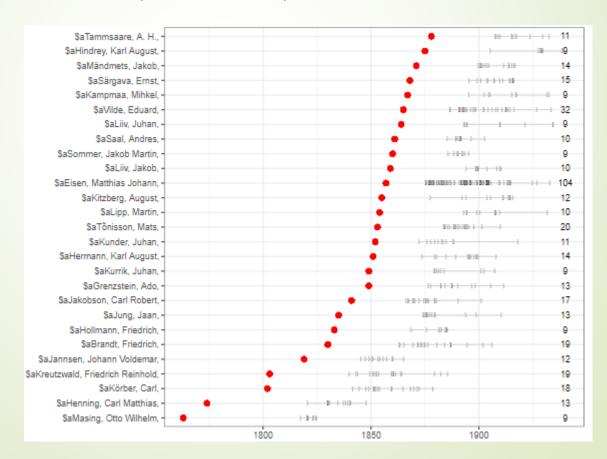
Publishers

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









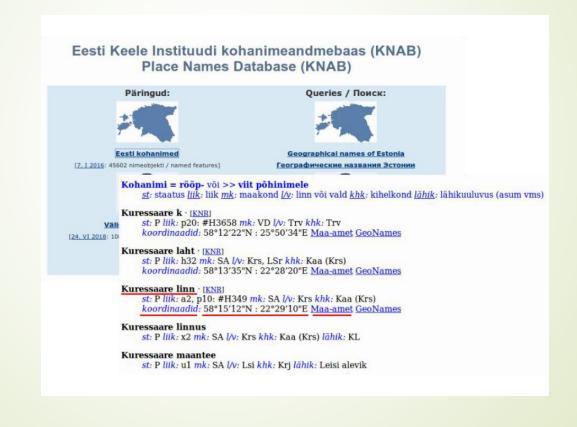
Eesti biograafiline andmebaas ISIK

ıvalehekülg - nimekiri ja otsing - ametikohad [108] - asutused [468] - isikud [4485] - kaastöö väljaannetele [1131] - nime rirkiyind [925] - pseudonüümid [9179] - sugulased [309] - sünnikohad [624] - surmakohad [186] - tegevusala [599] - lühendid ja ulikad - admin

SAAL, ANDRES	
Nimi	
Nime normitud kuju	Saal, Andres
Allikad	Postimees 1857-1907, 235-236; EAE-40, Ik. 200, EE (2000), XIV, Ik. 447.
Eludaatumid	
Sünniaeg	1861-05-21
Sünnikoht	Selja k. Tori v. Pärnumaa
Surmaaeg	1931-06-23
Surmakoht	Los Angeles
Üldandmed	
Tegevusala	kirjanik, <mark>õpetaja</mark>
Biograafilised andmed	SAAL, Andres. 21.V 1861 Tori khk Tori v Laasii. 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 Wildelmir (1904–65, ins.) [Selja v-k, Tori khkk, vallak.op eksam 1880, ITU vabakuulaja 1886–89] Eronenbergi fototehnikak Baieris, Viini paljundustehnikainst. [Selja v-k op 1880–84] Oleviku toim. 11884–90] at isinkograaf 1893–97, Tri Jaani kirk öp 1884–90, toiduainete kaupmees Trt-s 1890–93, [rotograaf Frankfurdis 1897–98, topograaf Jaaval 1898–1920, Los Angeleses 1920–31] jäi piimedaks 1928. Kirjanik. Srn 23.VI 1931 Hollywoodis, tuhastati (urn 1932 Trt Maarja kir-s, 1944 kirj.nuus-s). PmA, EBL, EE, EBL, ERDI, EKRI, EKRI, EKRI, EKRI,

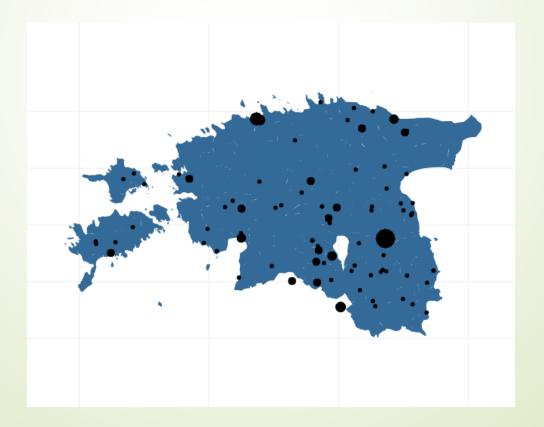
Person	Aamisepp, Julius
Geschlecht	männlich
Quelle	Pressearchiv des Herder-Instituts Marburg Wikipedia (Stand: 25.09.2018): https://en.wikipedia.org /wiki/Julius_Aamisepp
Zeit	Lebensdaten: 1883-1950
Land	Estland (XA-EE)
Geografischer Bezug	Geburtsort: Karilepa (Gem. Lääne-Harju) Sterbeort: Jõgeva
Beruf(e)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

Placename locations



Birthplaces of writers

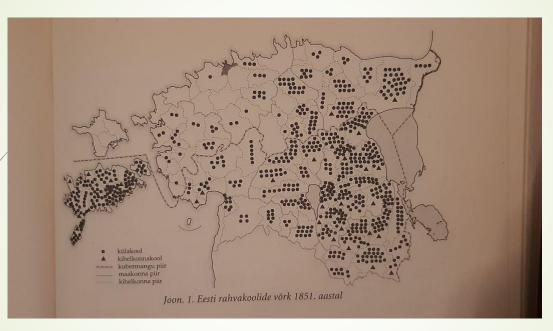
Some birthplaces of writers in corpus



Birthplaces by decade

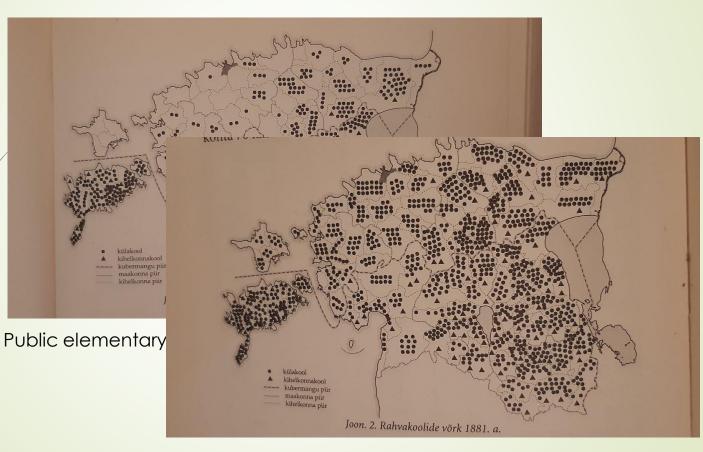


Education by decade



Public elementary school network in 1851

Education by decade



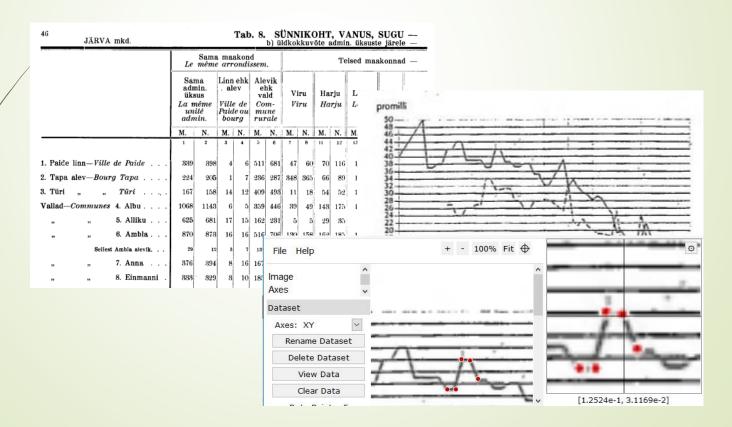
Public elementary school network in 1881

Life histories of writers

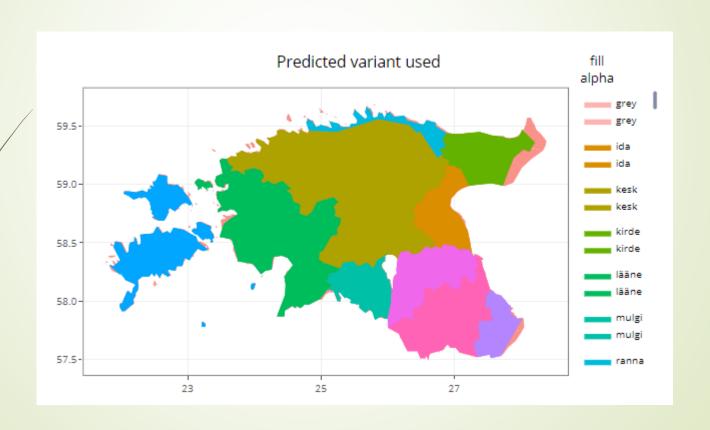


Demographic data

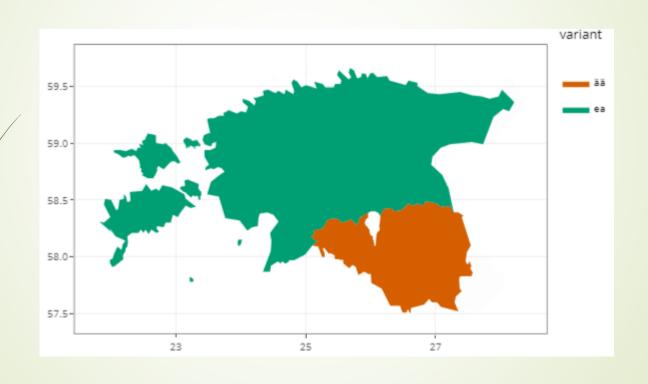
Censuses and analyses



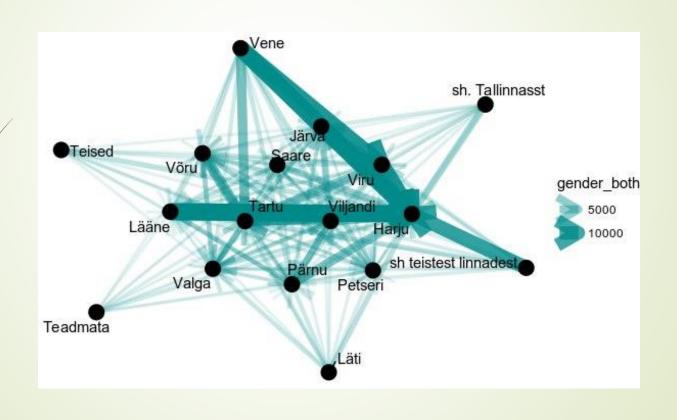
Dialect geography



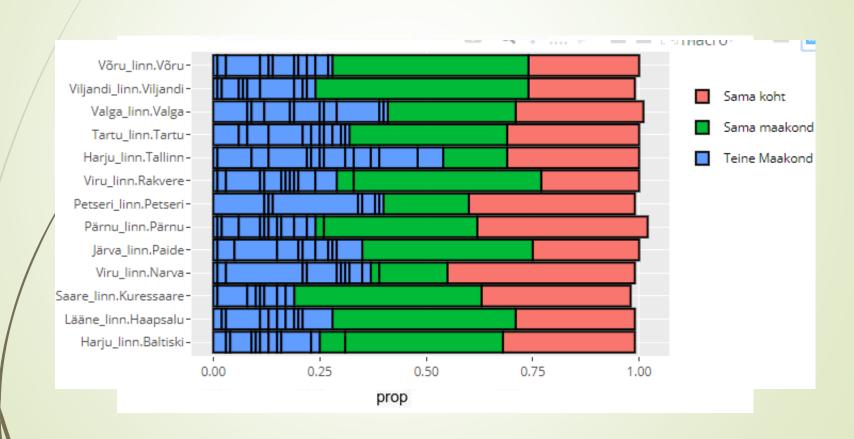
Dialect predicted values



Dialect compositions in cities



Migration and dialects



Getting to research

Operationalizing the question

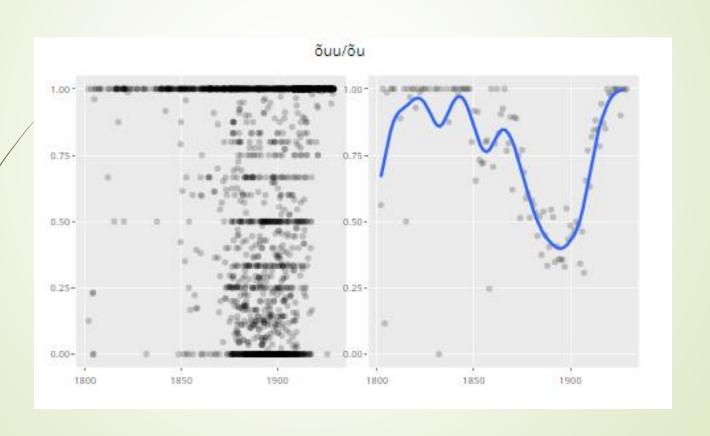
Making the study

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

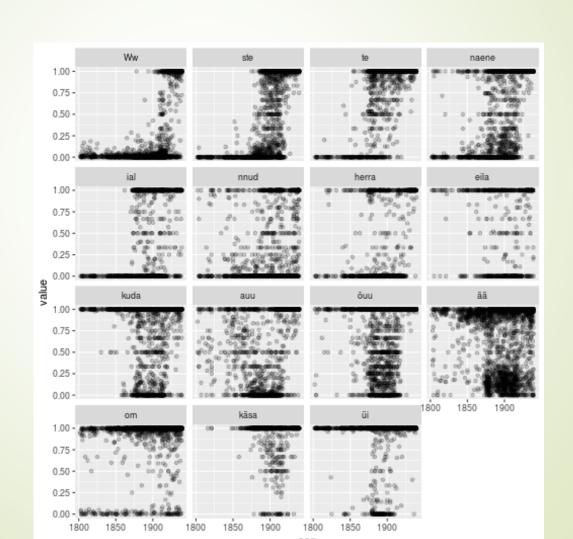
Variants and variation

- 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

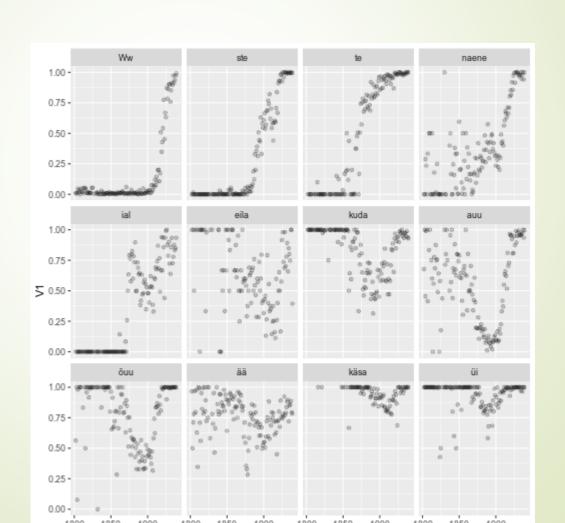
Variants and variation



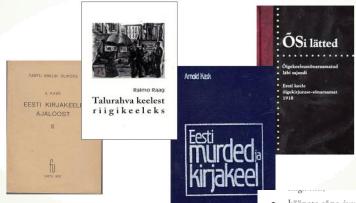
Study on particular variants



Average trends



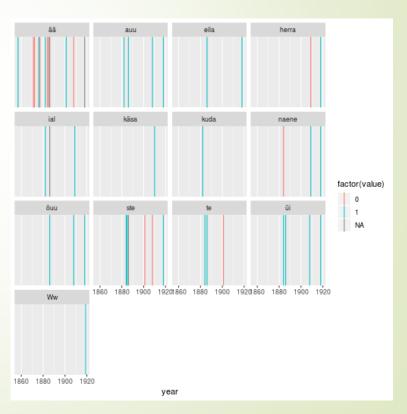
Historian's overviews



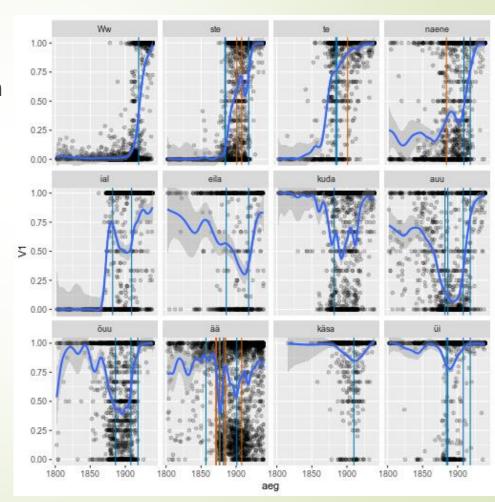
- käänata sõna juus (=juuksekarv) järgmiselt: juukse, juuksed, juuste;
- käänata sõna kodu ainsuse omastavas kodu ja sisseütlevas koju;
- kirjutada üü, mitte üi sõnades nagu nüüd, hüüdma, tüütama, süütama;
- kirjutada kaine, naine, laine, mitte kaene, naene, laene, aga naeris, paene (sónast paas), vaene; 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;

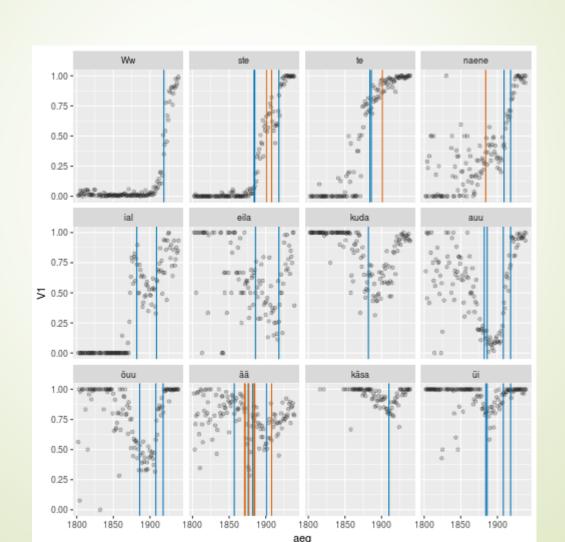
Overview of proposed norms in 1909 (Raag 2008)

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



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



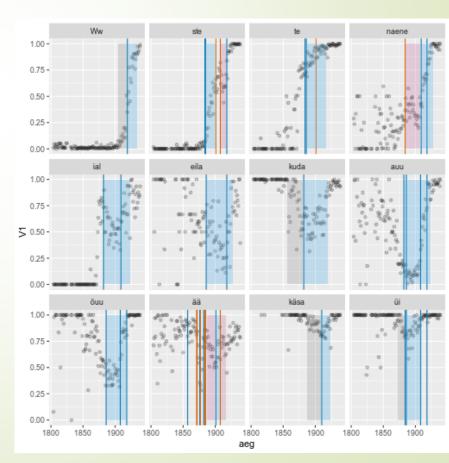


Mechanisms of standardization

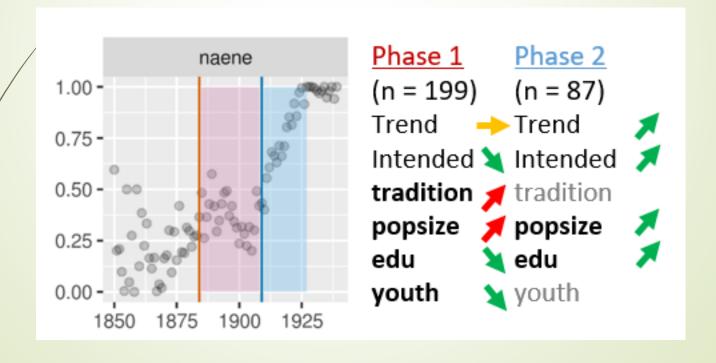
- Influences on language usage
 - Home dialect
 - Everyday language use
 - Education
 - Prescription
 - Etc
- What made a suggestion successful?
- Generational change vs individual change?

Phases of change

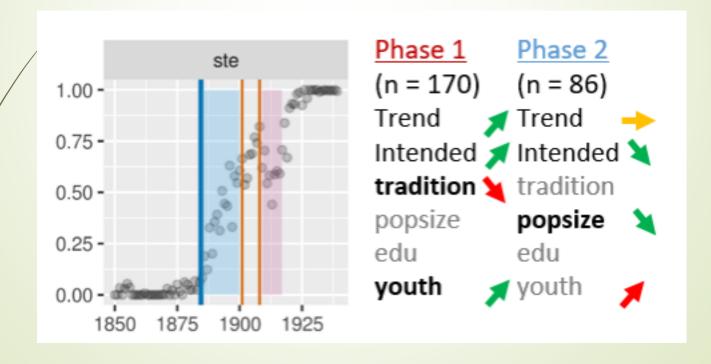
- Looking at the aftermath of prescriptive events
- Blue (suggested norm towards 1)
- Red (suggested norm towards 1)



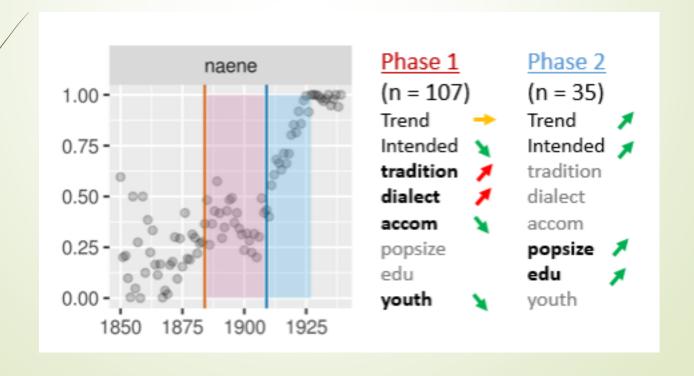
- Significant predictors (logistic regression)
 - Model does not include dialect info



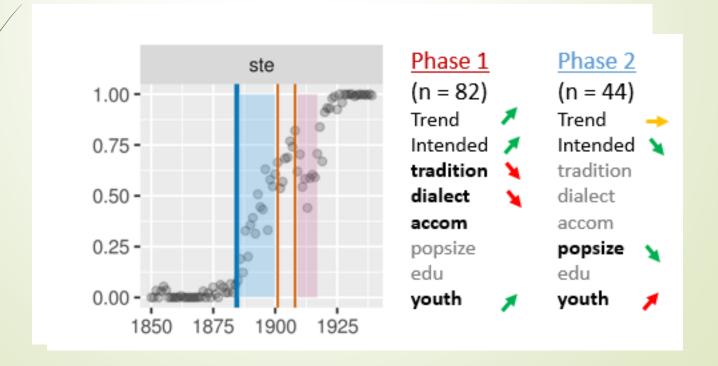
- Significant predictors (logistic regression)
 - Model does not include dialect info

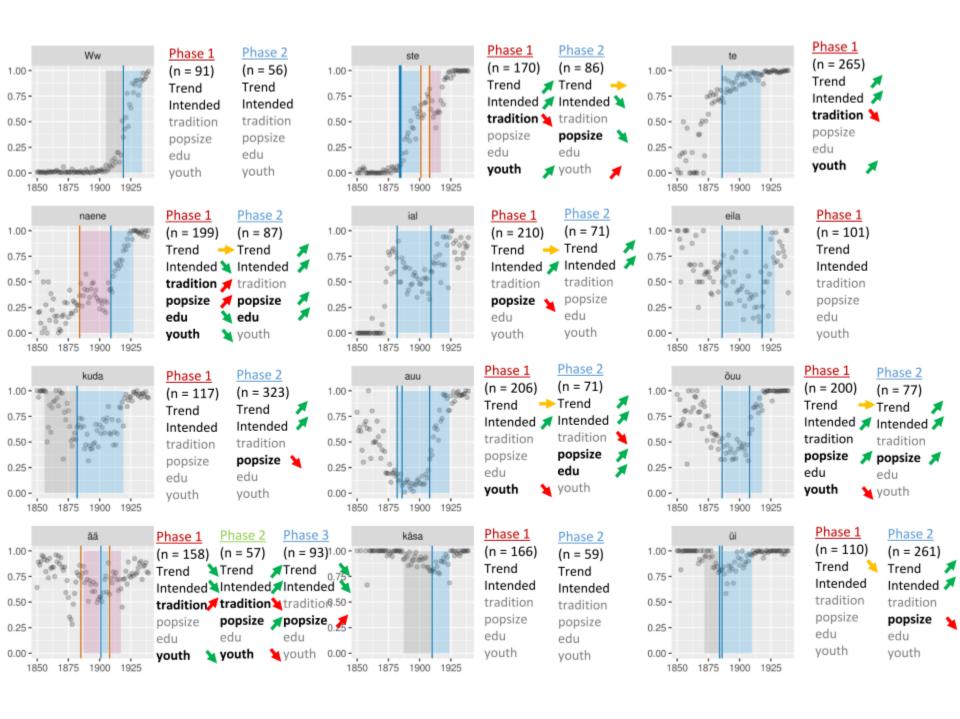


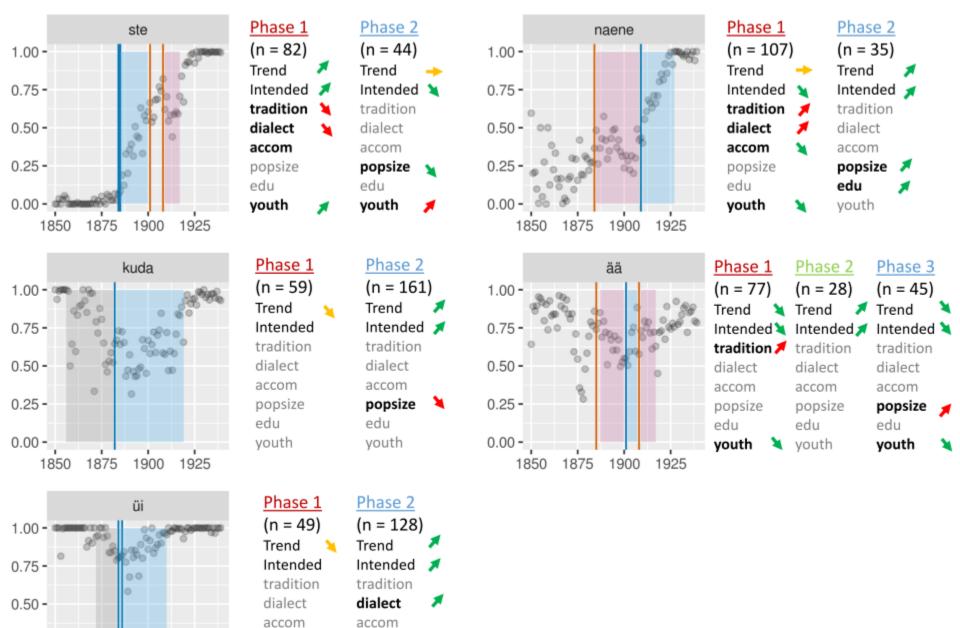
- Significant predictors (logistic regression)
 - Model includes dialect info



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







popsize

edu

youth

popsize edu

youth

0.25 -

0.00 -

1850

1875

1900

1925

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

- https://peetertinits.github.io/slides/DH-labs-2019/prese_DH-lab_2019.html#14
- https://peetertinits.github.io/slides/EDHC2018/prese E DHC_2018.html#45
- https://peetertinits.github.io/slides/EDHC2018/prese_E DHC_2018.html#50
- https://peetertinits.github.io/slides/plots/authorbirths.ht ml
- <u>https://peetertinits.github.io/slides/plots/samamaakon</u> <u>d.html</u>