



# Analysis tools for Danish newspapers

CLARIN-PLUS Workshop  
September 19-21

Lene Offersgaard  
Centre for Language Technology  
Faculty of Humanities

UNIVERSITY OF COPENHAGEN



# Analysis tools for Danish newspapers

## Outline

- Danish digitised newspapers
- Statsbibliotekets Lab
- CLARIN-DKs toolbox
  - Converting pdf's to text for tools
  - Linguistic annotation
- Plans and visions

# National Landscape in Denmark

DIGHUMLAB: national collaborative project: 6 partners:

2 national libraries: Royal Library and Statsbiblioteket (now merging)

4 universities: University of Copenhagen (CLARIN), University of Aarhus (DARIAH), University of Aalborg, and University of Southern Denmark

Statsbiblioteket has the responsibility of preserving the old newspapers

# Danish digitised newspapers

- Newspapers(still publishing) “digitising” their own newspapers

<http://www.kb.dk/da/nb/samling/ds/aviser/danskeaviser>  
“Politikken” digitised their newspapers, available for current subscribers

<http://www.nordjyske-avisarkiv.dk/> 10 newspapers

Also some indexes, selections of articles

- Mediastream, Statsbiblioteket

Now 24 mio. pages, targeting 32 mio. pagees

See list of digitised newspapers:

<http://www2.statsbiblioteket.dk/mediastream/avis/list>  
(sorry for Danish interface)

# Danish digitised newspapers: Mediastream

Søg i 24.139.466 avissider

Søg  Søg 

Søg kun i det, du har adgang til.

Periode fra til

[Se liste over aviser i Mediastream](#)

## Aviser i Mediastream

Fri adgang til 3.860.709 avissider  
Begrænset adgang til 20.226.515 avissider

[Læs mere om adgang](#)



[SE LISTE OVER AVISER I MEDIESTREAM >](#)



SÅDAN FÅR DU ADGANG TIL AVISSIDERNE

[Læs mere om mulighederne >](#)

ANTAL AVISSIDER I MEDIESTREAM

24.139.466

MÅLET ER 32 MIO. SIDER I ALT

# Statsbibliotekets Labs: EXPERIMENTAL & BETA



## SB LABS

### ABOUT

SB Labs seeks to find new ways to combine the library's digital cultural heritage collections and research, with the latest state of the art methods within machine learning. The lab is an initiative taken by the IT department at the State and University Library in Aarhus in 2016. SB Labs consists of the library's own IT professionals - in collaboration with researchers within the field digital humanities.

If you are interested in getting in touch with SB Labs please contact us at:

[sblabs@statsbiblioteket.dk](mailto:sblabs@statsbiblioteket.dk) or find us on twitter ([@sbtechlab](https://twitter.com/sbtechlab))

### SMURF

Smurf visualises how use of language in Danish newspapers has evolved since the 18th century.



# Statsbibliotekets Labs: Smurf (beta)

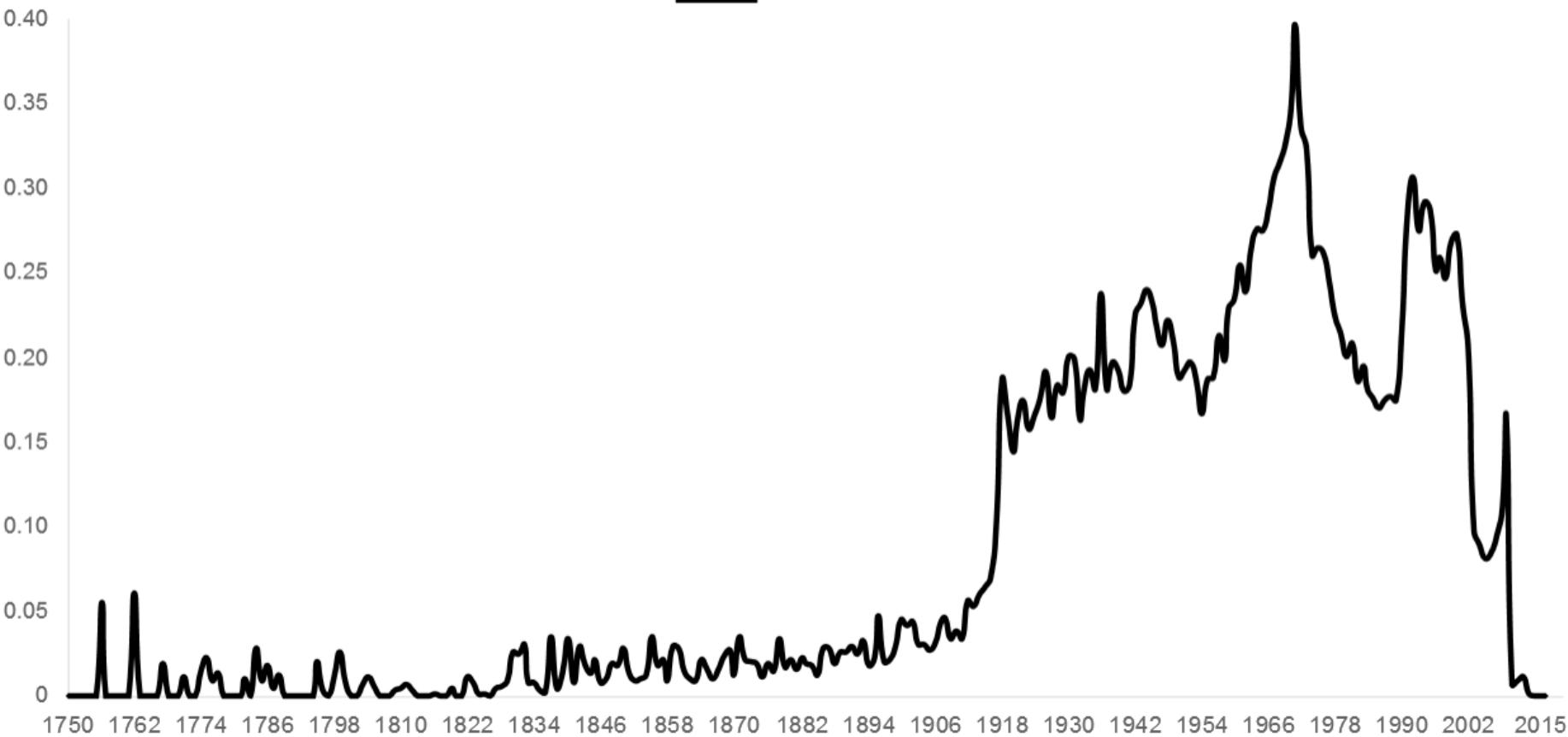
Grindsted

Search

Reset

% of total articles

Grindsted



# Statsbibliotekets Labs: Smurf (beta)

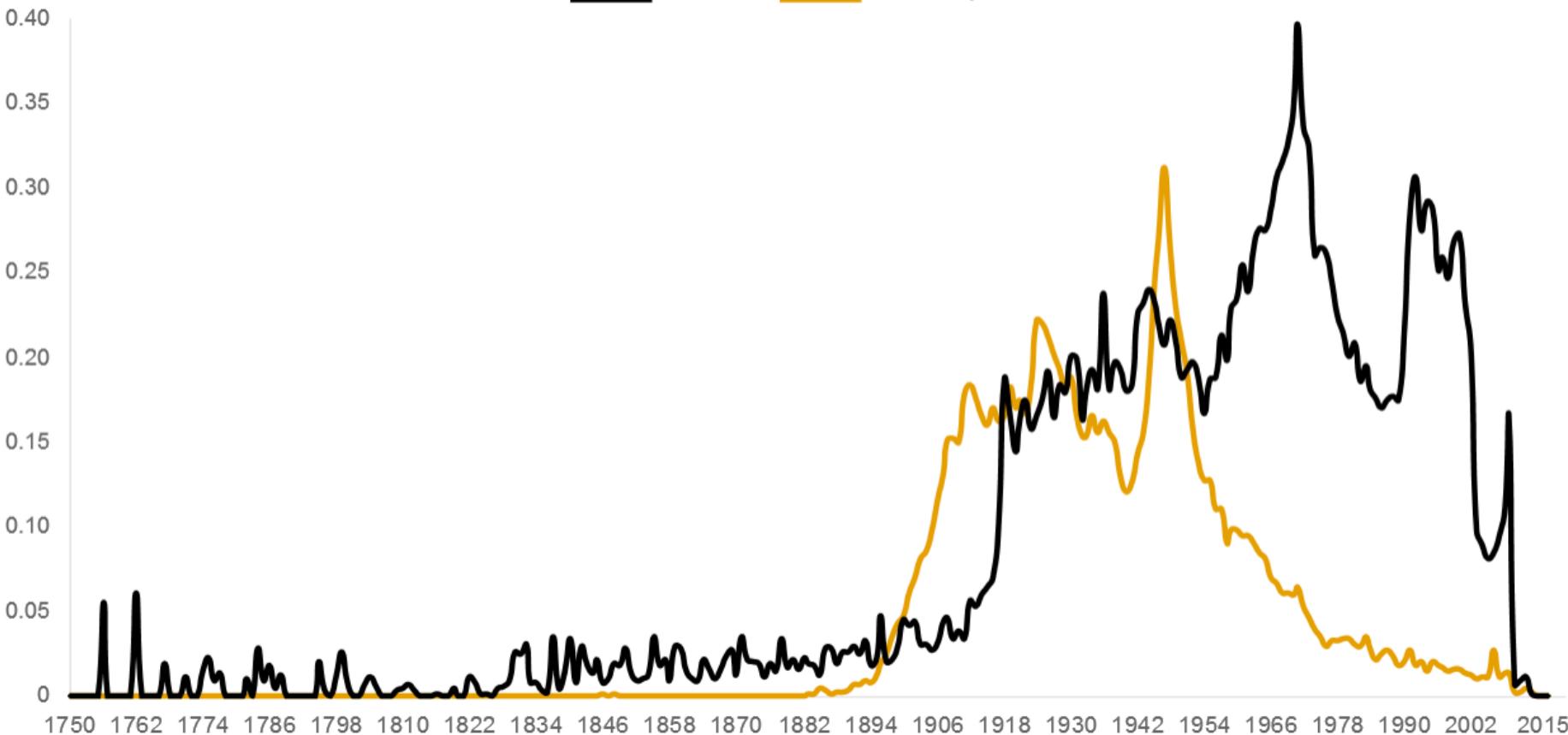
Stationsby

Search

Reset

% of total articles

Grindsted Stationsby



# Statsbibliotekets Labs: Smurf (beta)

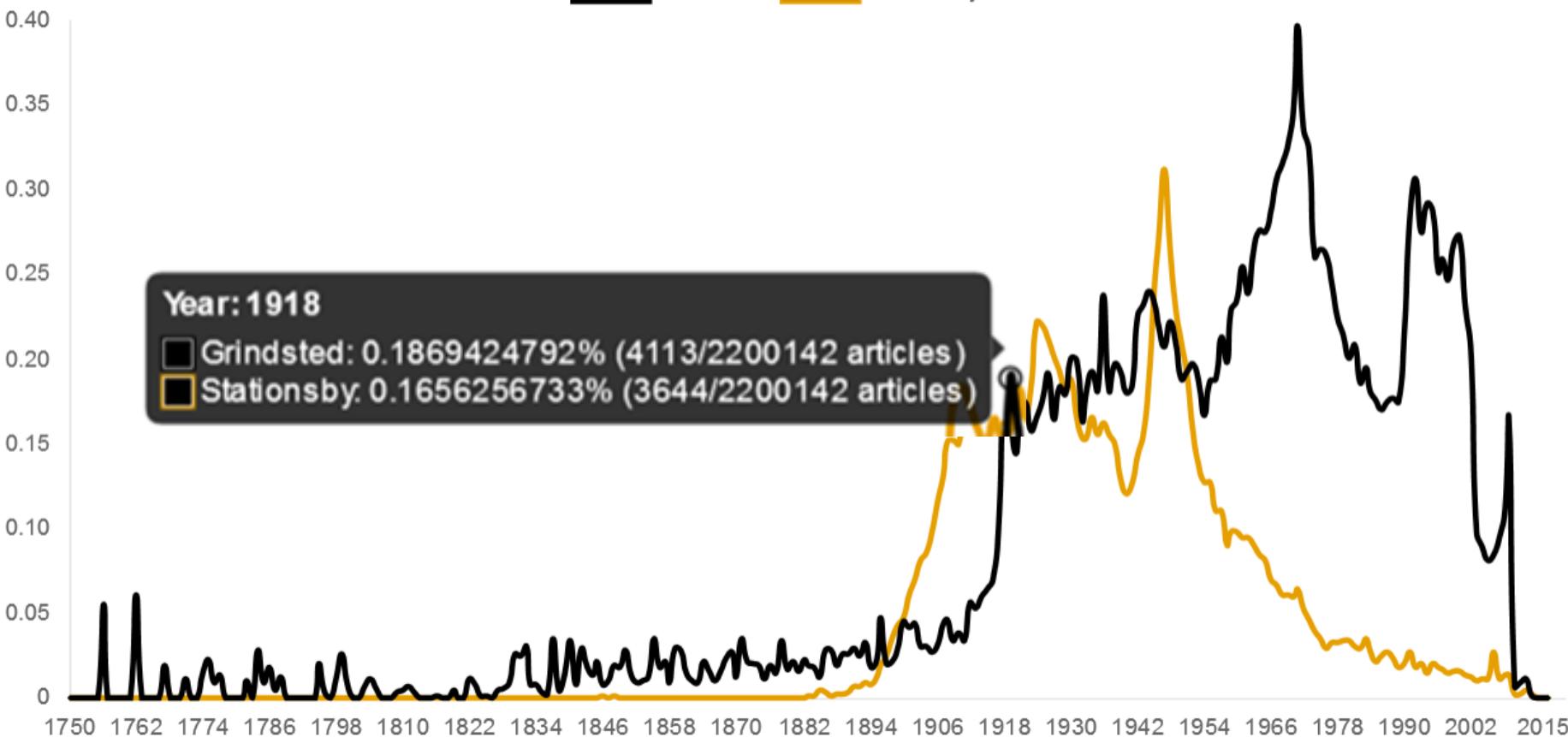
Stationsby

Search

Reset

% of total articles

Grindsted Stationsby



# Statsbiblioteket: Mediastream

Søg i 24.139.466 avissider

py:1916 AND fulltext:(Grindsted)



Søg

Søg kun i det, du har adgang til.

Periode fra

til



[Se liste over aviser i  
Mediastream](#)

Din søgning på **py:1916 AND fulltext:(Grindsted)** gav 1.486 hits

[Sortér efter relevans](#)

[Sortér efter dato](#)

## Afgræns din søgning

### Avistitel

Folkebladet Sydjylland (1889-1993)  
(433)

Jyllandsposten (1871-1937) (270)

Den til Forsendelse med de  
Kongelige Brevposter privilegerede  
Berlingske Politiske og  
Avertissementstidende  
(1833-1935) (97)

Aalborg Amtstidende (1889-1971)  
(76)

Ribe Stifts-Tidende (1849-1960)  
(74)



### JYLLANDSPOSTEN (1871-1937)

**AVIS** 4. februar 1916

I Grindsted er en Butik med Inventar samt stort | Side 4 - 1. sektion

Aalborg — Hadsund N. Fra Aalborg. . . . 7«o | Side 2 - 1. sektion



### JYLLANDSPOSTEN (1871-1937)

**AVIS** 19. februar 1916

Trælast. | Side 6 - 1. sektion

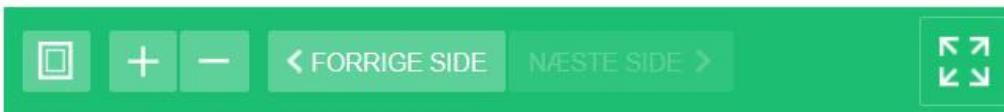


### RIBE STIFTS-TIDENDE (1849-1960)

**AVIS** 22. marts 1916

# Statsbiblioteket : Mediastream

## Jyllandsposten (1871-1937)

[◀ Forrige udgave](#)[Næste udgave ▶](#)

... siger Plads til 1. Marts, helst i eller Aalborg.  
let. mrkt. »3077«, modtager »Jyllands Kontor.

**En solid Mand,**  
erende Kjøbmand med Formue,  
Beskjæftigelse snarest.  
let. mrkt. »Kontor eller Lager  
modt. »Jyllandspostens Kontor.

dannet, livlig og musikalsk  
**ing Pige**  
godt Hjem ønskes optaget i et vel-  
t Hjem som Husholdningselev.  
verken gives eller ønskes. Pladsen  
gjørne til 15. ds. Billet, mrkt.  
modtager »Jyllandspostens Kon-  
øbet af 8 Dage.

**Plads søges.**  
ungt, velvoxent Menneske, som  
gen Skyld er ledig, siger Plads i  
retning, hvor der foruden Expedi-  
nde blive en Del Kontorarbejde.  
eret 4 Aar ved Fazet og har Han-  
olexamen. Pladsen kan tiltrædes  
Marts.  
(3411)  
ibud med Opgivelse af Len pr.  
d bedes sendt til  
Bestyrer A. Iversen,  
Lillegade 24, Grenaa.

### Butik til Leje.

En større Butik med Inventar og god  
Lejlighed paa det bedste Streg i Middel-  
fart (der har i flere Aar været Isenkram-  
og Galanterihandel) er til Leje fra April  
Flyttedag eller mulig før. Anvises af  
Hr. Sagterer Beck. Middelfart. (20175)

### Butik til Leje.

I Grindsted er en Butik med Inventar  
samt stort Bagværelse og Lagerkjælder  
til Leje strax. (2792)

Apotheket i Grindsted anviser.

### Et Værksted og Garage faaes til Leje.

Eventuelt kan medfølge en Udlej-  
ningsbil.

Billet, mrkt. »2778«, modtager »Jyl-  
landspostens Kontor.

### FORRETNINGSAPSTAELSE.

### Forretning til Afstaaelse.

For en Mand eller 2 dygtige Damer  
er en i god Gang værende **Fedevare-** og  
**Paalegsforretning** samt Udsalg af Slag-  
teriaffald og Flæsk til Afstaaelse. Til  
Butiken er anskaffet Beregningsvægt og

**AVIS** 4. februar 1916

DOWNLOAD AVIS SOM PDF

### HITS I DENNE AVIS

S.2 | Aalborg — Hadsund N. Fra Aalborg. . . . 7»o

S.4 | I Grindsted er en Butik med Inventar samt stort

### ANTAL SIDER I AVISEN

4

### INFORMATION OM AVISEN

Udgivelsessted: Aarhus

Periode dækket af denne titel: 16-09-1871 - 21-12-1937

[Læs mere](#)

# CLARIN-DKs toolbox

- Standalone tools wrapped as web services
- Workflow manager to make tools easier to choose and chains of tools easier to select for non-skilled NLP users
- NLP segmentation and annotation tools
- Corpus search: supporting special corpora
- Integration with CLARIN Language Switchbord

# Converting pdf's to TEI text with Metadata - ready to use with toolbox



The image shows the CLARIN-DK website interface. At the top, there is a navigation bar with links: Front page, Introduction, Overview, Find resources, Find in text, Deposit, Tool box, Seneste søgning, About Clarin, and a language switcher (dansk). To the right of the navigation bar are user profile icons for Lene Offersgaard and a basket icon labeled "Show basket 1". Below the navigation bar is a "Tool box" section with tabs: Prepare text resources, Create CMDI metadata, Linguistic annotation (which is currently selected), and Tool administration. The main content area is titled "Upload files and convert to TEI". It includes a file upload input field with the placeholder "Gennemse..." and "Ingen filer valgt.", a dropdown menu for "Text language" set to "Danish", and a note about receiving an email with workflow results. There is also an empty "E-mail:" input field.

**CLARIN-DK**  
Common Language Resources and Technology Infrastructure

Ressource ID  
En forsknings-  
infrastruktur  
for humanister

Front page    Introduction    Overview    Find resources    Find in text    Deposit    Tool box    Seneste søgning

About Clarin    dansk    Lene Offersgaard    Show basket 1

**Tool box**

Prepare text resources    Create CMDI metadata    Linguistic annotation    Tool administration

**Upload files and convert to TEI**

Choose one or more files to convert to TEIP5-DKCLARIN. All files must have the same file format (pdf, rtf, doc, docx, odt, txt or html).

File(s):  Ingen filer valgt.

Text language:

You will receive an email with a link to the results of the workflow when the workflow has been executed. Make sure that the email address field below is not empty.

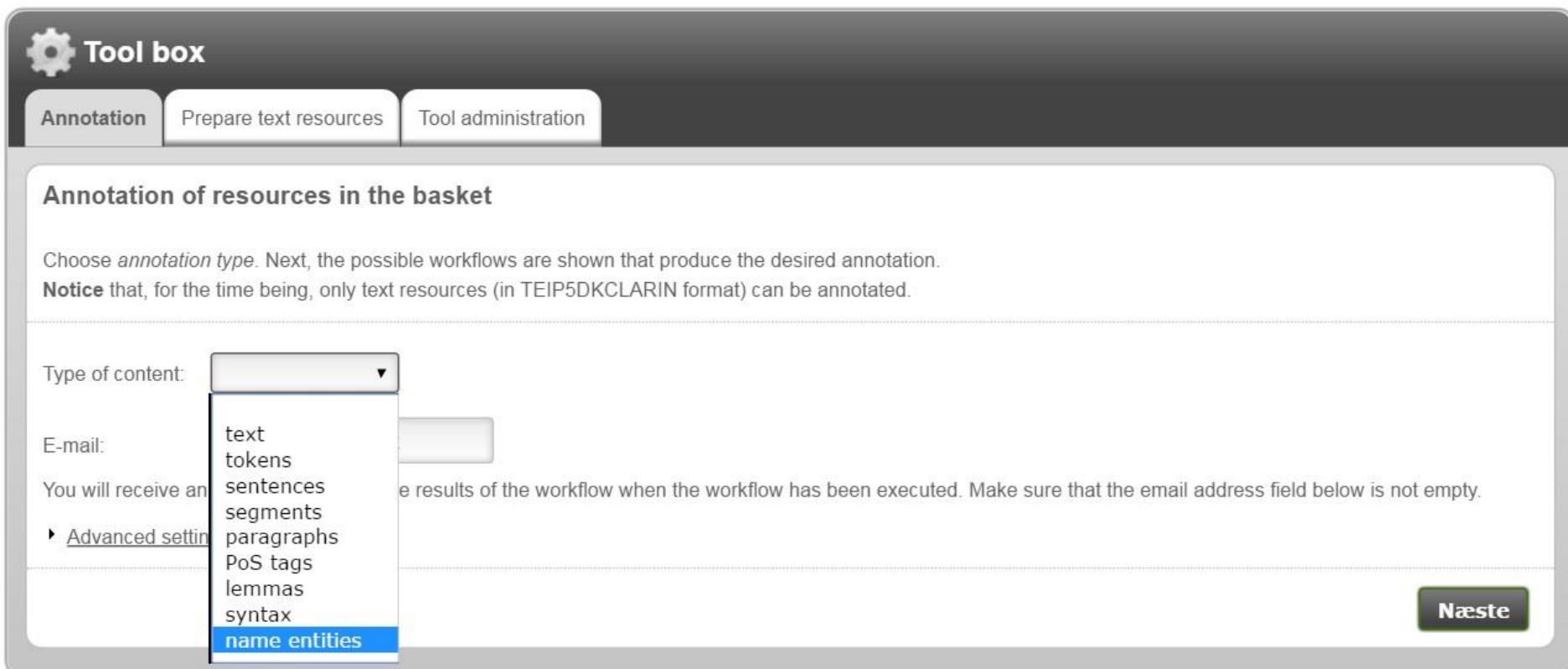
E-mail:

Næste

# Linguistic annotation and workflow manager



The image shows the CLARIN-DK website interface. At the top left is the CLARIN-DK logo with the text "Common Language Resources and Technology Infrastructure". To its right is a graphic of three blue spheres connected by lines. Further right is a map of Europe with yellow stars, one of which contains the text "En forsknings-infrastruktur for humanister". Below the logo is a navigation bar with links: "Front page", "Introduction", "Overview", "Find resources", "Find in text", "Deposit", "Tool box", "Seneste søgning", "About Clarin", "dansk", "Lene Offersgaard" (with a user icon), and "Show basket 1" (with a shopping cart icon). The "Tool box" link is highlighted.



The image shows the "Tool box" section of the CLARIN-DK interface. The title "Tool box" is at the top left, next to a gear icon. Below it is a navigation bar with tabs: "Annotation" (which is active and highlighted in blue), "Prepare text resources", and "Tool administration".

### Annotation of resources in the basket

Choose *annotation type*. Next, the possible workflows are shown that produce the desired annotation.

**Notice** that, for the time being, only text resources (in TEIP5DKCLARIN format) can be annotated.

Type of content:

E-mail:

You will receive an

**Advanced settings**

Text  
tokens  
sentences  
segments  
paragraphs  
PoS tags  
lemmas  
syntax  
**name entities**

**Næste**

# CLARIN - DK

Common Language Resources and Technology Infrastructure

[Front page](#)[Introduction](#)[Overview](#)[Find resources](#)[Find in text](#)[Deposit](#)[Tool box](#)[Seneste søgeresultater](#)[About Clarin](#)[dansk](#)[Lene Offersgård](#)

## Tool box

[Annotation](#)[Prepare text resources](#)[Tool administration](#)

## Workflow

Choose the workflow that you want to use with the resources in your basket

- 1  TEIP5-tokeniser/sentence extractor(normal) → CST-Lemmatiser(normal ; flat, alphabetic list)
- 2  TEIP5-tokeniser/sentence extractor(normal) → CST-Lemmatiser(normal ; flat, frequency list)
- 3  TEIP5-tokeniser/sentence extractor → CST-Lemmatiser(TEIP5DKCLARIN\_ANNOTATION)

# CLARIN-DK

Common Language Resources and Technology Infrastructure



Ressource ID



En forsknings-  
infrastruktur  
for humanister



Front page   Introduction   Overview   Find resources   Find in text   Deposit   Tool box   Seneste søgning

About Clarin

dansk

Lene Offersgaard

Show basket 1

## Tool box

Annotation

Prepare text resources

Tool administration

### Your job has started

The job with the chosen workflow is started. Processing your resources can take some time. You will receive an email when the job has finished. The email will contain links to the produced resources. You don't have to wait for the results. If you want you can go back to the basket and start another job.

[Tilbage til kurv](#)

CLARIN  
CENTRE B



# Plans and visions

## **Personalized annotations:**

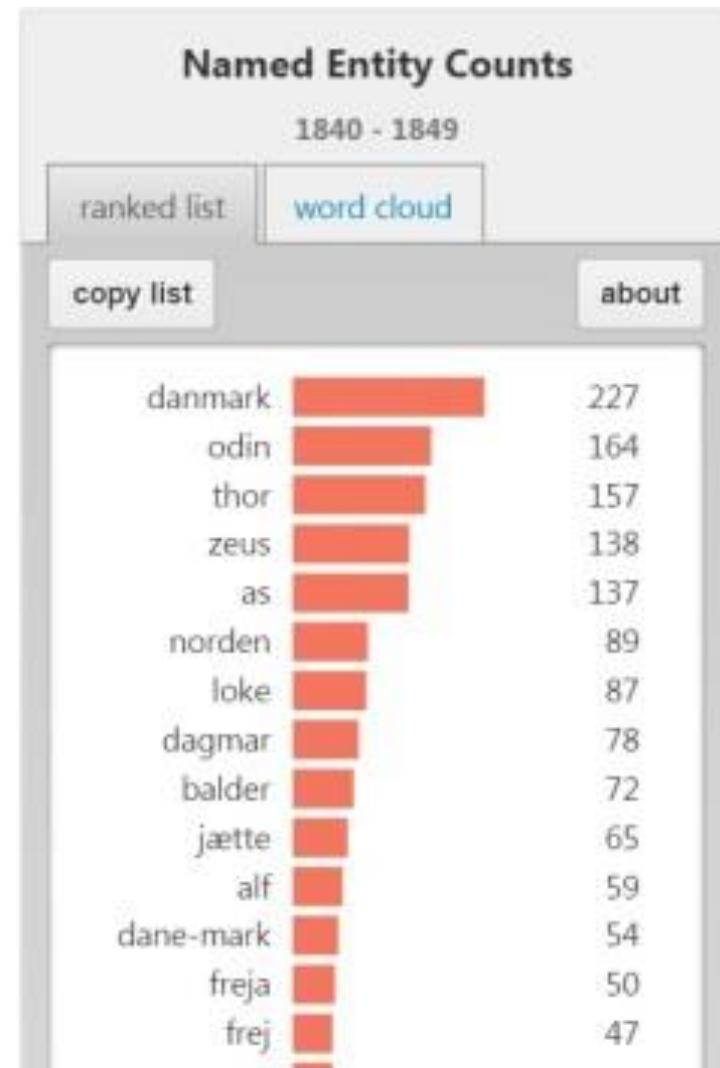
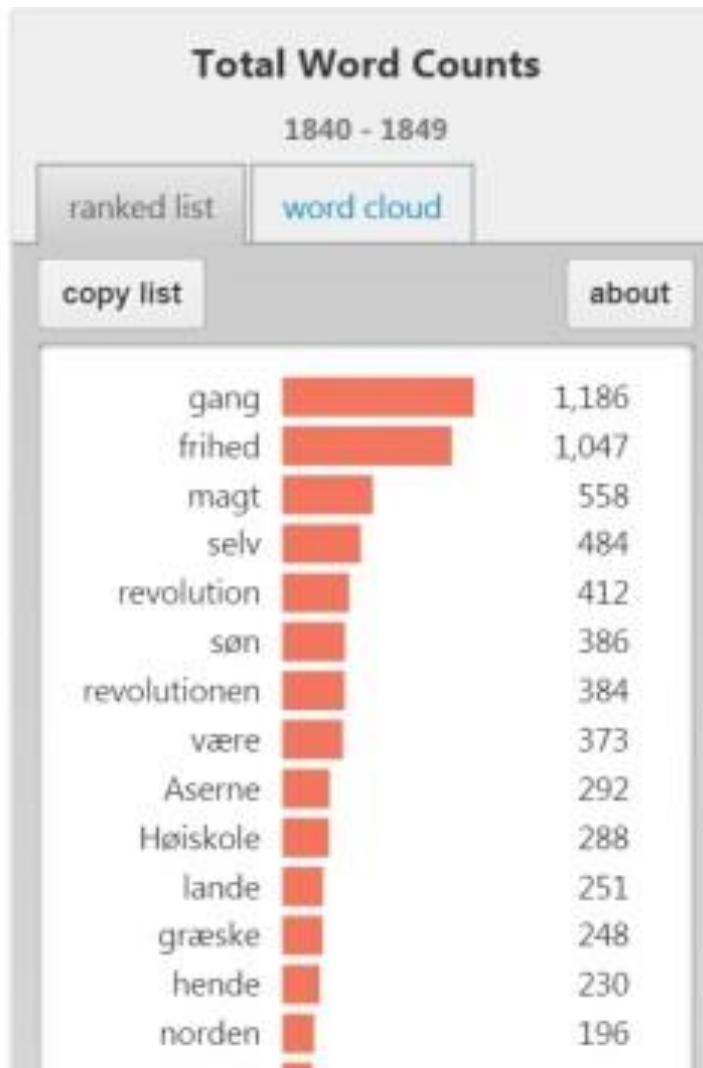
Working with formats which are easy for researchers to add information into

.... and for us to easily configure for other use cases

TEI and JSON

# Plans and visions

## Timeline selection of data:



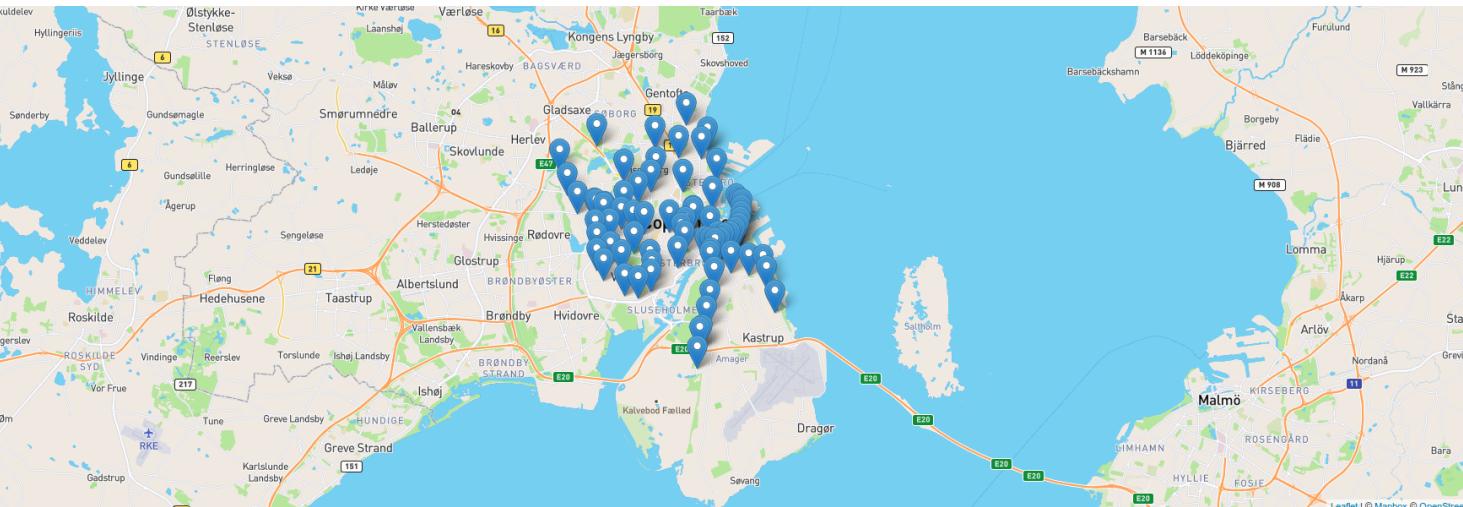
# Plans and visions

## Visualization of places:

- Working on general solution, so it can be re-used for selected subcorpus

## Collaboration:

Interested in fitting current and new tools to new material and research questions





# Thanks for listening!

**Work done together with:**

Mitchell Seaton

Bart Jongejan

Dorte Halstrup Hansen

About Smurf: contact:

[sblabs@statsbiblioteket.dk](mailto:sblabs@statsbiblioteket.dk)

UNIVERSITY OF COPENHAGEN

