Metadata from ENB

ENB metadata

Abstract

Estonian National Bibliography is a metadata set that aims to collect information on all publications written in any language in Estonia and all texts written in Estonian in whichever country. The dataset has been compiled in digital format since 2002 and aggregates the work of multiple institutions and generations in collecting the publication information.

The data here has been extracted on Jan 27 2022.

This dataset presents the Estonian National Bibliography dataset in wide instead of long format used in Marc21, with some of the variables that may be useful for text-mining studies. It includes the following information

- publication ID in National Library
- time of publication (aeg)
- place of publication (koht_raw)
- partly standardized place of publication (koht)
- publisher (kirjastus_raw)
- partly standardized publisher information (kirjastus)
- title (title, subtitle, comptitle)
- author (name, date of birth,id)
- other associated authors (translator, editor etc)
- number of copies printed
- font information
- keyword
- genre
- link to fulltext

Information on coding the variables can be found here: [meta file on github]

Helpful information on the metadata available can be found here: http://data.digar.ee/#page5

The rules followed in adding information on older books can be found here: https://www.elnet.ee/wiki/lib/exe/fetch.php?media=kataloogimine:vanaraamatmarc_2019.pdf

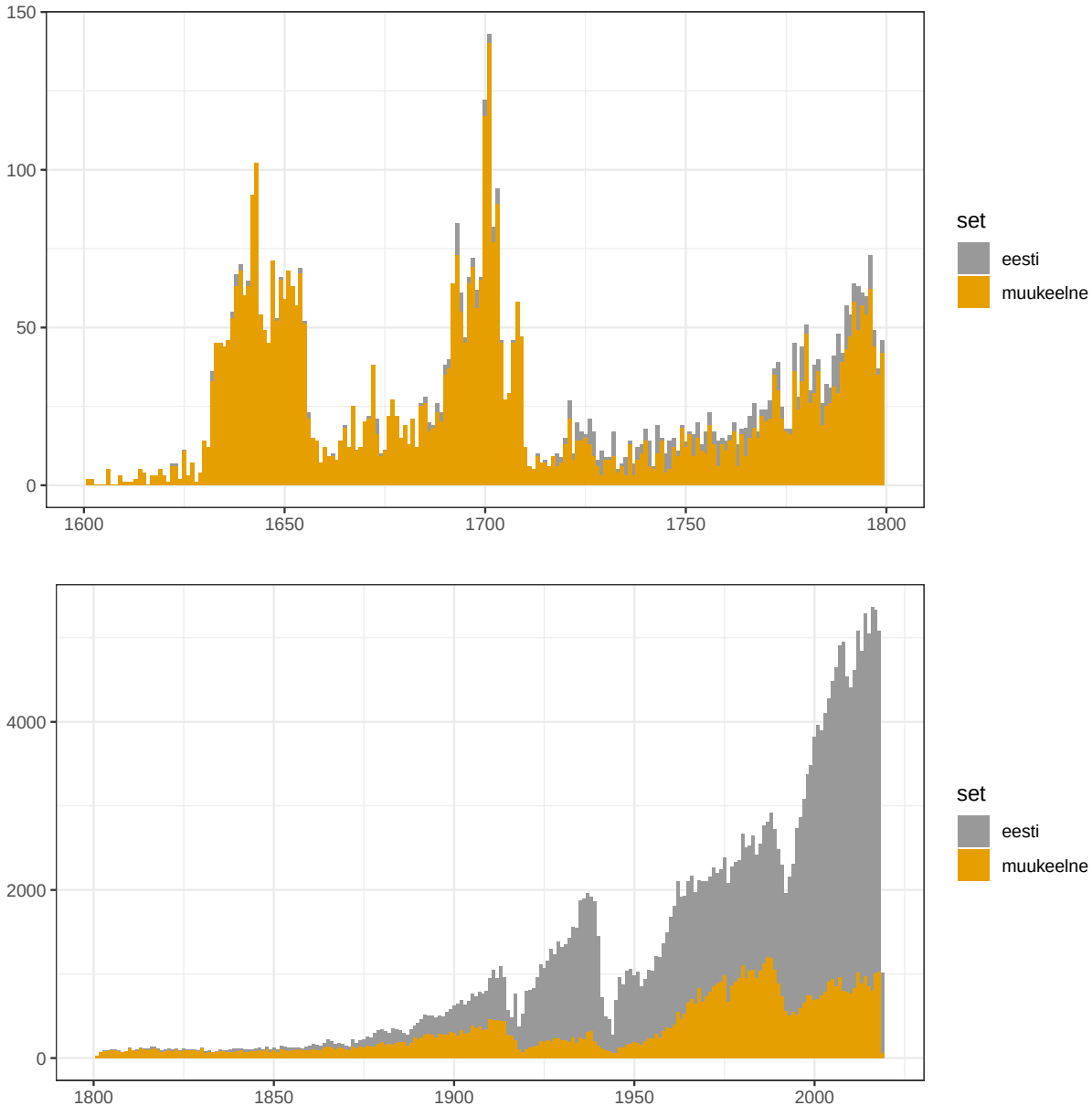
Intro

The current code organizes the ENB metadata into a tidy format: with one publication per line, and with the basic metadata information distributed into few relevant categories. This has been done to facilitate further data analytic steps, particularly when someone may be unfamiliar with Marc21 database structures.

Summary

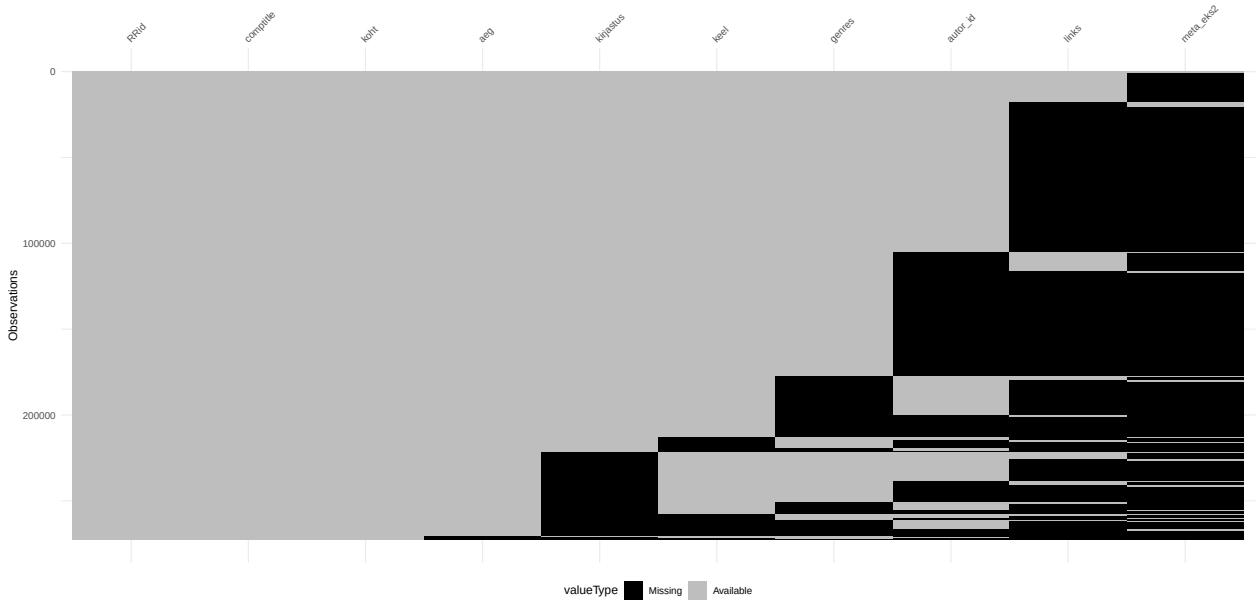
The dataset has altogether information on 272411 printed items. The coverage has been estimated to be better than 95% for all of the relevant published works.

The dataset is divided into two: works in Estonian language, and works in other languages. The two sets are displayed over the year of publication here in different colors.



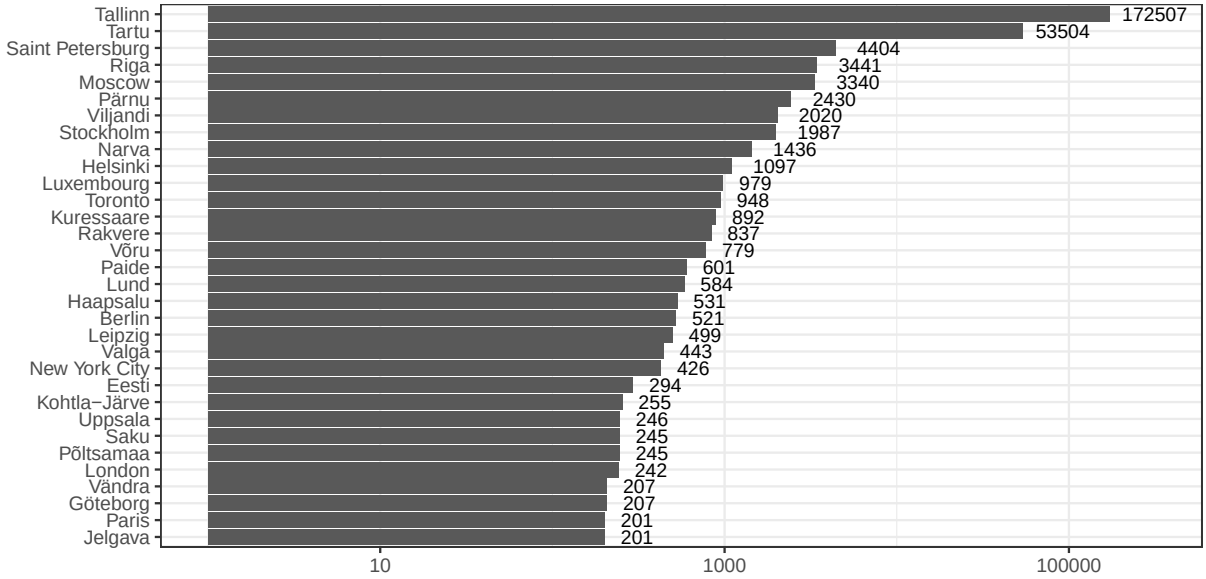
Coverage of metainformation

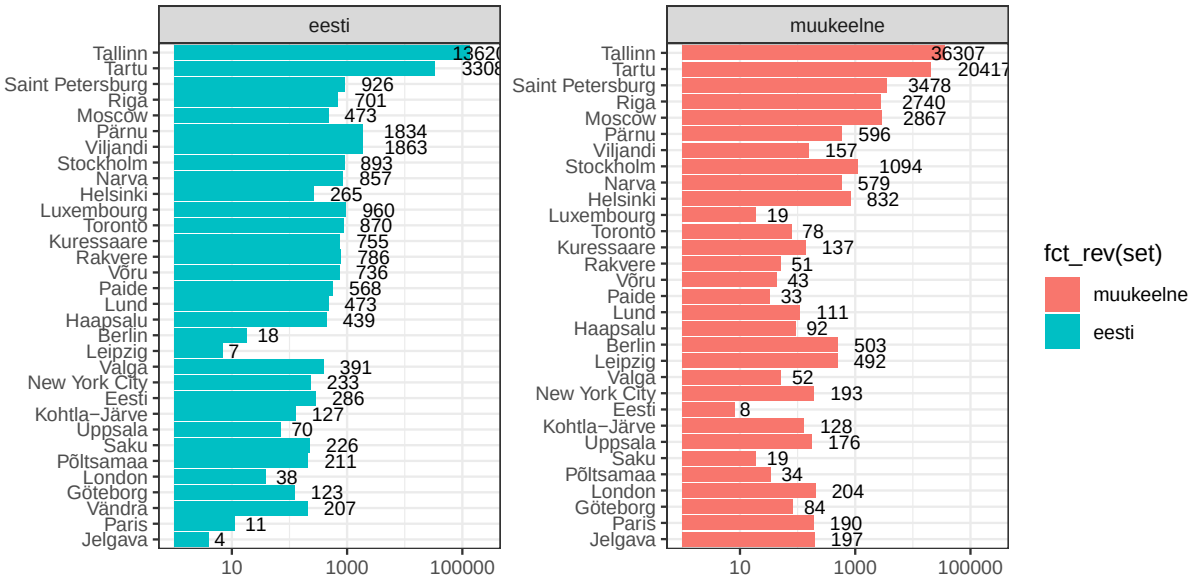
An overview of the coverage of the data is given below. Grey areas indicate datapoints that do have the information of that column, black areas the datapoints that do not have the information of that column. For example 82% of the books have information on the publisher, however just 59% of the books have information on the author. 16% of the books have a link to an online digital copy.



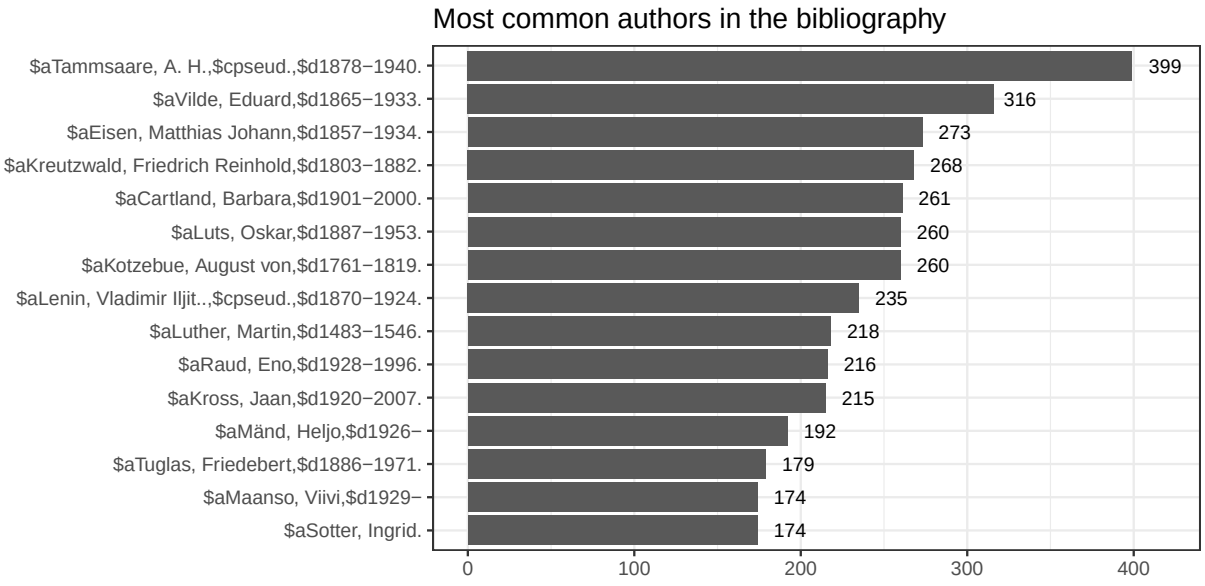
Cities

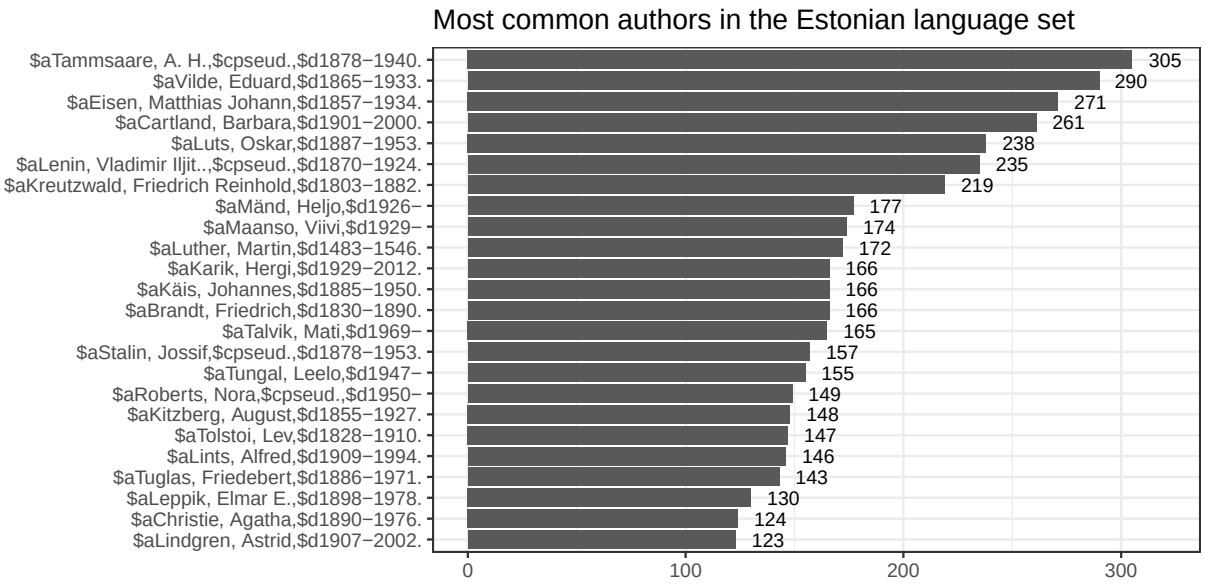
The city names where the works have been published have been harmonized manually and through a few heuristic algorithms. The tokens that appear more than 40 times should be mostly harmonized, while rarer tokens have been included only through algorithmic processing. Depicted below are the most common cities of publication and the number of publications in them separately for language sets.

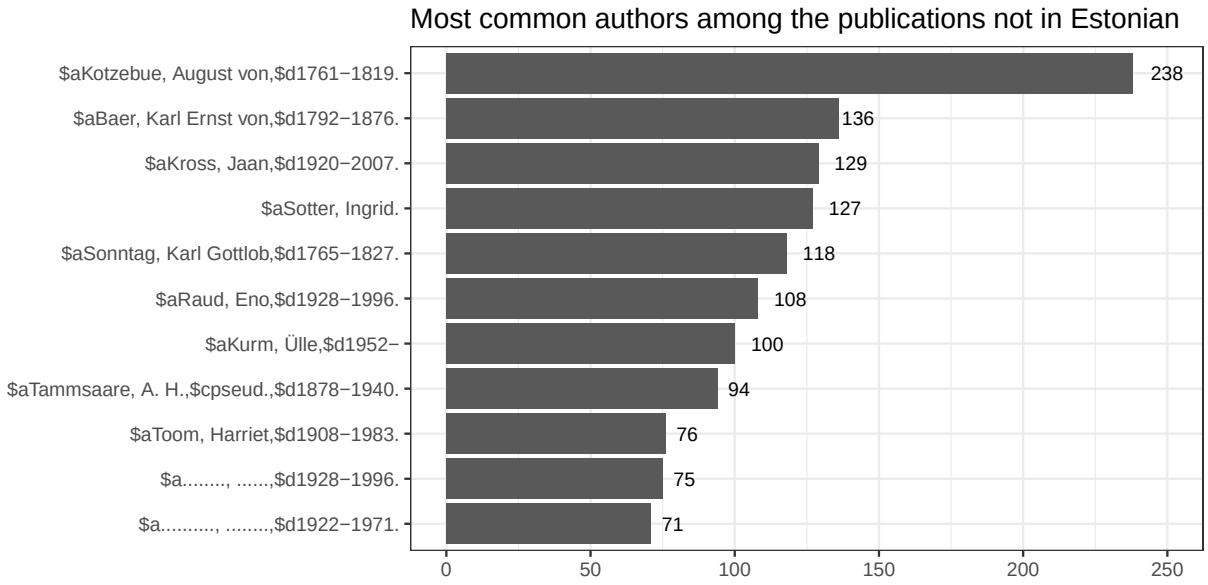



eesti
muukeelne
Tallinn
Tartu
Saint Petersburg
Riga
Moscow
Pärnu
Viljandi
Stockholm
Narva
Helsinki
Luxembourg
Toronto
Kuressaare
Rakvere
Võru
Paide
Lund
Haapsalu
Berlin
Leipzig
Valga
New York City
Eesti
Kohtla−Järve
Uppsala
Saku
Põltsamaa
London
Göteborg
Vändra
Paris
Jelgava
13626
3308
926
701
473
1834
1863
893
857
265
960
870
755
786
736
568
473
439
18
7
391
233
286
127
70
226
211
38
123
207
11
4
36307
20417
3478
2740
2867
596
157
1094
579
832
19
78
137
51
43
33
111
92
503
492
52
193
8
128
176
19
34
204
84
190
197
10
1000
100000
fct_rev(set)
muukeelne
eesti

Authors

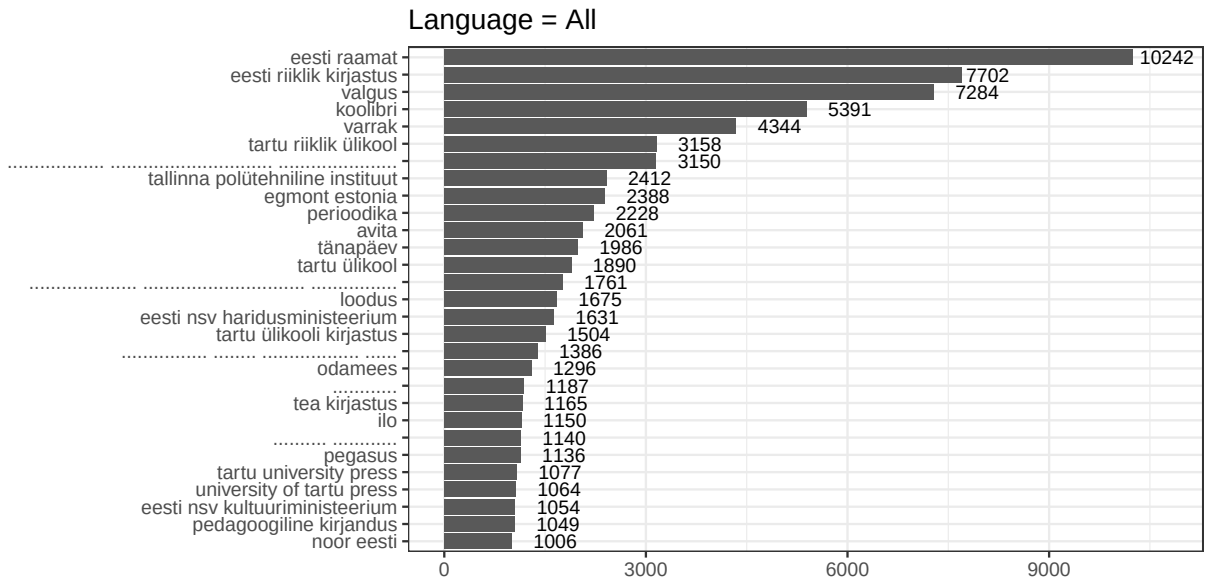

Most common authors in the bibliography
$aTammsaare, A. H.,$cpseud.,$d1878−1940.
$aVilde, Eduard,$d1865−1933.
$aEisen, Matthias Johann,$d1857−1934.
$aKreutzwald, Friedrich Reinhold,$d1803−1882.
$aCartland, Barbara,$d1901−2000.
$aLuts, Oskar,$d1887−1953.
$aKotzebue, August von,$d1761−1819.
$aLenin, Vladimir Iljit..,$cpseud.,$d1870−1924.
$aLuther, Martin,$d1483−1546.
$aRaud, Eno,$d1928−1996.
$aKross, Jaan,$d1920−2007.
$aMänd, Heljo,$d1926−
$aTuglas, Friedebert,$d1886−1971.
$aMaanso, Viivi,$d1929−
$aSotter, Ingrid.
399
316
273
268
261
260
260
235
218
216
215
192
179
174
174
0
100
200
300
400

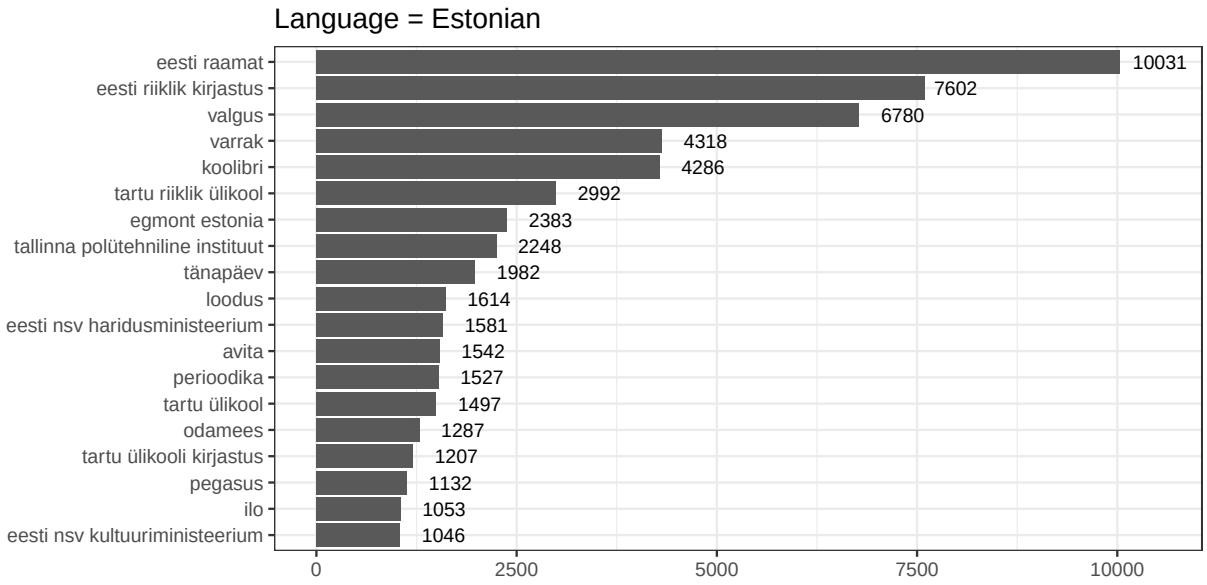

Most common authors in the Estonian language set
$aTammsaare, A. H.,$cpseud.,$d1878−1940.
$aVilde, Eduard,$d1865−1933.
$aEisen, Matthias Johann,$d1857−1934.
$aCartland, Barbara,$d1901−2000.
$aLuts, Oskar,$d1887−1953.
$aLenin, Vladimir Iljit..,$cpseud.,$d1870−1924.
$aKreutzwald, Friedrich Reinhold,$d1803−1882.
$aMänd, Heljo,$d1926−
$aMaanso, Viivi,$d1929−
$aLuther, Martin,$d1483−1546.
$aKarik, Hergi,$d1929−2012.
$aKäis, Johannes,$d1885−1950.
$aBrandt, Friedrich,$d1830−1890.
$aTalvik, Mati,$d1969−
$aStalin, Jossif,$cpseud.,$d1878−1953.
$aTungal, Leelo,$d1947−
$aRoberts, Nora,$cpseud.,$d1950−
$aKitzberg, August,$d1855−1927.
$aTolstoi, Lev,$d1828−1910.
$aLints, Alfred,$d1909−1994.
$aTuglas, Friedebert,$d1886−1971.
$aLeppik, Elmar E.,$d1898−1978.
$aChristie, Agatha,$d1890−1976.
$aLindgren, Astrid,$d1907−2002.
305
290
271
261
238
235
219
177
174
172
166
166
166
165
157
155
149
148
147
146
143
130
124
123
0
100
200
300



Publishers

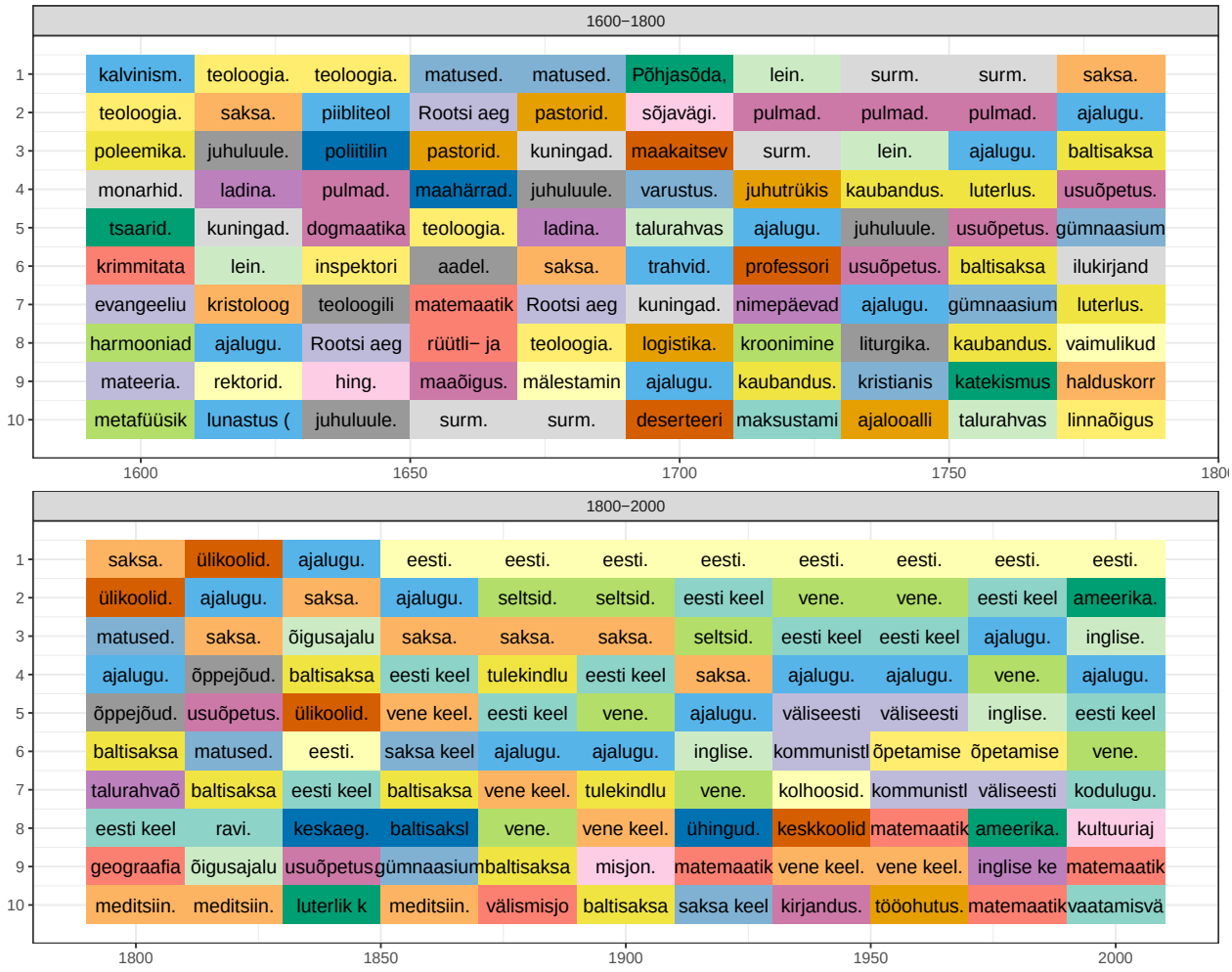
Publisher names have also been harmonized manually and algorithmically with more frequent names in focus.





Topics

The bibliography has each publication marked by some general topics. Some works have no topic marking, many works have several topic markings. By 20-year intervals, here are the most common topics in the set. For visualization purposes, only the first 10 characters of the topic are shown.

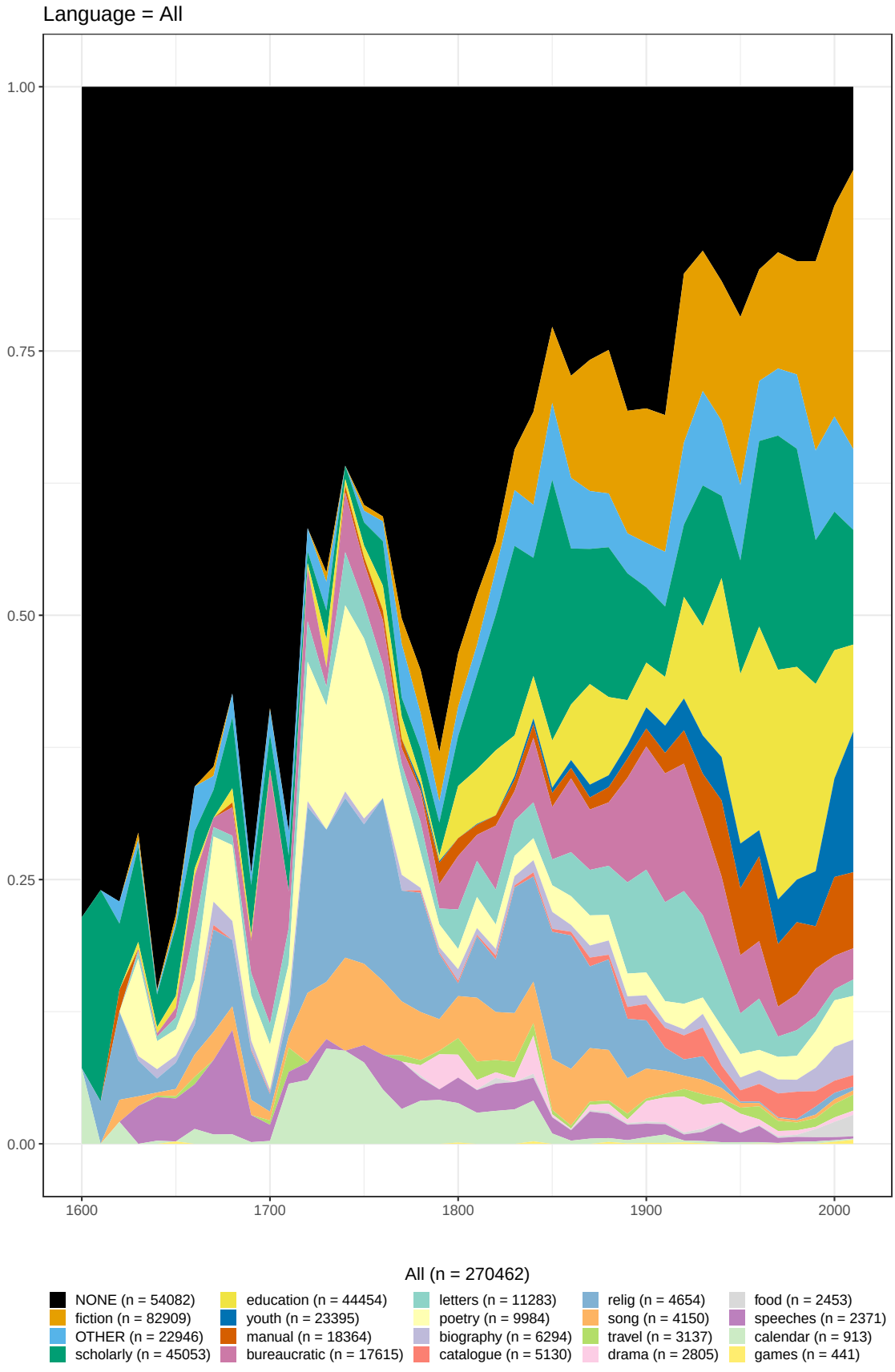


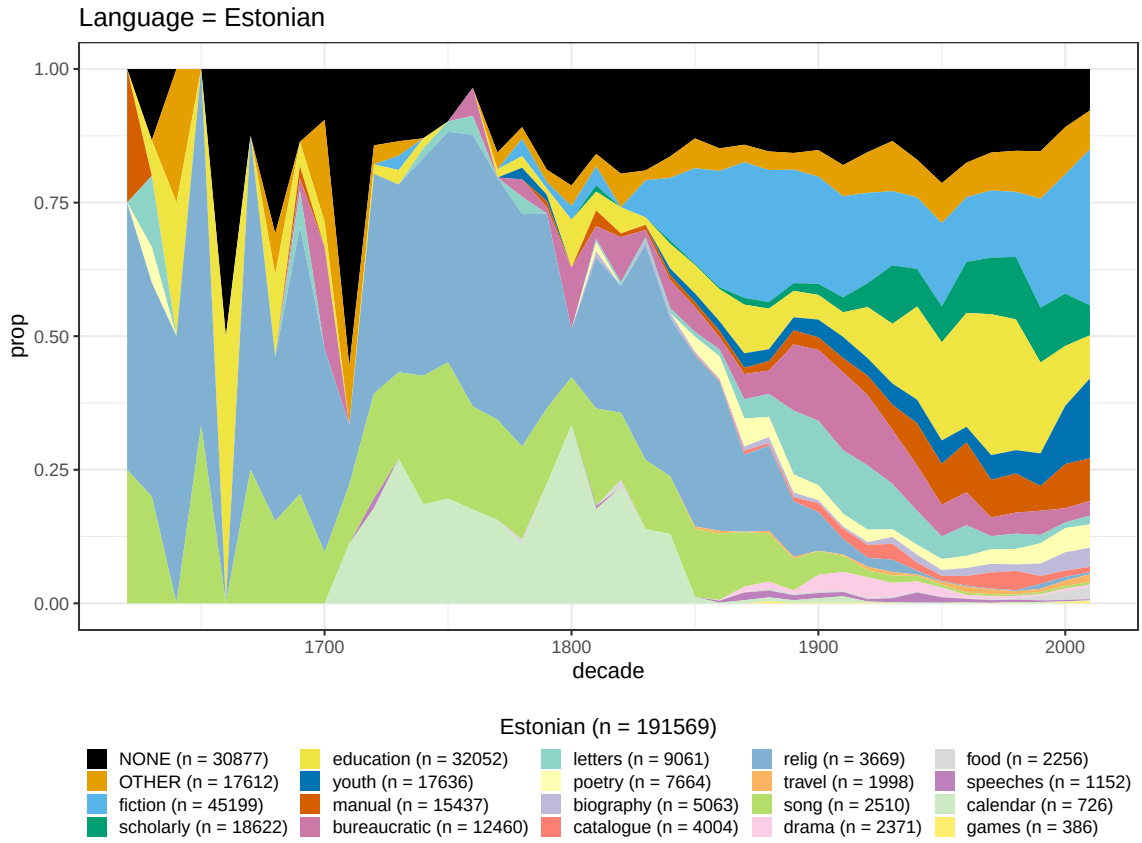
Genres

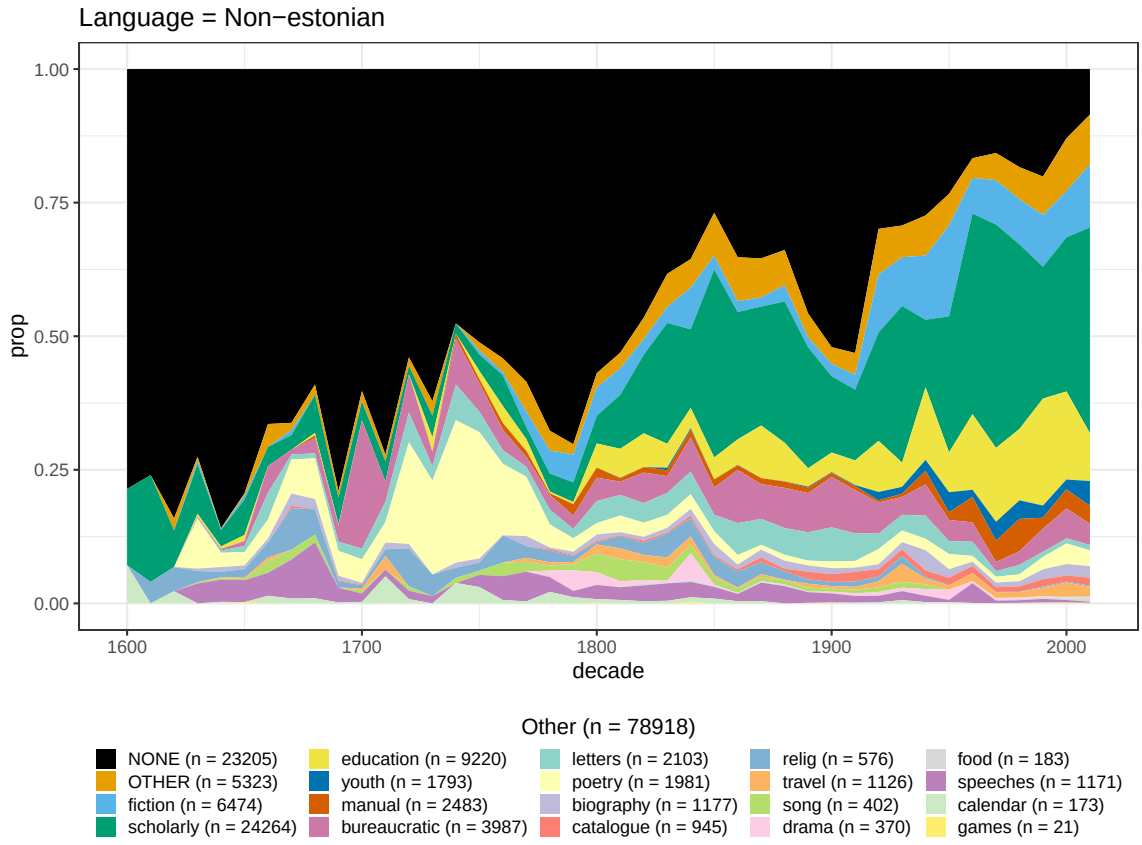
The bibliography also includes genre markers on many works. Some works have no markers, many have several. Here, we have built a few larger categories based on the dataset and visualized their frequencies over time. NONE means that the works did not have any genre markers.

Genres have been grouped into 18 major categories. Many remain uncategorized. A total of 559 unique keywords are currently classified as OTHER. Top genres there are given in the table below. 54647 works in the dataset had no genre marking.

genre_orig	N
mälestused.	2621
e-raamatud.	1258
võrguväljaanded.	948
juubeliväljaanded.	936
fotoalbumid.	854
kavad.	826
statistilised andmed.	813
infotrükised.	749
sõnaraamatud.	643
oskussõnastikud.	465

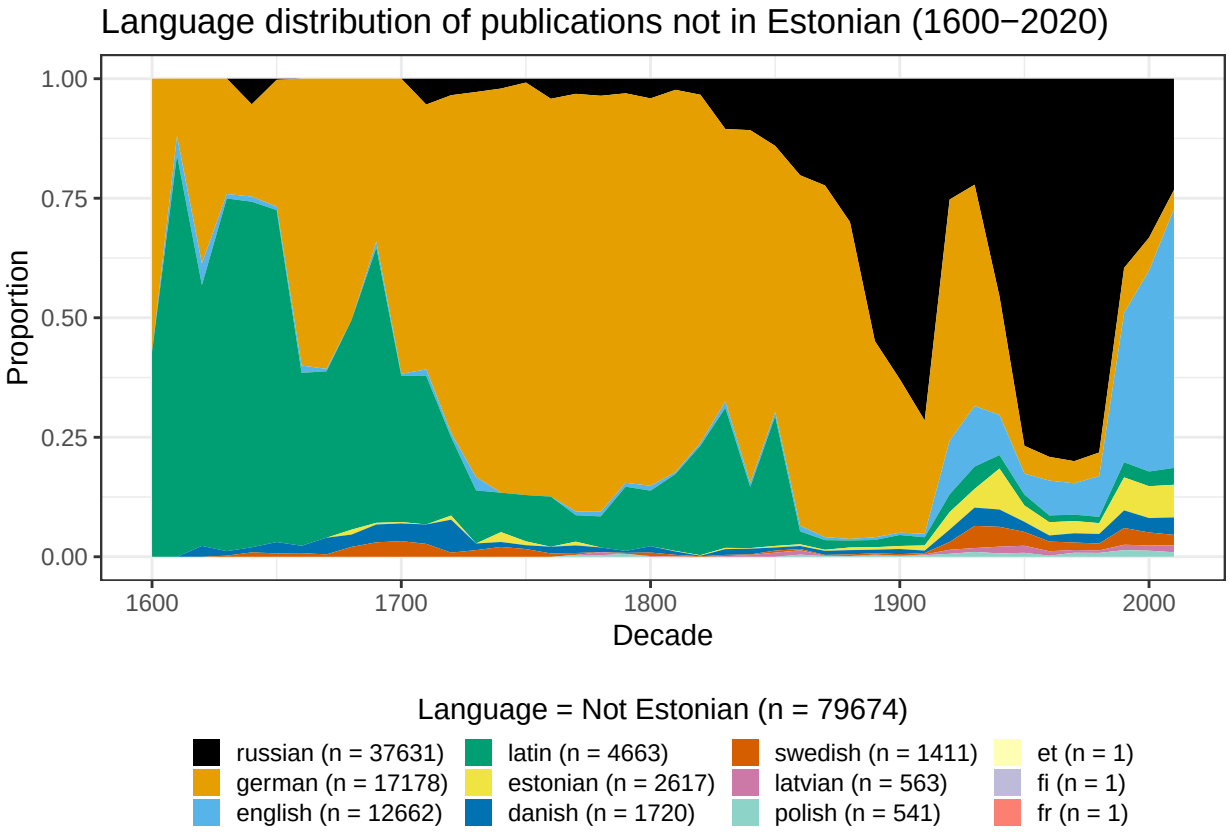

Language = All
1.00
0.75
0.50
0.25
0.00
1600
1700
1800
1900
2000
All (n = 270462)
NONE (n = 54082)
fiction (n = 82909)
OTHER (n = 22946)
scholarly (n = 45053)
education (n = 44454)
youth (n = 23395)
manual (n = 18364)
bureaucratic (n = 17615)
letters (n = 11283)
poetry (n = 9984)
biography (n = 6294)
catalogue (n = 5130)
relig (n = 4654)
song (n = 4150)
travel (n = 3137)
drama (n = 2805)
food (n = 2453)
speeches (n = 2371)
calendar (n = 913)
games (n = 441)

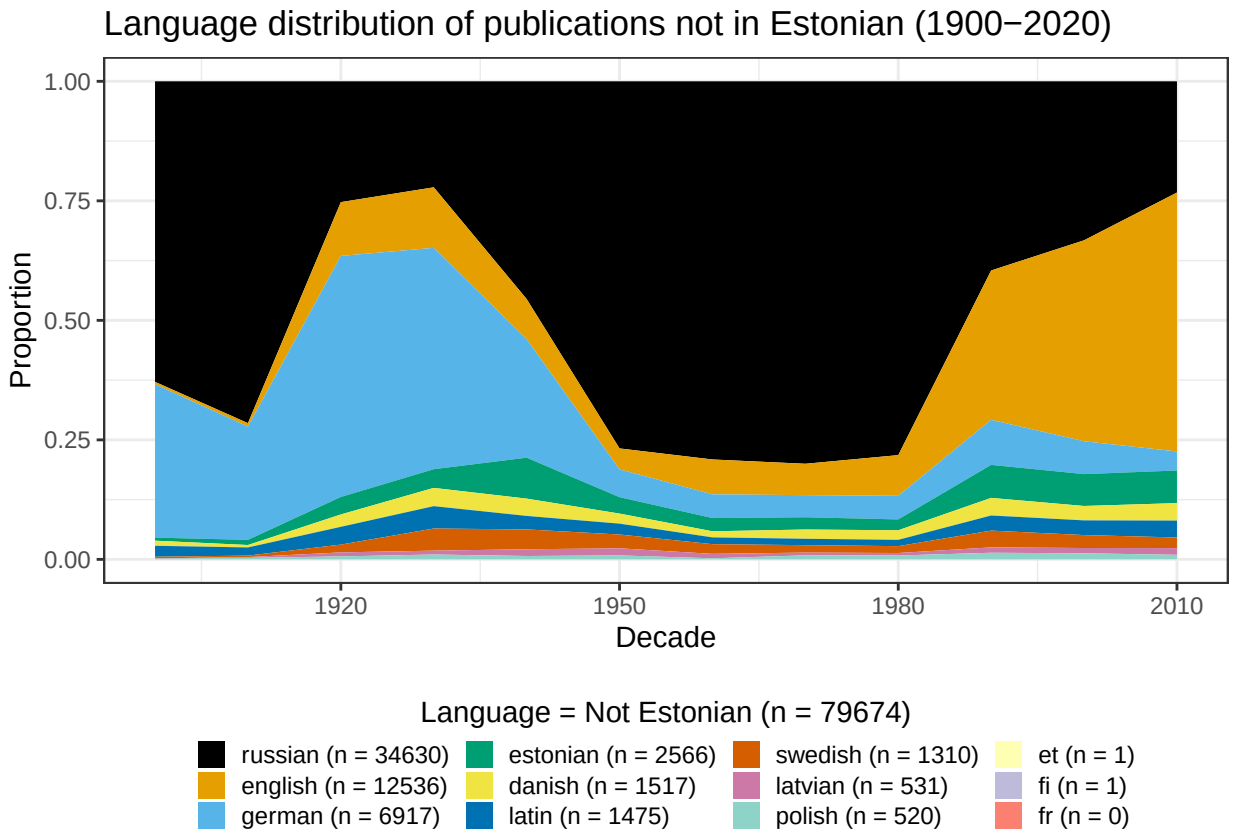

Language = Estonian
1.00
0.75
0.50
0.25
0.00
prop
1700
1800
1900
2000
decade
Estonian (n = 191569)
NONE (n = 30877)
OTHER (n = 17612)
fiction (n = 45199)
scholarly (n = 18622)
education (n = 32052)
youth (n = 17636)
manual (n = 15437)
bureaucratic (n = 12460)
letters (n = 9061)
poetry (n = 7664)
biography (n = 5063)
catalogue (n = 4004)
relig (n = 3669)
travel (n = 1998)
song (n = 2510)
drama (n = 2371)
food (n = 2256)
speeches (n = 1152)
calendar (n = 726)
games (n = 386)


Language = Non−estonian
1.00
0.75
0.50
0.25
0.00
prop
1600
1700
1800
1900
2000
decade
Other (n = 78918)
NONE (n = 23205)
OTHER (n = 5323)
fiction (n = 6474)
scholarly (n = 24264)
education (n = 9220)
youth (n = 1793)
manual (n = 2483)
bureaucratic (n = 3987)
letters (n = 2103)
poetry (n = 1981)
biography (n = 1177)
catalogue (n = 945)
relig (n = 576)
travel (n = 1126)
song (n = 402)
drama (n = 370)
food (n = 183)
speeches (n = 1171)
calendar (n = 173)
games (n = 21)

Languages

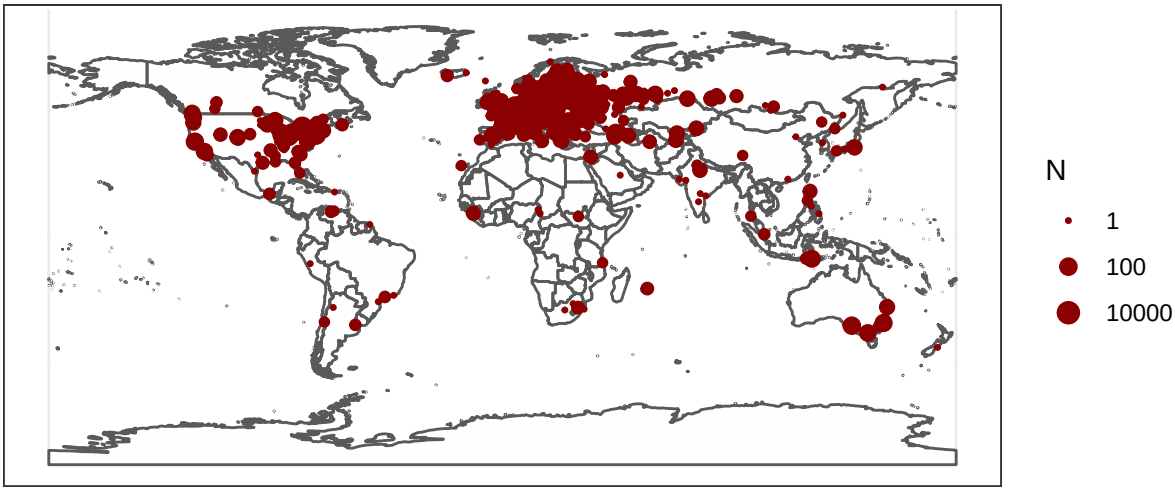
For the dataset that is not in Estonian, language has been marked in combination of existing metainformation and automatic language extraction based on the title of the work. An overview of the languages is given below. Each work is just given one language here and multilinguality is not included in this measure.

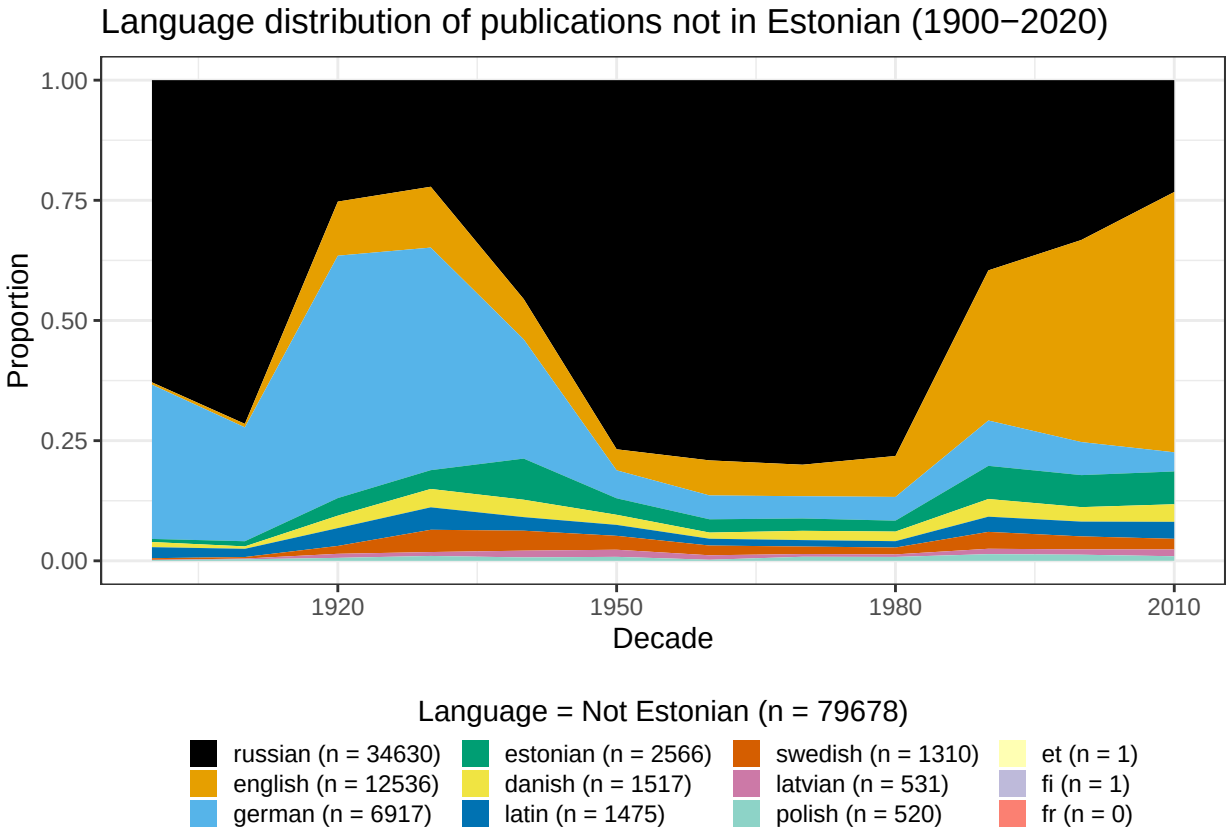


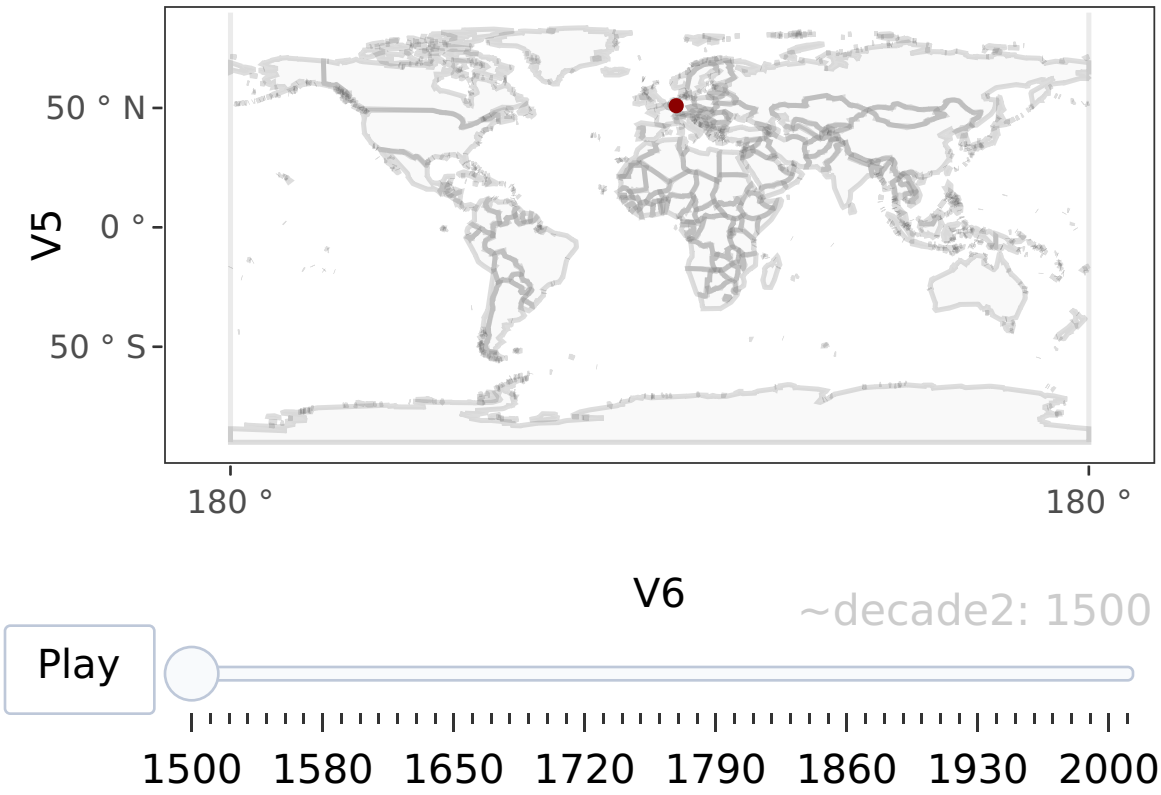


Publication locations

Publication locations have also been joined with geographic information in the GeoNames database. This allows the publication locations also to be displayed on a map. To see the interactive version of the map, use the html overview.


N
1
100
10000


Language distribution of publications not in Estonian (1900−2020)
Proportion
1.00
0.75
0.50
0.25
0.00
1920
1950
1980
2010
Decade
Language = Not Estonian (n = 79678)
russian (n = 34630)
english (n = 12536)
german (n = 6917)
estonian (n = 2566)
danish (n = 1517)
latin (n = 1475)
swedish (n = 1310)
latvian (n = 531)
polish (n = 520)
et (n = 1)
fi (n = 1)
fr (n = 0)


50 ° N
0 °
50 ° S
V5
180 °
180 °
V6
~decade2: 1500
Play
1500
1580
1650
1720
1790
1860
1930
2000