



Newspaper collections and data mining access in Estonia

Peeter Tinitis (University of Tartu, National Library of Estonia)

MEDAL Summer School 2023

Texts and digital humanities

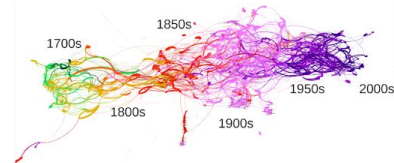
So much text

Every 60 sec ~500k tweets, fb-comments, 16M messages, 156M e-mails, 103M spam e-mails

A total of ~135M printed works ever published, 20-30% digitized



Close reading: what the text says, how and why

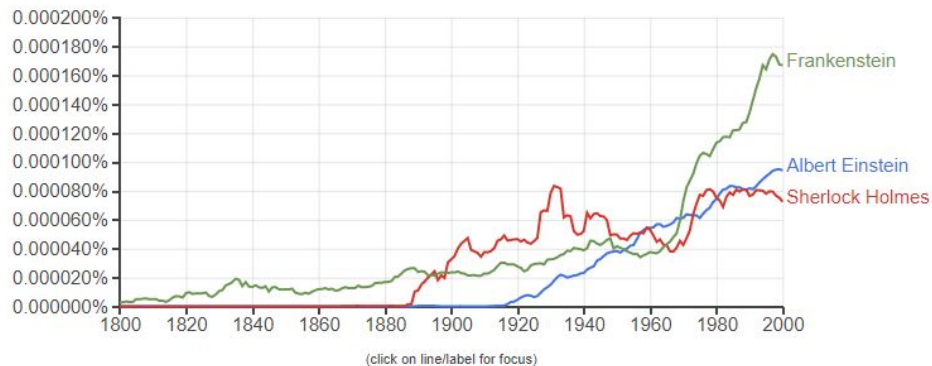


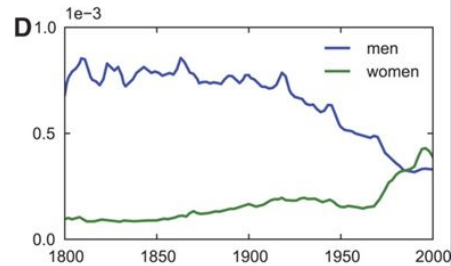
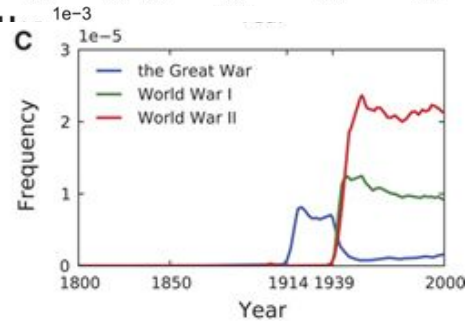
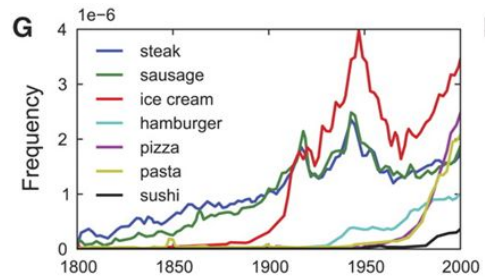
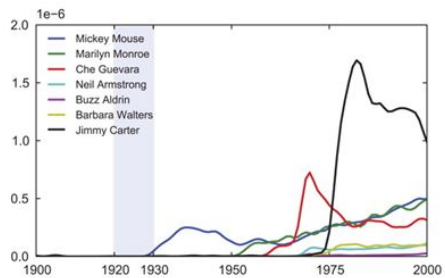
Distant reading - formal characteristics of many texts, their comparison, evolution etc

Google Books Ngram Viewer

Graph these comma-separated phrases: ☐ case-insensitive

between and from the corpus with smoothing of [Search lots of books](#)





Large digital collections



The collage illustrates the digitization process. It includes a person at a desk with a scanner, a hand operating a scanner, a stack of old documents, and a computer screen displaying a scanned document with OCR text overlays.

Varastati jalgratas

8. skp. ööl on Wõhmas warastatud Johan Saewerle kuuluvast rauttoöstusest „Edu“ jalgratas. Waras on tulnud akna kaudu sisse ja tahkunud uksest, seda seestpoolt awades. Warastatud jalgratas oli toodud töökotta paranduseks ja kuulus Johannes Tikkale Jmawere wallast. Omaniku kahju 30 krooni. Heinamaa läks põlema

Ilus pidupäev Karkise

Siinse linnepäeva tujunes nialalaste ja selle ümbruskonna rahwaste ühiseliste pidupäewaks, mis korraldati Karkise kihelkonna kooli poolt loomiseaegse ühiseliskooli lastepäew, mis oli ühikult ja kooli lastepäew 50 aastat juba ümber. Üks päetepäew ümber oli rahwast lohesi meelitanud muusikate. Siisupäew algas laste rongikäiguga aw. õigepidi kirita paratigi läbi aiaid wõmale loomiseaegse kooli — koolimäe, siia awas kihelkonna wõmaleid õpetajad, kes praegu on full penionist, hr. Koots wõmalekate kasega. Selle järele fanti õpilaste poolt ette mitmed tegawõimad ja wõimlemise harjutusi, mis enam-wõhem õnnestatud. Siis jättis õigepidi kooli. Kõik oli alawig, ühtetõrje ühingu uues ma- ja testitõrjenduse wõmalekate poolt. Karkise ette W. Karkise „Karkise“.

Pangutase revolvril lastepool

Kihelkondliku lastepool Karkise loomiseaegse kooli poolt

ABSRDITY OF PRIDE

1. EVERY man, let his state and condition in life be what they may, depends on those around him for assistance and support.
2. Men in a very low estate, may do us a great deal of good, and we often want their help. Many animals save us much labour and trouble, and supply us with many comforts.

Digital Humanities at the NLE

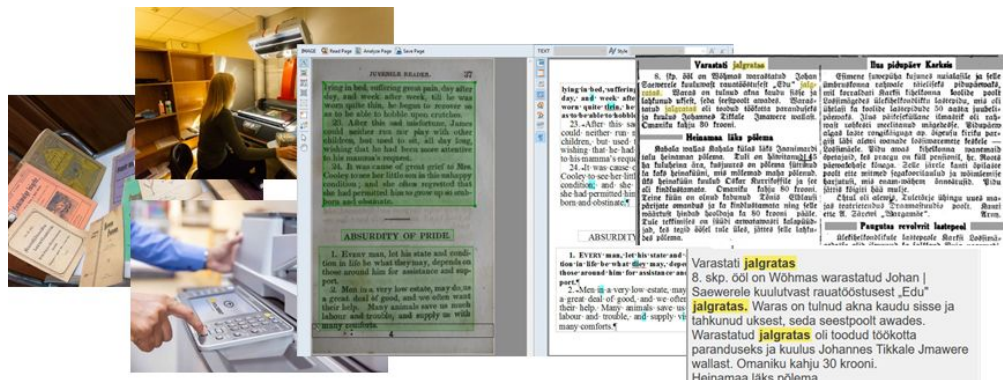
Large digital collections

- Storage and preservation
- Text search and access

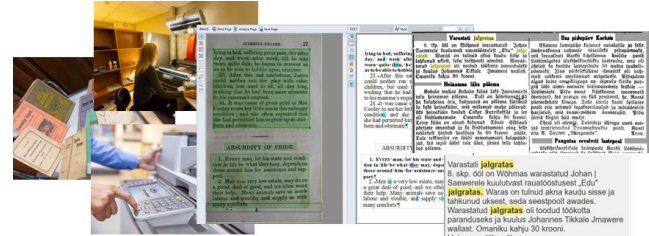


New types of users

- Tech-savvy humanities
- Data geek with interest
- Teachers, journalists etc



New user needs



Texts



Metadata



Tools



GLAM Labs

GLAM Labs community

(galleries, libraries, archives, museums)

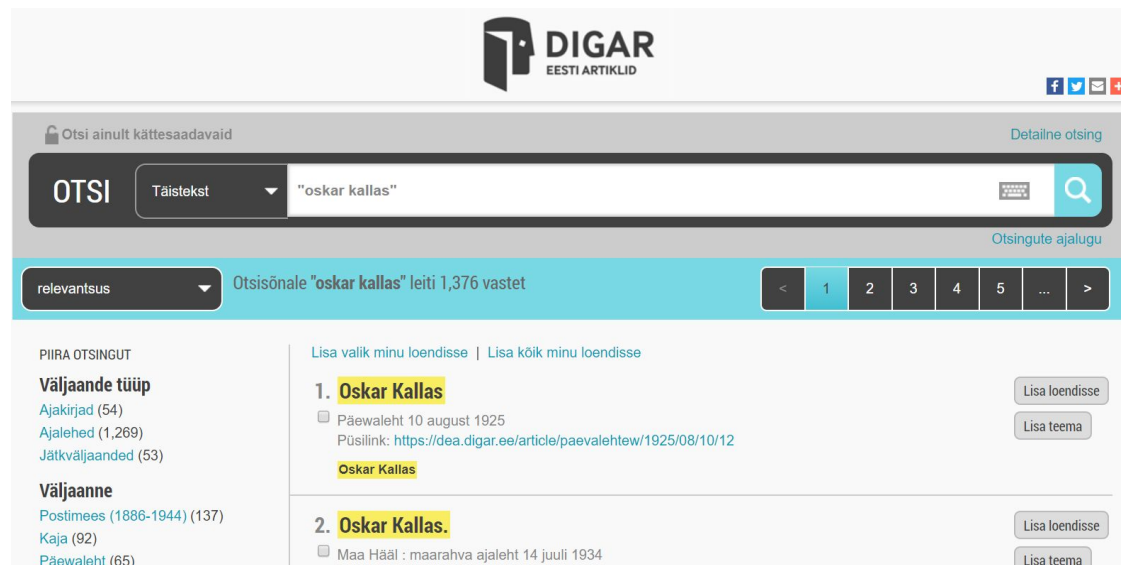
Forerunner in experimental projects in using data.

Computational access to digital collections



Findable 🔍
Accessible 🖱️
Interoperable ⚙️
Reusable ♻️

Search engine



The screenshot displays the DIGAR search engine interface. At the top, the DIGAR logo (EESTI ARTIKLID) is visible alongside social media icons for Facebook, Twitter, Email, and a general share button. A navigation bar indicates that only accessible content is shown and provides a link to detailed search options. The search bar contains the query "oskar kallas" and shows a dropdown menu for "Taistekst". Below the search bar, a filter for "relevantsus" is set, and the results count is 1,376. A pagination bar shows the current page is 1 of 5. The left sidebar lists categories: "Väljaande tüüp" (Ajakirjad, Ajalehed, Jätkväljaanded) and "Väljaanne" (Postimees, Kaja, Päevaleht). The main results area lists two items, both titled "Oskar Kallas", with options to view the full text or the topic.

DIGAR
EESTI ARTIKLID

Otsi ainult kättesaadavaid [Detaalne otsing](#)

OTSI Taistekst "oskar kallas"

Otsingute ajalugu

relevantsus Otsisõnale "oskar kallas" leiti 1,376 vastet

< 1 2 3 4 5 ... >

PIIRA OTSINGUT

Väljaande tüüp
[Ajakirjad](#) (54)
[Ajalehed](#) (1,269)
[Jätkväljaanded](#) (53)

Väljaanne
[Postimees \(1886-1944\)](#) (137)
[Kaja](#) (92)
[Päevaleht](#) (65)

[Lisa valik minu loendisse](#) | [Lisa kõik minu loendisse](#)

- Oskar Kallas** [Lisa loendisse](#)
☐ Päevaleht 10 august 1925
Püsiliik: <https://dea.digar.ee/article/paevalehtew/1925/08/10/12>
Oskar Kallas [Lisa teema](#)
- Oskar Kallas.** [Lisa loendisse](#)
☐ Maa Hääl : maarahva ajaleht 14 juuli 1934 [Lisa teema](#)

<https://dea.digar.ee>

Search engine

AVALEHT	OTSING	VÄLJAANDED	ILMUMISAEG	TEEMAD	ABI	LOGI SISSE	ENG	EST	PYC
---------	--------	------------	------------	--------	-----	------------	-----	-----	-----

[Kaja](#) > 13 aprill 1935 tr. 1

[Väljaanne PDF \(37.46 MB\)](#)

Väljaanne	Artikkel
-----------	----------

Dr. Oskar Kallas
<https://dea.digar.ee/article/kaja/1935/04/13/1/69>

Tekst

NB! Tekst võib sisaldada vigu. Loe lähemalt...
 Paranda seda teksti. Logi sisse raamatukogu kasutajatunnuse, ID-kaardi või Mobiil-ID-ga

Dr. Oskar Kallas
 MI / kutusutud Helsingi Kaalvala seltsi välisaiseks (Ulkomainen) liikmeks.

Teemad (0)

otsingu tulemused

illede näitus.

Dr. Oskar Kallas

Keeditupanga peakoosolek.

Clearing Rootsiiga kinn

Rahvakultuuri nõuk

ehmehetks prof. P. Crei

<https://dea.digar.ee>

More complex queries?



The screenshot shows the DIGAR (Eesti Artiklid) search interface. At the top, the logo "DIGAR EESTI ARTIKLID" is displayed. Below the logo, there is a search bar with the text "names that cooccur with oskar kallas". To the left of the search bar, there is a dropdown menu labeled "Täistekst" and a button labeled "OTSI". To the right of the search bar, there is a magnifying glass icon. Below the search bar, there is a message: "Otsisõnale names that cooccur with oskar kallas leiti 0 vastet". In the top right corner, there are social media icons for Facebook, Twitter, Email, and a plus sign. In the bottom right corner, there is a link "Otsingute ajalugu".

Otsi ainult kättesaadavaid

Detailne otsing

OTSI Täistekst names that cooccur with oskar kallas

Otsingute ajalugu

Otsisõnale names that cooccur with oskar kallas leiti 0 vastet

Interfaces & libraries

Delpher



jupyter nbviewer

JUPYTER

FAQ

</>

≡

🔄

🔍

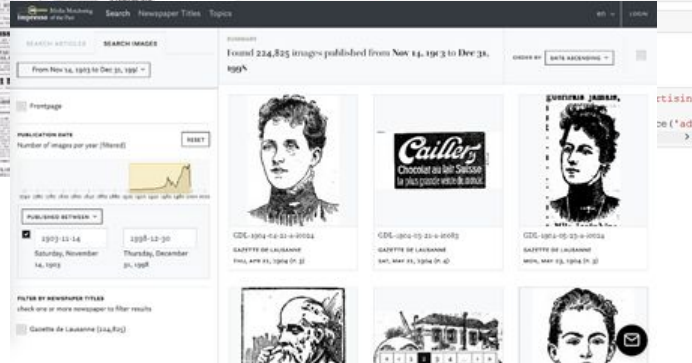
📄

Show when the articles were published

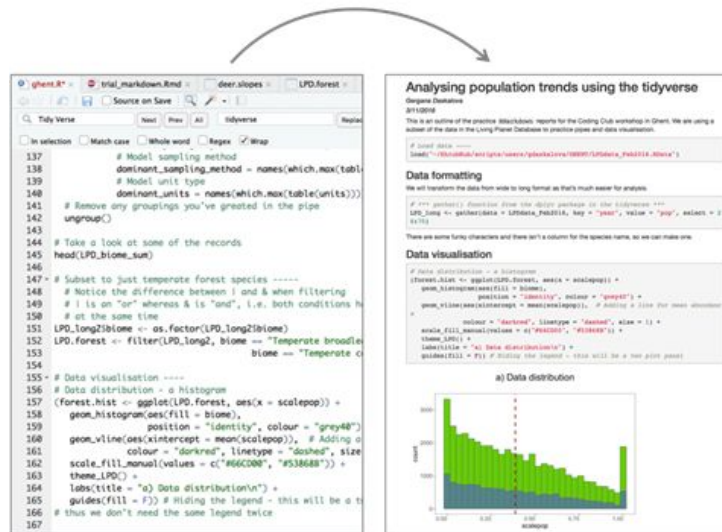
```
In [ ]: alt.Chart(df).mark_line().encode(  
    x='year(date):T',  
    y='count()',  
    tooltip=[alt.Tooltip('year(date):T', title='Year'), alt.Tooltip('cou  
    ).properties(width=600)]
```

Find the longest article

```
In [ ]: # Which is the longest article(s)?  
df[df['words'] == df['words'].max()]
```



Open Science practice



Data and code availability

Data and code to reproduce the analysis and figures are available at <https://osf.io/6ysda/>



(Heunis 2020)

GLAMs with data

Access points to data via open code (e.g. GLAMworkbench)

TROVE

ABOUT HELP NEWS PARTNERS SIGN UP LOGIN

Explore Categories Community Research First Australians

Enter your search query

Use the [Trove web interface](#) to construct your search. Remember that the harvester will get all of the matched results, not just the first 2,000 you see in the web interface. Once you're happy with your search, just copy the url and paste it below.

Query url:

Set harvest options

By default the harvester only saves the metadata (date, page, title, newspaper etc) from the search results. If you want to save the full text content of each article, just check the 'Text' box. You can also save PDF copies of every article by checking the 'PDF' option, but be warned that this will slow down your harvest and generate large download files. If you want to save PDFs from large harvests, you're probably better off installing and running the harvester on your own computer.

☒ Save full text

☐ Save PDFs (this can be slow)

Start harvest

47% 100/213 [00:03-00:03, 29.02article/s]

Once your harvest is complete a link will appear to download the results as a single, zipped file. See [this notebook](#) for more information about the contents and format of the results folder.

You can also start to explore your results [using this notebook](#).

Created by [Tim Sheratt](#) (@tshagg) as part of the [GLAM Workbench project](#).

Jupyter nbviewer

JUPYTER FAQ </> [Menu] [Refresh] [Close] [Download]

Show when the articles were published

```
In [ ]: alt.Chart(df).mark_line().encode(
        x='year(date):T',
        y='count()',
        tooltip=[alt.Tooltip('year(date):T', title='Year'), alt.Tooltip('count()', title='Count')],
        properties(width=600)
    )
```

Find the longest article

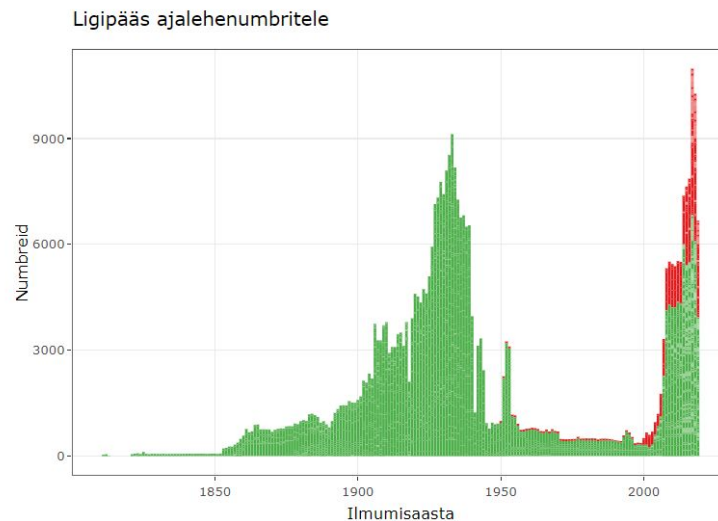
```
In [ ]: # Which is the longest article(s)?
df[df['words'] == df['words'].max()]

In [ ]: df.loc[df['title'].str.contains('cyclone', case=False, na=False)]
```

Make a simple word cloud

```
In [ ]: df_titles = df[(df['title'] != 'No Title') & (df['title'] != 'Advertisement')]
# Get all the articles titles and turn them into a single string
title_text = df_titles['title'].str.lower().str.cat(sep=' ').replace('ad', ' ')
# Create word cloud
```

Open materials at NLE





Interactive overviews

http://data.digar.ee/text/dea_info.html

http://data.digar.ee/text/dietrich_digar.html



Open code

Andmekogu

Andmekoguna kasutame Eesti Rahvusraamatukogu digiarhiivi Eesti artikleid, millele on olemas tekstikaavevõlgipääs. Kollektsoonil materjalidest saab ülevaate siit http://data.digar.ee/text/dea_info.html. Ligipääs on hetkel ainult koodi läbi

```
```{r}
Loe sisse metaandmete fail hpc serverilt.
all_issues <- fread("unzip -p /gpfs/hpc/projects/digar_txt/text/all_issues_access.zip", sep="\t")[access_now==T]

Valime AJALEHED, 1920 ja 1940 vahel, kus on väljaande koodiks postimeesew
subset <- all_issues[str_detect(DocumentType,"NEWSPAPER")%year>1920%year<1940%keyid=="postimeesew"]

Meile vajalike failide nimekirj
files <- subset[zippath_sections!="",unique(zippath_sections)]
collectionname <- "/gpfs/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname,"/text_sections/", files)

```
```

Tekstide metafailid on samamoodi indekseeritud. Järgmine koodijupp kogub kokku meie otsinguga seotud metainfo.

```
```{r}
metafiles <- subset[zippath_sections!="",unique(zippath_sections_meta)]
metafilelist <- paste0(collectionname,"/meta_sections/", metafiles)

subset_meta <- rbindlist(lapply(paste0("unzip -p ",metafilelist),fread,fill=T),idcol=T)

write_tsv(subset_meta,"subset_meta_postimeesew1.tsv")

```
```

Open data

Files at local computing cluster at the Information System of Estonian Science Agency (ETAIS)



Sign in

Username:

Password:

Sign In

```
## Andmekogu

Andmekoguena kasutame Eesti Rahvusraamatukogu digiarhiivi Eesti artikleid, millele on olemas tekstikaevelgipääs.
Kollektsiooni materjalidest saab ülevaate siit http://data.digar.ee/text/dea_info.html. Ligipääs on hetkel ainult koodi labi

'''[r]
# Loe sisse metaandmete faili hpc serverilt.
all_issues <- fread("unzip -p /gpfs/hpc/projects/digar_txt/text/all_issues_access.zip", sep="\t")[access_now==T]

# Valime AJALEHD, 1928 ja 1940 vahel, kus on väljaande koodiks postimeesew
subset <- all_issues[str_detect(DocumentType, "NEWSPAPER")$year-1928$year-1940$keyId=="postimeesew"]

# Meile vajalikke failide nimekirj
files <- subset(zippath_sections!="", unique(zippath_sections))
collectionname <- "/gpfs/hpc/projects/digar_txt/text"
filelist <- paste0(collectionname, "/text_sections/", files)

'''

Tekstide metafailid on samamoodi indekseeritud. Järgmine koodijupp kogub kokku meile otsinguga seotud metainfo.

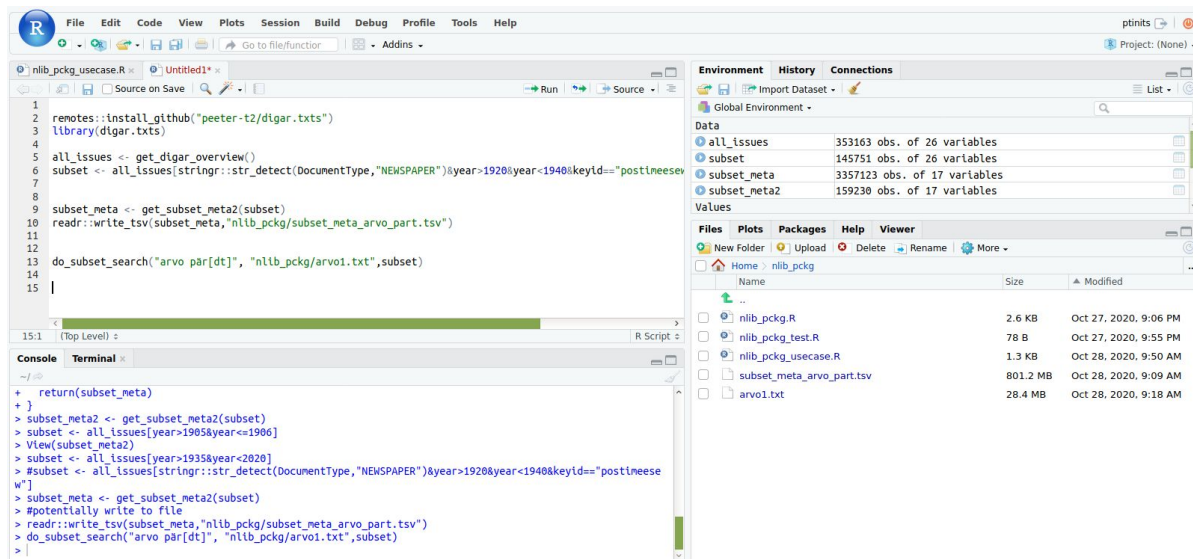
'''[r]
metafiles <- subset(zippath_sections!="", unique(zippath_sections_meta))
metafilelist <- paste0(collectionname, "/meta_sections/", metafiles)

subset_meta <- rbindlist(lapply(paste0("unzip -p ", metafilelist), fread, fill=T), idcol=T)

write_tsv(subset_meta, "subset_meta_postimeesew1.tsv")
'''
```

Access points

RStudio, Jupyter



The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains an R script for data processing. The code includes installing a GitHub package, loading the 'dgar' library, filtering 'all_issues' by document type and year, and saving the results to a TSV file.
- Environment Pane:** Lists the objects in the global environment: 'all_issues' (353163 obs. of 26 variables), 'subset' (145751 obs. of 26 variables), 'subset_meta' (3357123 obs. of 17 variables), and 'subset_meta2' (159230 obs. of 17 variables).
- Files Pane:** Shows the file structure, including 'nlib_pckg.R', 'nlib_pckg_test.R', 'nlib_pckg_usecase.R', 'subset_meta_arvo_part.tsv', and 'arvo1.txt'.
- Console:** Shows the execution output of the script, including the return value of 'subset_meta' and the results of various subset and write operations.

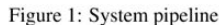
```
1 renotes::install_github("peeter-t2/dgar.txts")
2 library(dgar.txts)
3
4 all_issues <- get_dgar_overview()
5 subset <- all_issues[stringr::str_detect(DocumentType,"NEWSPAPER")&year>1920&year<1940&keyId=="post/neese
6
7
8 subset_meta <- get_subset_meta2(subset)
9 readr::write_tsv(subset_meta,"nlib_pckg/subset_meta_arvo_part.tsv")
10
11
12 do_subset_search("arvo par[dt]", "nlib_pckg/arvo1.txt",subset)
13
14
15
```

Console Output:

```
+ return(subset_meta)
+ }
> subset_meta2 <- get_subset_meta2(subset)
> subset <- all_issues[year>1905&year<=1906]
> View(subset_meta2)
> subset <- all_issues[year>1935&year<2020]
> #subset <- all_issues[stringr::str_detect(DocumentType,"NEWSPAPER")&year>1920&year<1940&keyId=="post/neese
W")
> subset_meta <- get_subset_meta2(subset)
> #potentially write to file
> readr::write_tsv(subset_meta,"nlib_pckg/subset_meta_arvo_part.tsv")
> do_subset_search("arvo par[dt]", "nlib_pckg/arvo1.txt",subset)
>
```



Let's access the texts





Semantic change

Automobile -> Car (motor, wagon, limousine)

Cable car, streetcar, tram, tramcar, trolley, trolley bus, electric coach etc.

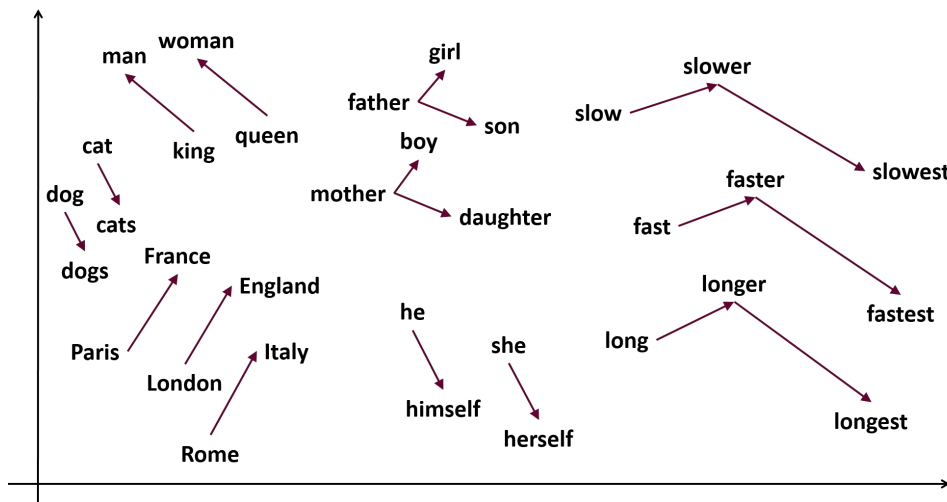
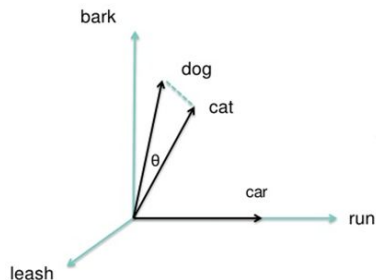
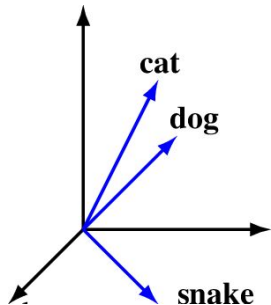
Word vectors

■ : Center Word
■ : Context Word

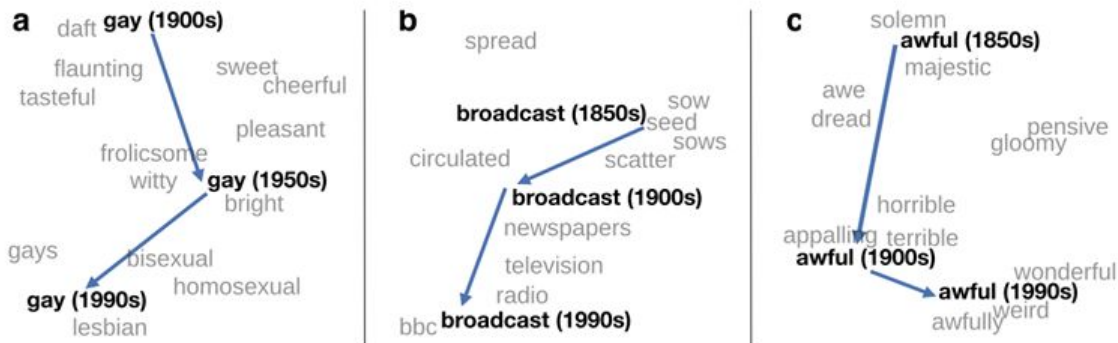
c=0 The cute **cat** jumps over the lazy dog.

c=1 The **cute** **cat** **jumps** over the lazy dog.

c=2 **The** **cute** **cat** **jumps** **over** the lazy dog.



Word vectors and meaning change



Balancing corpora and collections

Search by title

All titles are included in the search, but if there is no output in the plot and the table, the sidebar options such as year range and language should be loosened. The table below show summarised information for each newspaper.

Choose newspaper(s):

Päevaleht Postimees Õlewik

