

# **Eesti kirjakeele kihiline korpus 1800-1940**

Peeter Tinit

ERÜ aastakonverents. 2023, Tallinn



# Ajalooline sotsiolingvistika

## Mineviku keelekogukonnad ja korpused

- Kes, mida, miks rääkis/kirjutas
- Keelemuutuste mehhanismid
  - Varjatud keeled ja kogukonnad (Vandenbussche, Elspaß 2007; Havinga, Langer 2015)
  - Muutused läbi elu (Evans 2013, Petre et al. 2019, Schiegg 2022)
- Me ei tea kui erinev minevik oli (Labov 1994)
- Aga meil on hulk andmeid, mille põhjal seda uurida



## *The Handbook of* **Historical Sociolinguistics**



*Edited by*  
**Juan M. Hernández-Campoy and**  
**J. Camilo Conde-Silvestre**

Y Blackwell



The Historical  
Sociolinguistics Network

# Digiteeritud tekstid

~20-30% teostest digiteeritud



Tarto Piiblikogodusse Peaseltsi  
Arrotegginne Marahwa  
abbiseltsidele kirjutud  
Carl Körber

## TEHNILISED ANDMED

Ilumisaeg: 1845

Keel: eesti

Depositor:

Eesti Rahvusraamatukogu

Kasutusemärke:

Zeutschel book scanner

OS14000A2

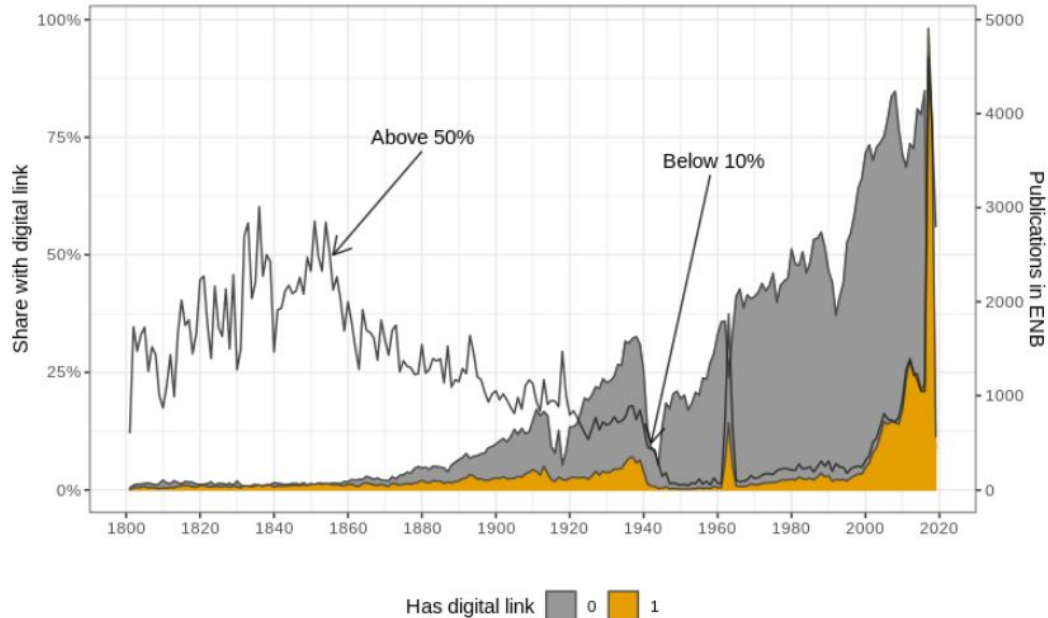
Laad: raamat

ESTER b16315017

ISBN 978-9949-897-07-0 (pdf)

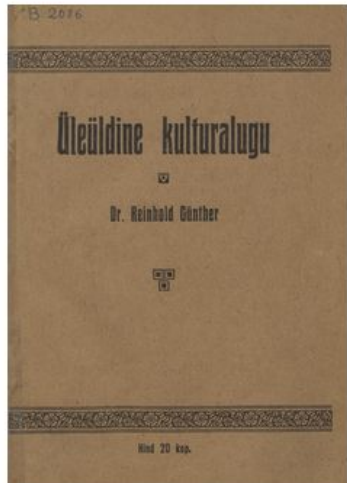
Püsilink: <http://www.digar.ee/id/nlib-digar:332604>

Digital copies linked in Estonian National Bibliography



Paiknevad paljudes eri kollektsioonides:  
DIGAR, KIVIKE, UT-DSPACE, VAKK, ETERA, VIKITEKSTID jne

# Digiteeritud tekst



## Üleüldine kulturalugu

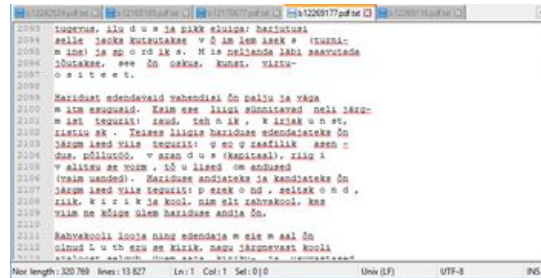
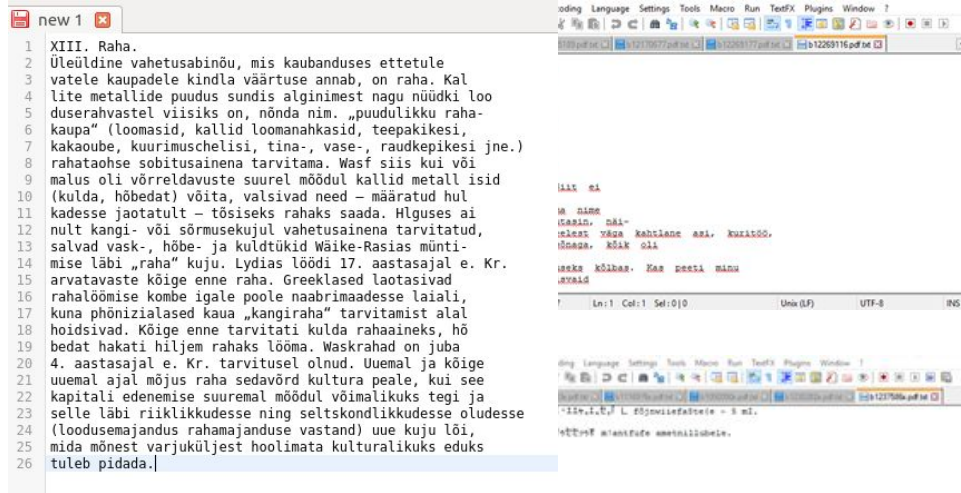
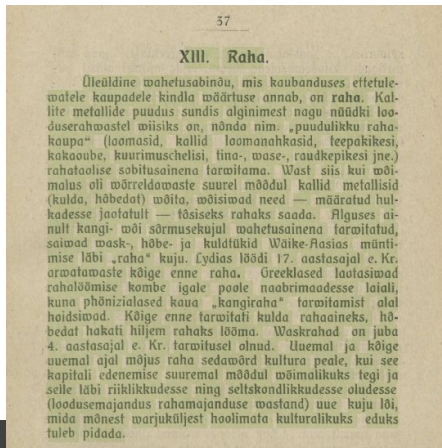
Reinhold Günther

### TEHNILISED ANDMED

Kirjastaja: [s.n.]  
Ilumumiseaeg: 1911  
Keel: eesti  
Depositoor:  
Eesti Kirjandusmuuseum  
Rahastus:  
Digiteerimist rahastas Euroopa  
Regionaalarengu Fond

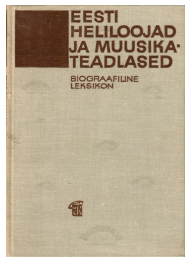
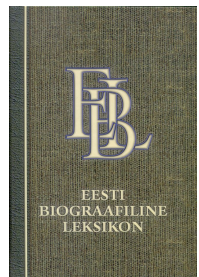
Laad: raamat  
Vabakasutus -  
https://creativecommons.org  
/publicdomain/mark/1.0/  
ESTER 1:4657459  
ISBN 9789916135457 (pdf)

Püsilink: <http://www.digar.ee/id/nlib-digar:450799>



# Metaandmed

Autorite kohta on infot kogutud bibliograafiates, biograafilistes leksikonides, entsüklopeediates.



Väljavõtte ISIK biograafilisest leksikonist:

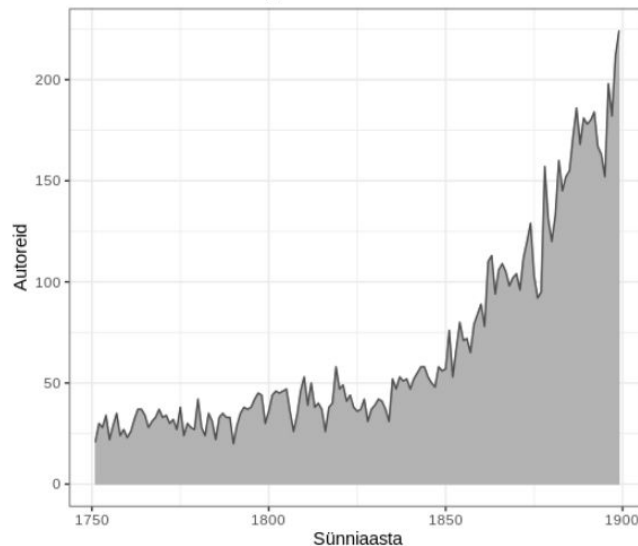
## Julius Aamisepp

1.IX 1883 Harjumaal Kloostri valla Karilepa-Tõnul. Taluperemees Siim A., Liisa Vraeger. 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, EBL, EAT2, ENE1, ENE2, EABL, Vjv, ENE2(14), ETeadBL – ISOTAMM 2

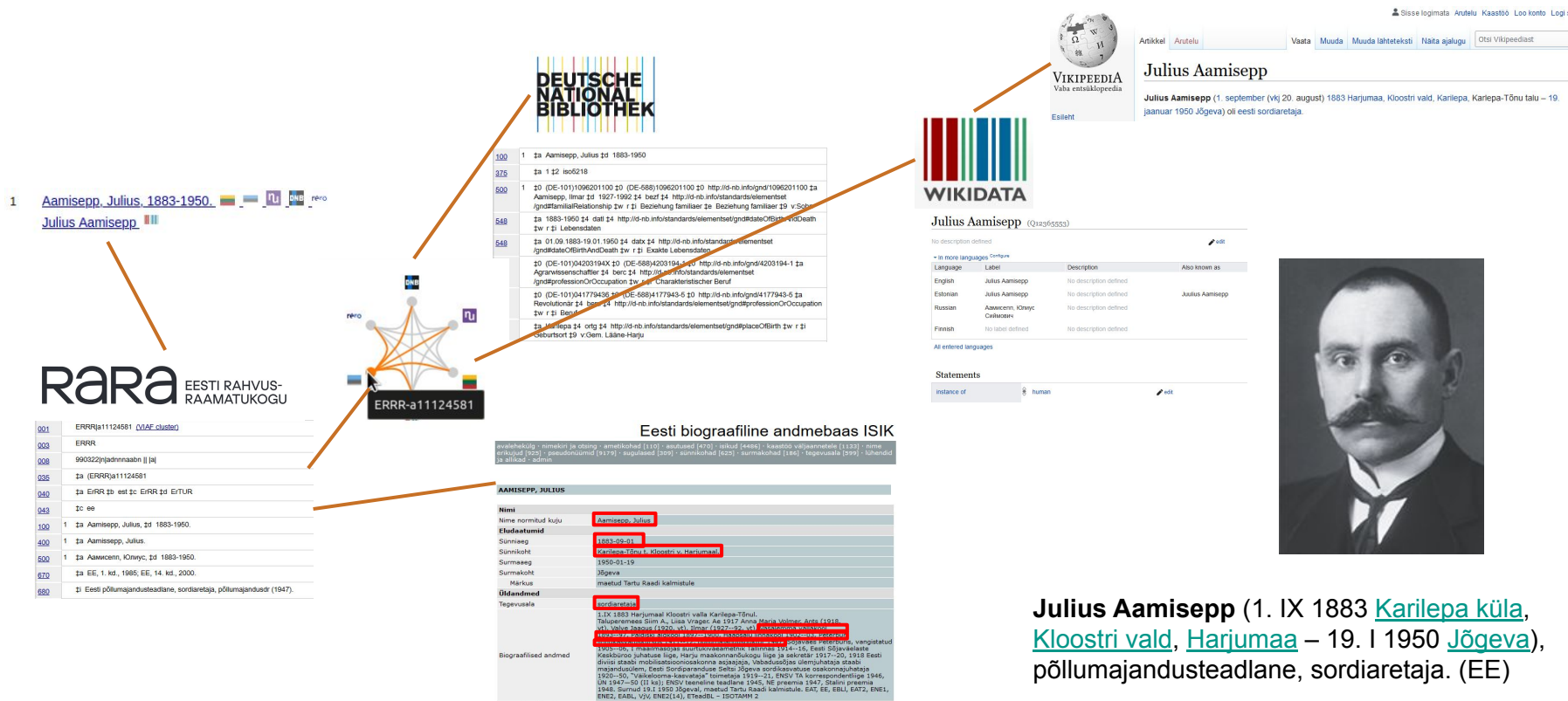
**RARA** EESTI RAHVUS-  
RAAMATUKOGU



Autorid rahvusbibliograafias



# Andmestike ühendamine



**Julius Aamisepp** (1. IX 1883 [Karilepa küla](#), [Kloostri vald](#), [Harjumaa](#) – 19. I 1950 [Jõgeva](#)), põllumajandusteadlane, sordiaretaja. (EE)

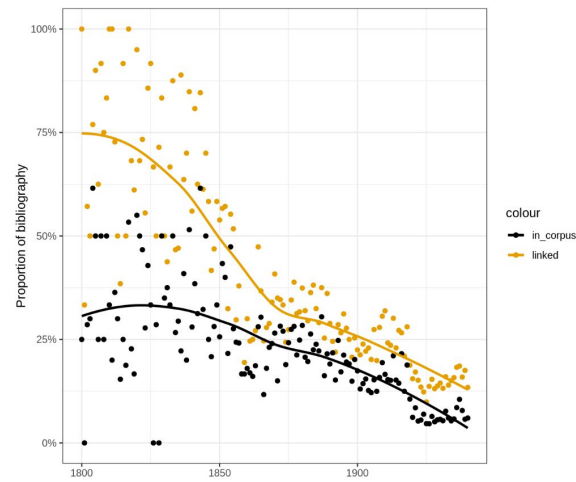
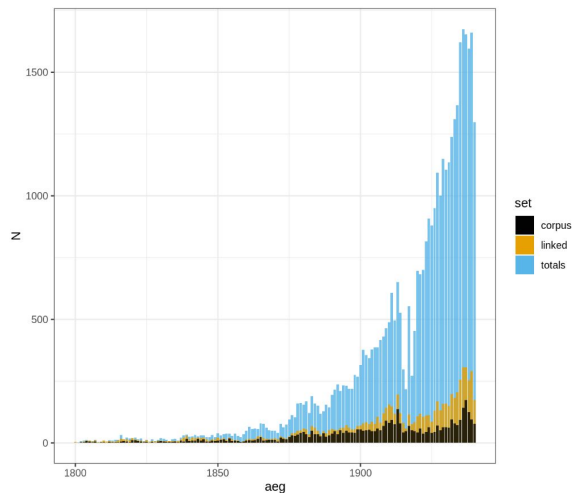
# Korpuse sisu 1

Kogutud tekstihulk:

- 4412 teksti
- 1188 autorilt
- 52M sõne
- 11% ajastu kate ERBis

Tehnilist:

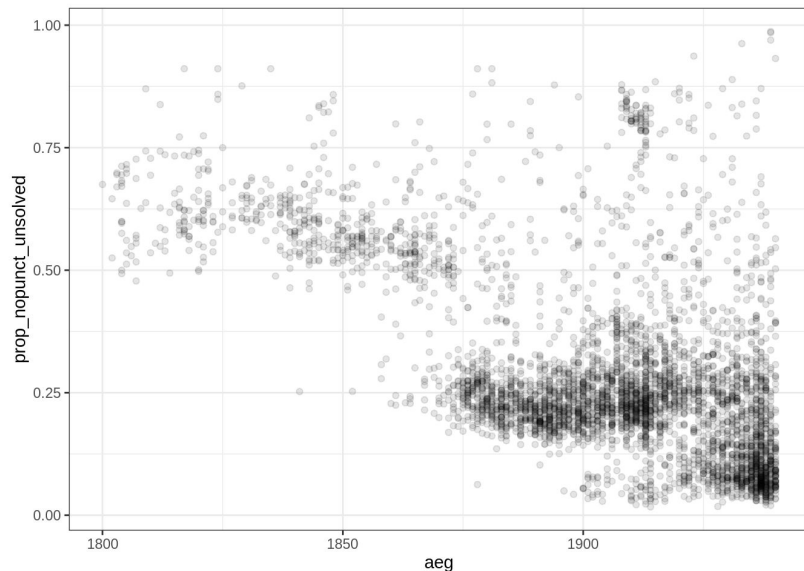
- 615Mb toortekstid
- 6,5Gb märgenduskihiga



# Korpuse sisu 2

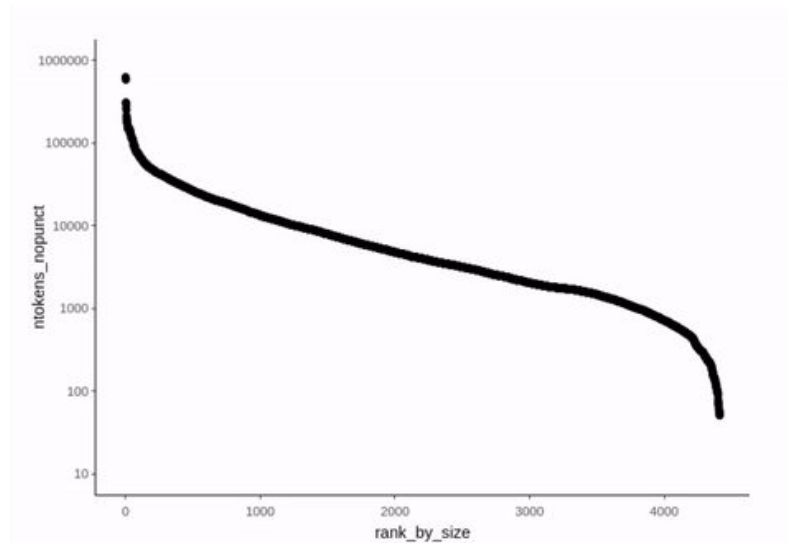
Kvaliteet ja kirjaviisid:

Tuvastatud sõnede hulk tekstis (EstNLTK)



Tekstide jaotus korpuses

pikim 662 000, mediaan 3994, lühim 51 sõne



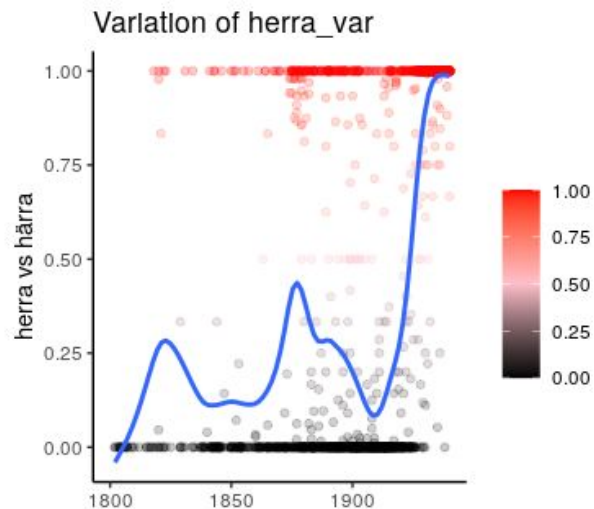
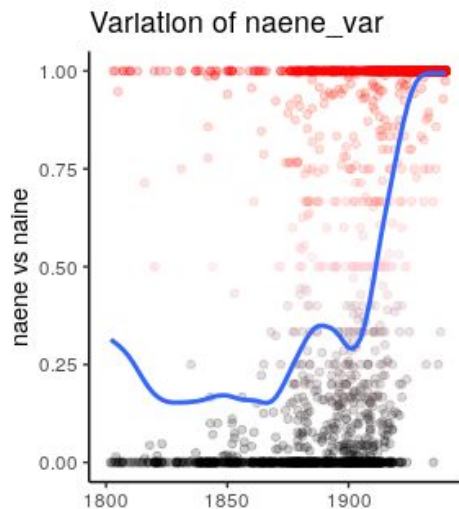
# Näidisjuhtumid

Mida saab uurida?

Nt keeleline muutuja  
(konkureerivad keelekujud):

- naene - naine
- herra - härra

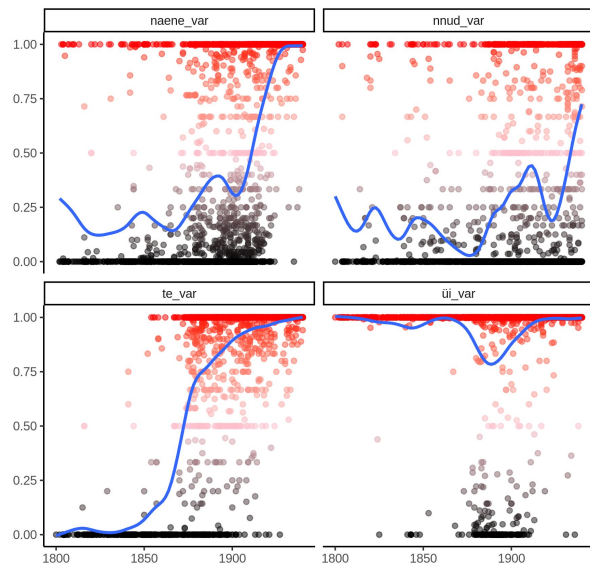
Kuidas vormi kasutus on  
muutunud ajas?



# Näidisjuhtumid 1

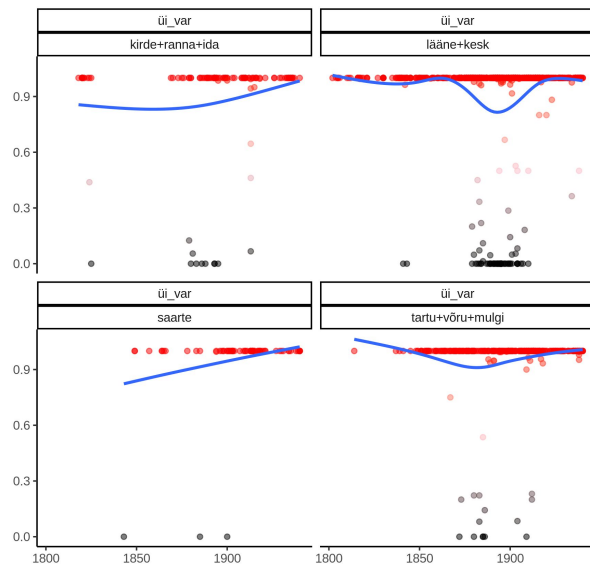
Pikaajalised muutused:

naene-naine, nnud-nud (annud-andnud), te-ti (õiete-õieti), üi-üü (nüid-nüüd)



Metainfoga kombineerimine:

Nüid vs nüüd: nüid on murretes lääne ja keskmurde tunnus. Võtame autori kodumurrete kaupa.



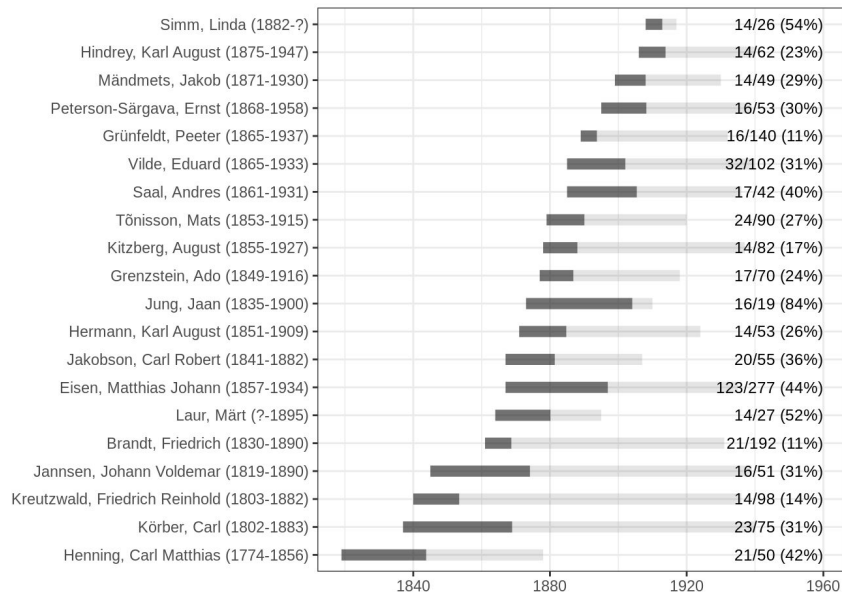
# Näidisjuhtumid 2

Mõned uuringud on püüdnud vaadelda, kuidas asetub indiviid keelemuutuste konteksti.

- Kuninganna Elisabeth I vahemikus 1544–1603 (Evans 2013)
- Top 50 kirjanikku inglise keeles 1623-1757 (Petre et al. 2018)

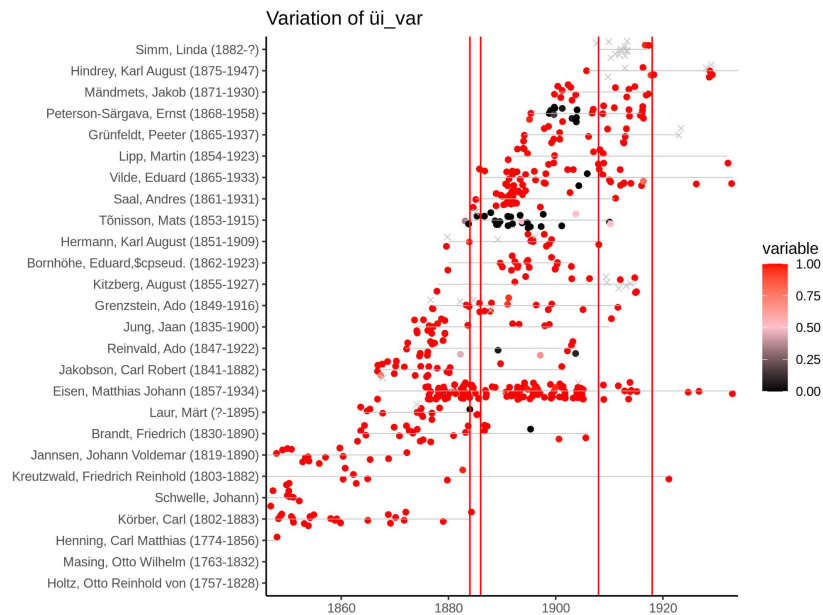
Kuidas muutub keelekasutus autori eluea jooksul?  
Mis tingib muutuseid?

Korpuse top 20 autori teoseid ilmunust

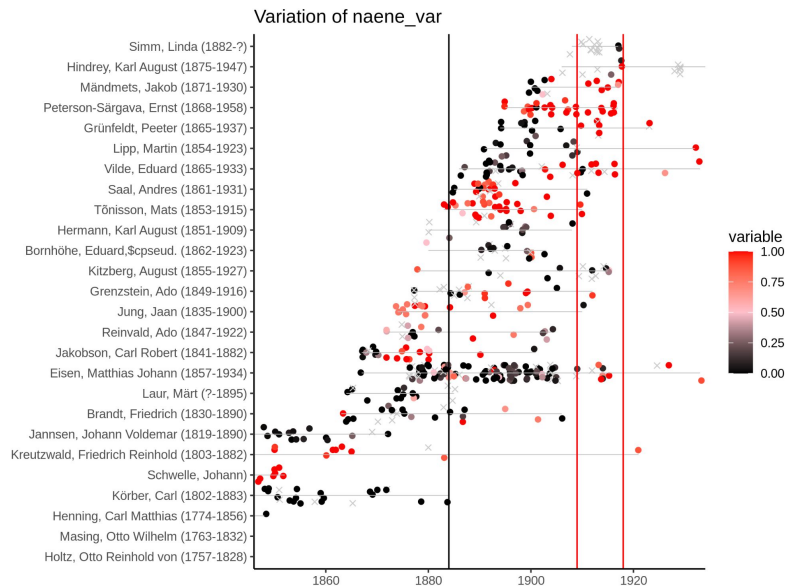


# Näidisjuhtumid 2

## Top 20 autorit ja nüüd-nüüd



## Top 20 autorit ja naene-naine



# Näidisjuhtumid 3

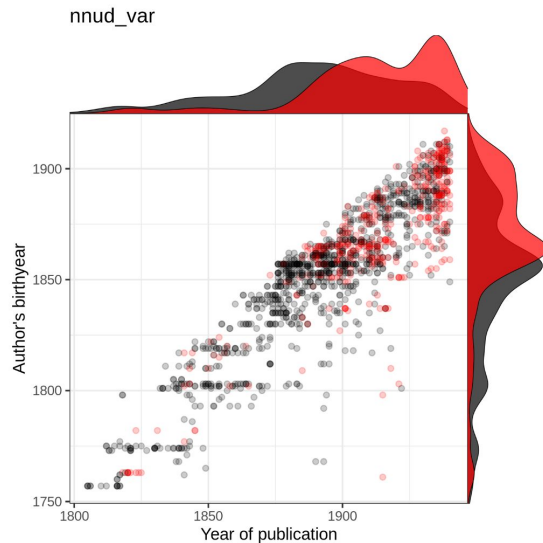
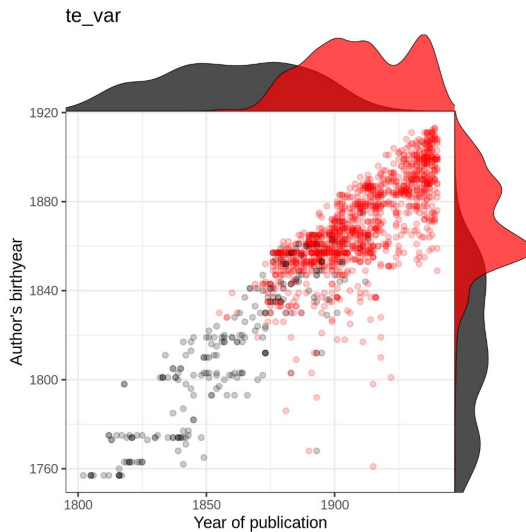
Keelemuutustes võib rolli mängida põlvkondade vahetus

- Muutused võivad toimuda sujuvalt põlvkondade üleselt
- Vanemad põlvkonnad võivad püsida konservatiivsed ja uued põlvkonnad veavad muutust

Kas kasutust näeme autori sünniaastalt või teose ilmumisaastalt?

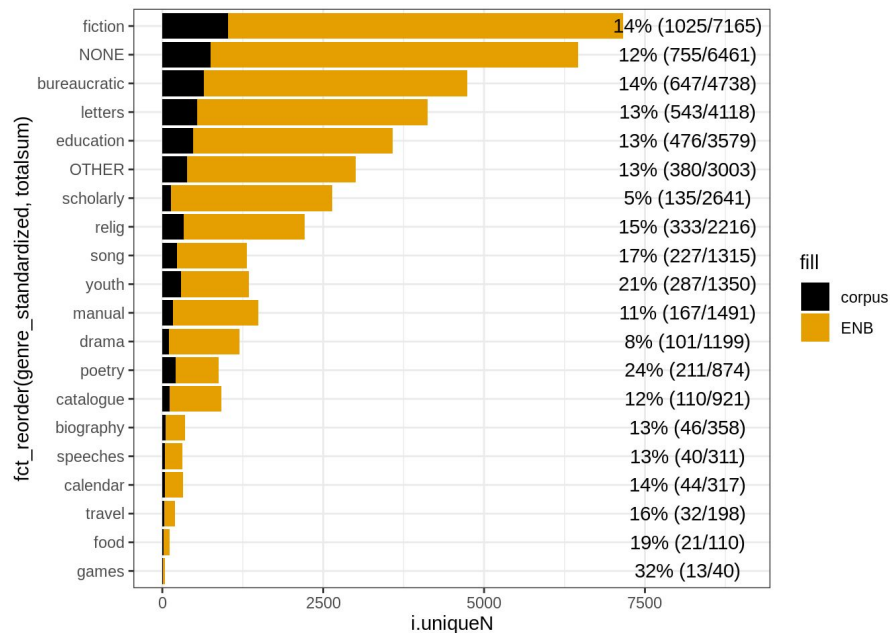
- 685 autoril on sünniaasta
- 2561 teosel on autori sünniaasta

Vaatame te-ti ja nnud-nud muutujaid.

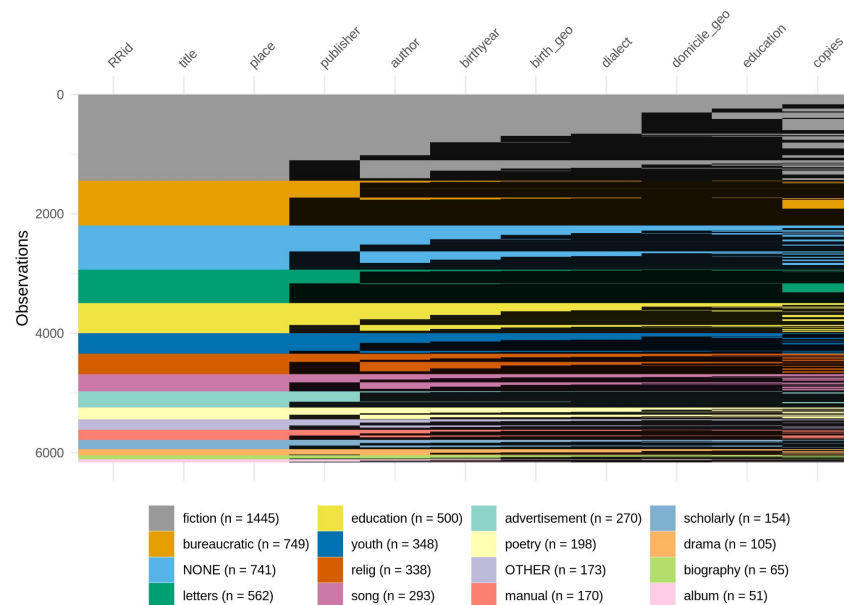


# Zanriline koosseis

## Žanride esindatus korpuses vs bibliograafia

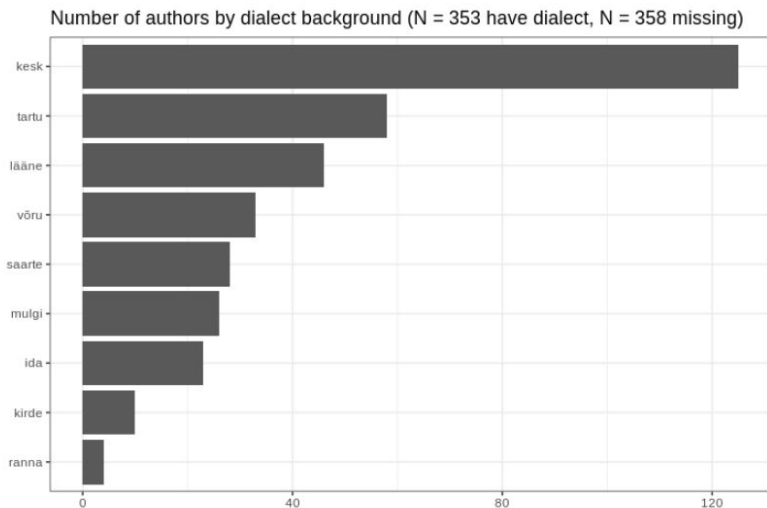


## Žanrid ja metainfo olemasolu

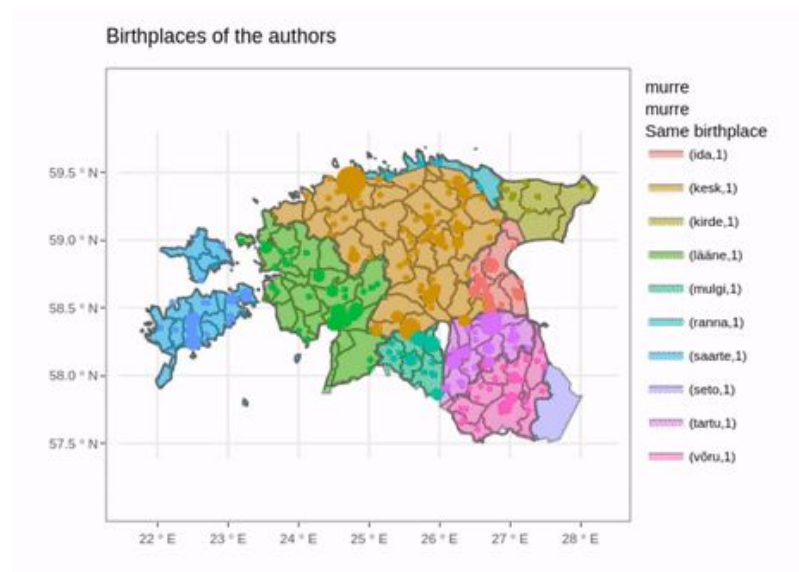


# Autorid korpuses

Autorite murdetausta jaotus

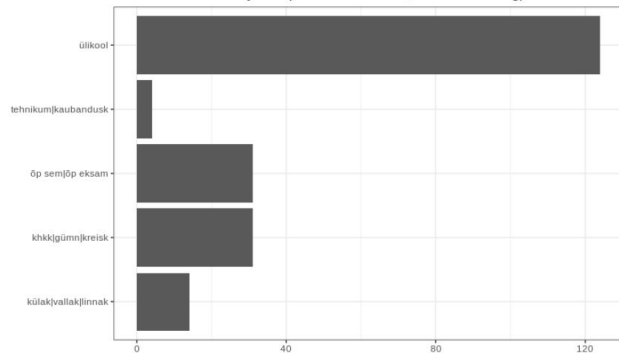


Autorite sünnikohad ja murdetaust kaardil

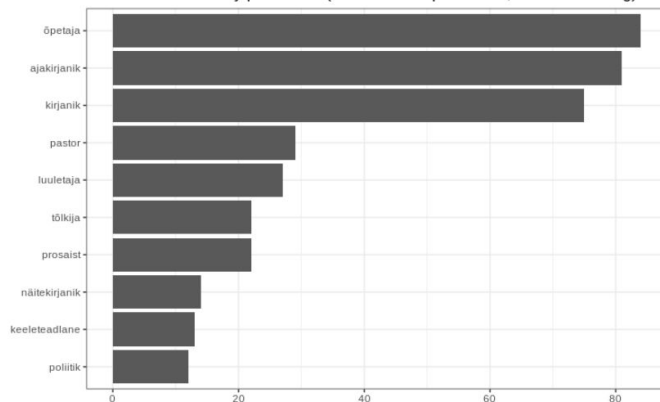


# Autorite elukäik

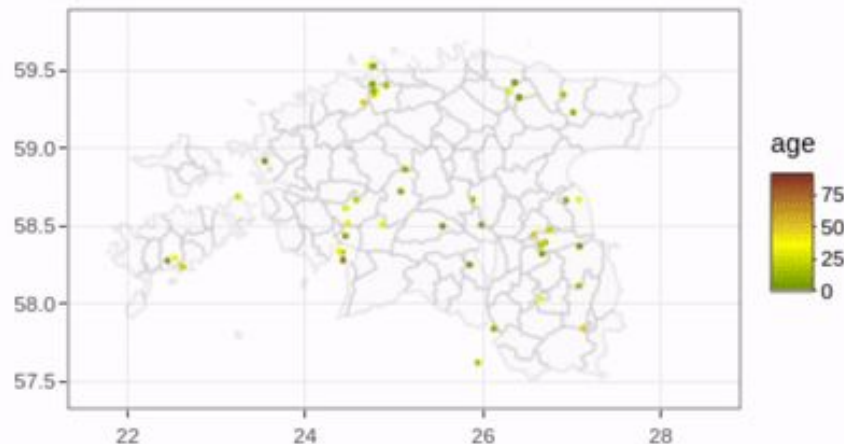
Number of authors by edu (N = 204 have edu, N = 507 missing)



Number of authors by profession (N = 407 have profession, N = 304 missing)



Korpuse autorite elukäik



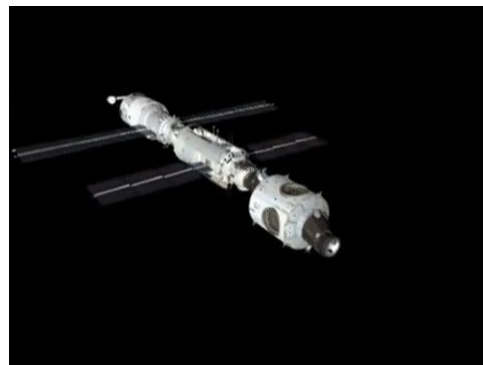
Play

~year: 1851

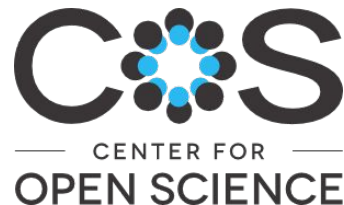
1801 1817 1833 1849 1865 1881 1897 1913 1929

# Arutelu avatud teadusest

- Kuidas saame kasvavad digikogud vastastikku tööle panna?
  - Kas suurte digikogudega peab uurija hindama ise ainese representatiivsust ja valima sobivad osad?
  - Kas ja kuidas saaksime metainfot parandada töö käigus?
- 
- Seda andmestikku on mõte jooksvalt parandada ja täiendada. Vaata <https://osf.io/zbup2/>



International Space Station 1998-2010  
([Internet Archive](#))



# Kokkuvõtteks

- Korpuses on praegu 4412 teksti, 1188 autorilt, 52M sõne, 11% esindatus 1800-1940.
- Korpus on tasakaalustamata, koondatud on kättesaadavat infot, piirangutega arvestades saab sellega juba teha hulk tööd.
- Andmekvaliteedi parandamisega on palju tööd ees. Abi, nõu ja jõud on väga oodatud. Kirjutage [peeter.tinits@ut.ee](mailto:peeter.tinits@ut.ee)!

Täna tähelepanu eest!

doi:10.5128/ERYa19.11

## STRATIFIED HISTORICAL CORPUS OF ESTONIAN 1800–1940

Peeter Tinits

**Abstract.** The article introduces a stratified historical corpus of Estonian 1800–1940. A stratified corpus will allow for sociolinguistic comparisons of language use between past authors, considering their background and biographical details (e.g. native dialect area, age cohort, attained education) or the publication details (e.g. genre of publication or publisher). The corpus assembles texts from a number of different public archives and combines it with metadata on their publication details and the author's background. The corpus at the moment of publication consists of 4,412 works from 1,188 author names, constituting 11% of the works registered in the Estonian National Bibliography from 1800–1940. The author names are associated with biographical information where possible. Three use cases on studying orthographic variation are introduced as examples where the corpus can help study past language communities. The corpus is published online to allow updates as data is improved and more texts are digitized.



Korpus: <https://osf.io/zbup2/>