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

Statsbibliotektet 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/mediestream/avis/list

  (sorry for Danish interface)

Danish digitised newspapers: Mediestream

Søg i 24.139.466 avissider

Søg Søg

Søg kun i det, du har adgang til.

Periode fra til

Se liste over aviser i Mediestream

Aviser i Mediestream

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

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 or find us on twitter (@sbtechlab)

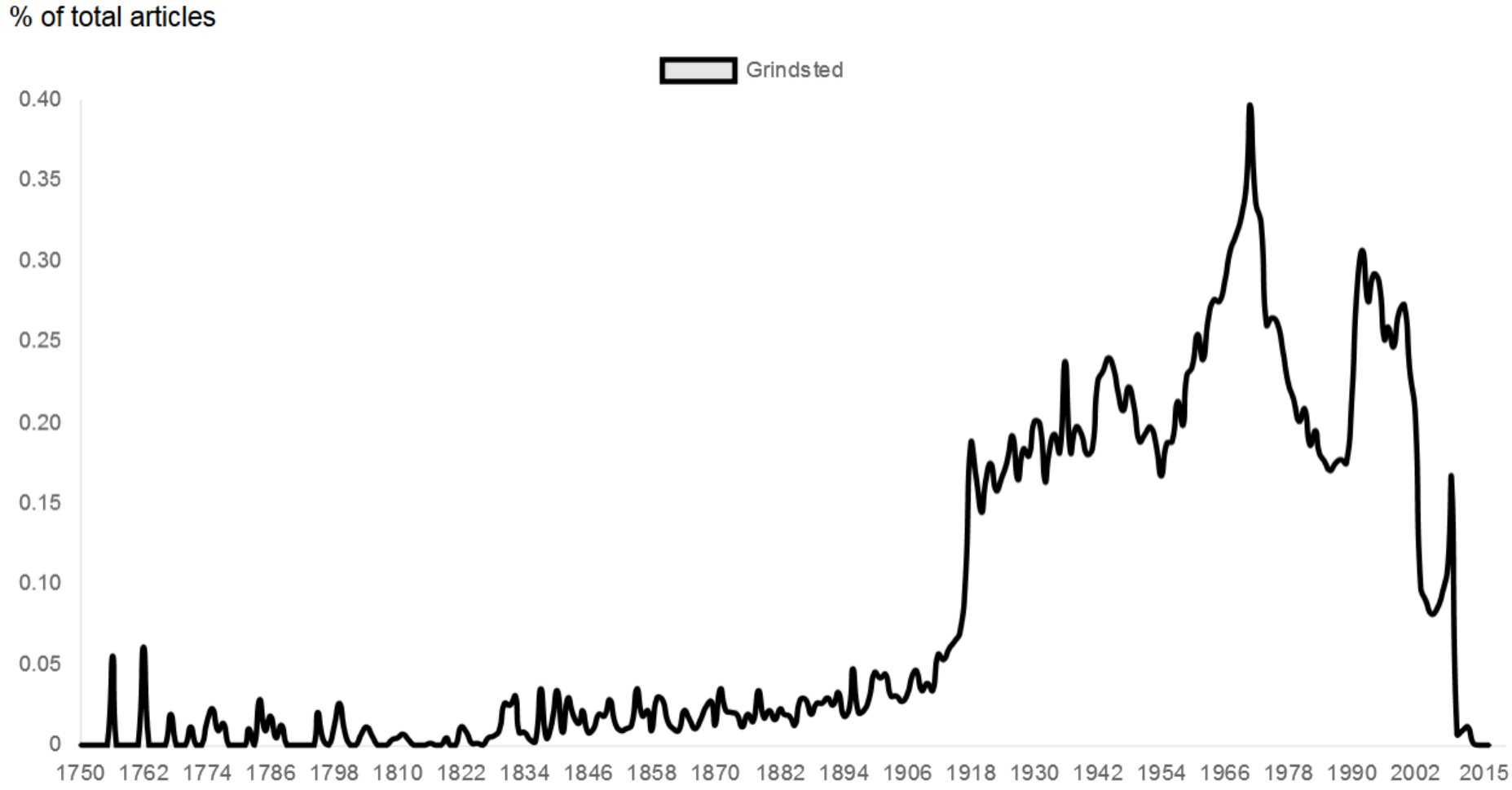
SMURF

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

Statsbibliotekets Labs: Smurf (beta)

Grindsted Search

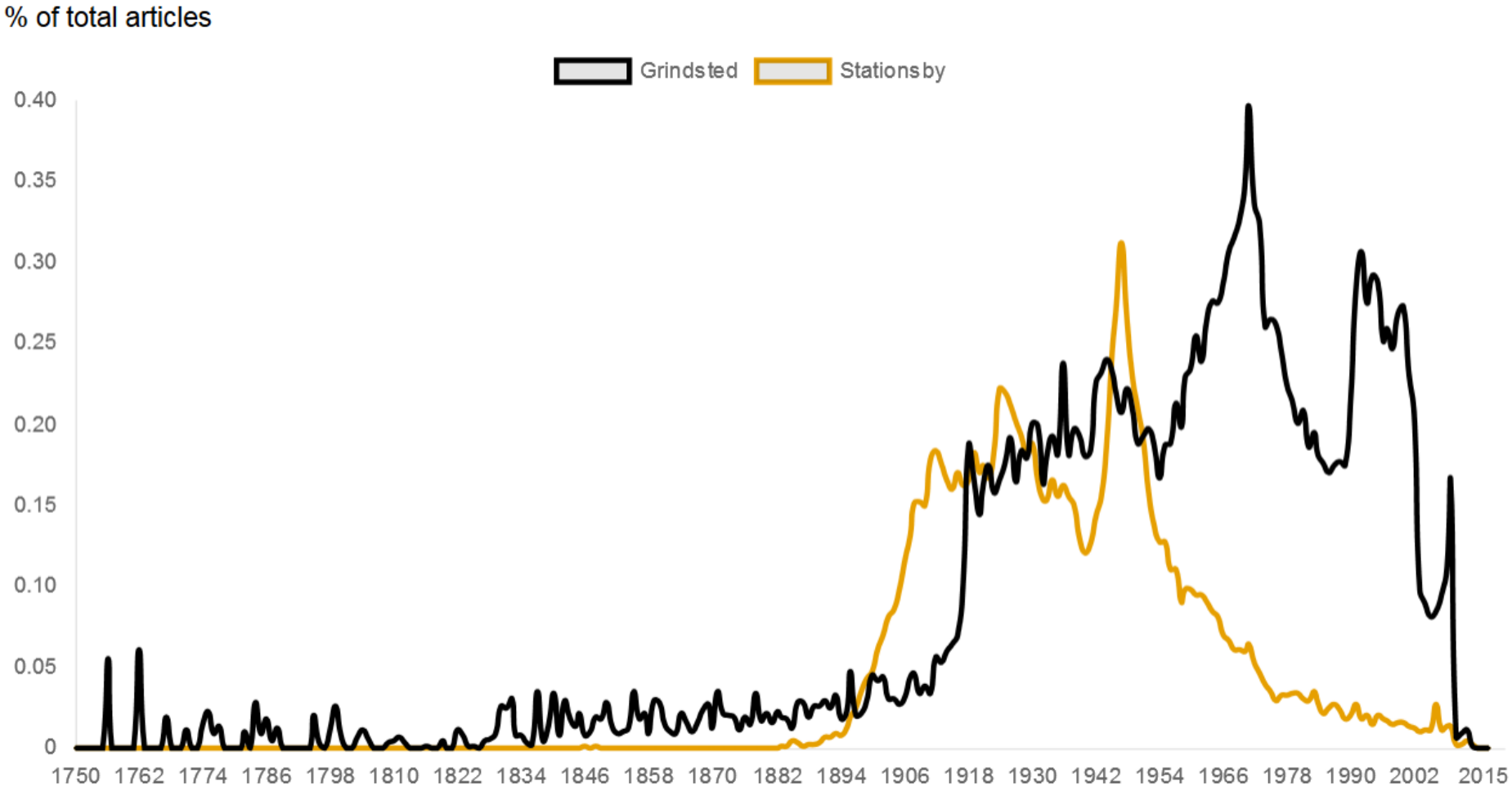
Reset



Statsbibliotekets Labs: Smurf (beta)

Stationsby Search

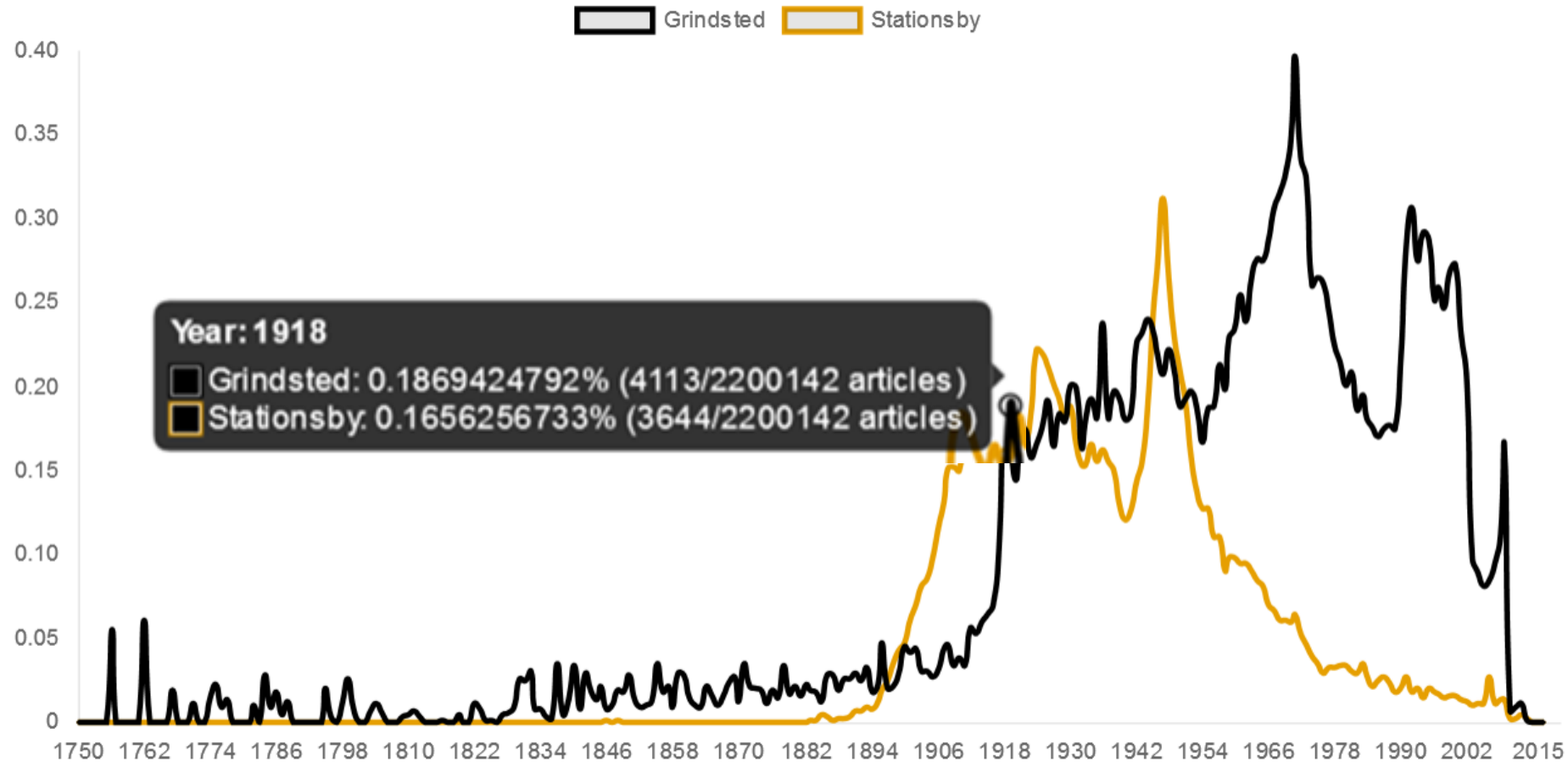
Reset



Statsbibliotekets Labs: Smurf (beta)

Stationsby Search

Reset



Statsbiblioteket: Mediestream

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 Mediestream

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 : Mediestream

Jyllandsposten (1871-1937)

Forrige udgave | Næste udgave

FORRIGE SIDE | NÆSTE SIDE

søger Plads til 1. Marts, helst i eller Aalborg.
let. mrkt. »3077«, modtager »Jyl-sten«s Kontor.

En solid Mand,

rende Kjøbmand med Formue, Beskjæftigelse snarest.
let. mrkt. »Kontor eller Lager modt. »Jyllandsposten«s Kontor.

dannet, livlig og musikalsk

ung Pige

godt Hjem ønskes optaget i et vel- t Hjem som Husholdningselev. erken gives eller ønskes. Pladsen gjærne til 15. ds. Billet, mrkt. modtager »Jyllandsposten«s Kon- øbet af 8 Dage.

Plads søges.

ungt, velvoxent Menneske, som gen Skyld er ledig, søger Plads i retning, hvor der foruden Expedi- nde blive en Del Kontorarbejde. eret 4 Aar ved Faget og har Han- leexamen. Pladsen kan tiltrædes larts. (3411)
bud med Opgivelse af Løn 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 Strøg i Middelfart (der har i flere Aar været Isenkram- og Galanterihandel) er til Leje fra April Flyttedag eller mulig før. Anvises af Hr. Sagfører **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 Udlejningsbil.
Billet. mrkt. »2778«, modtager »Jyllandsposten«s Kontor.

FORRETNINGSAFSTAAELSE.

Forretning til Afstaaelse.

For en Mand eller 2 dygtige Damer er en i god Gang værende **Fedevare- og Paalægsforretning** samt Udsalg af Slagteriaffald 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

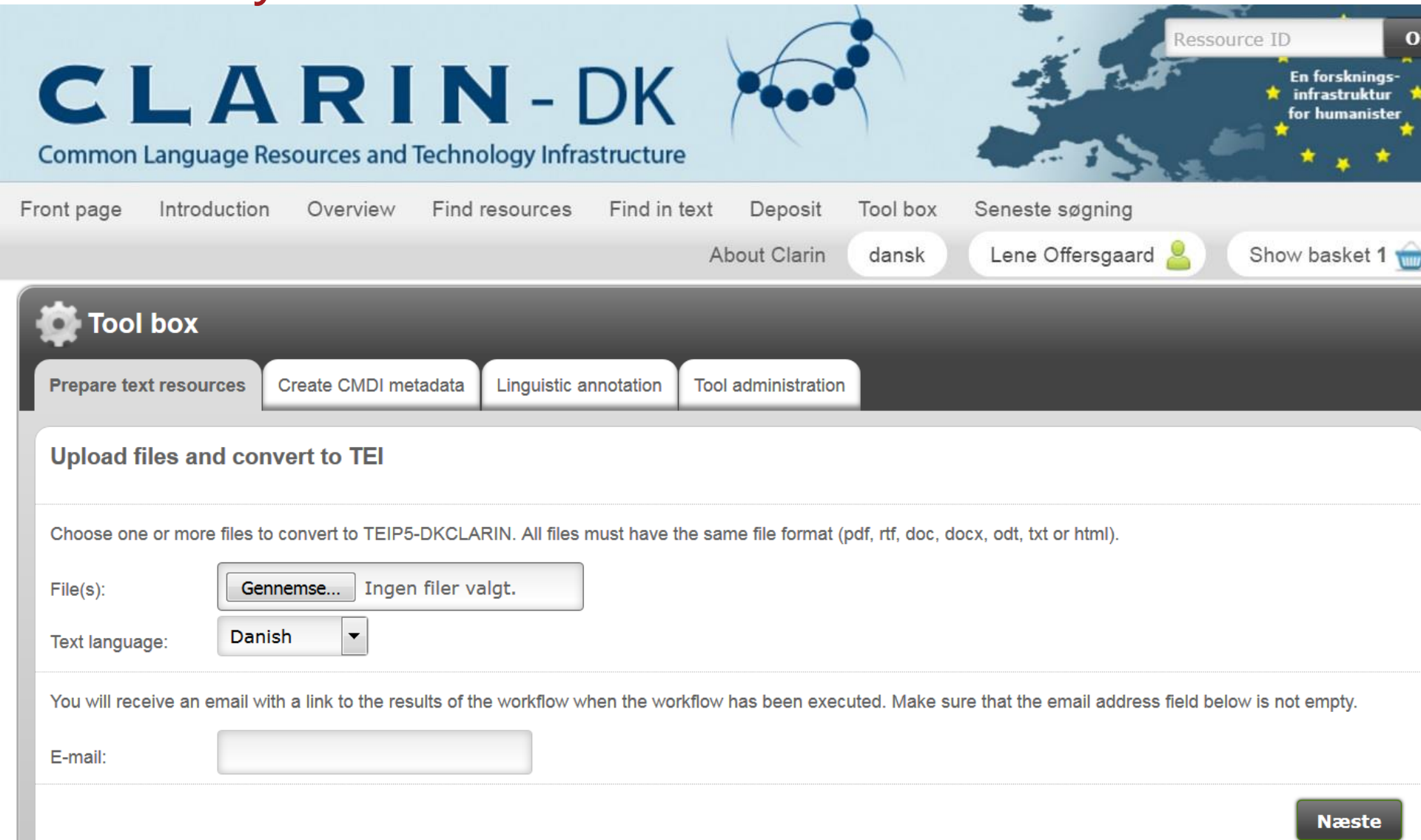
Periode dækket af denne titel: 16-09-1871 - 21-12-1937

Læs mere

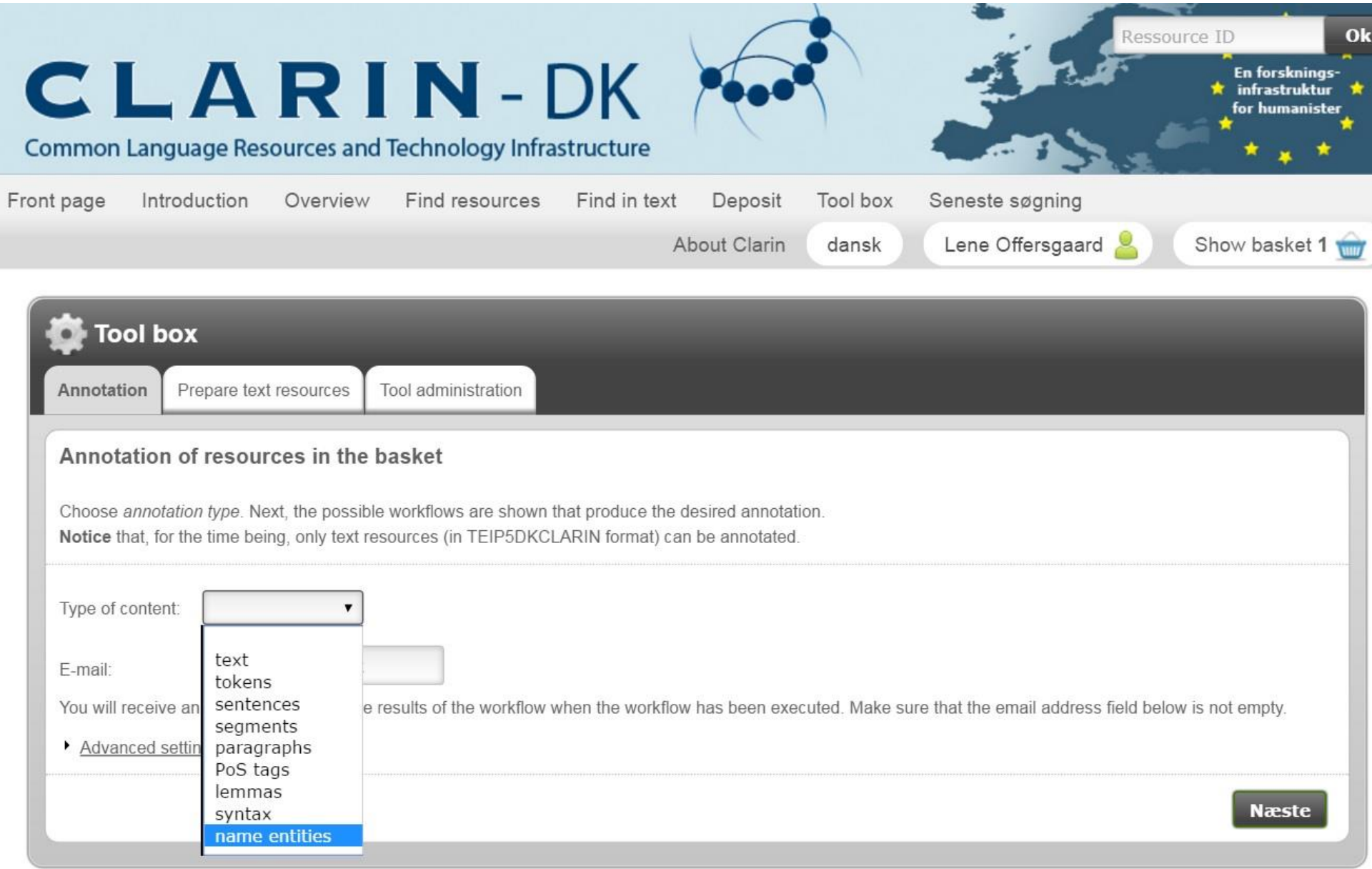
CLARIN-DKs toolbox

- 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



Linguistic annotation and workflow manager



CLARIN-DK

Common Language Resources and Technology Infrastructure

Tool box

Annotation | Prepare text resources | Tool administration

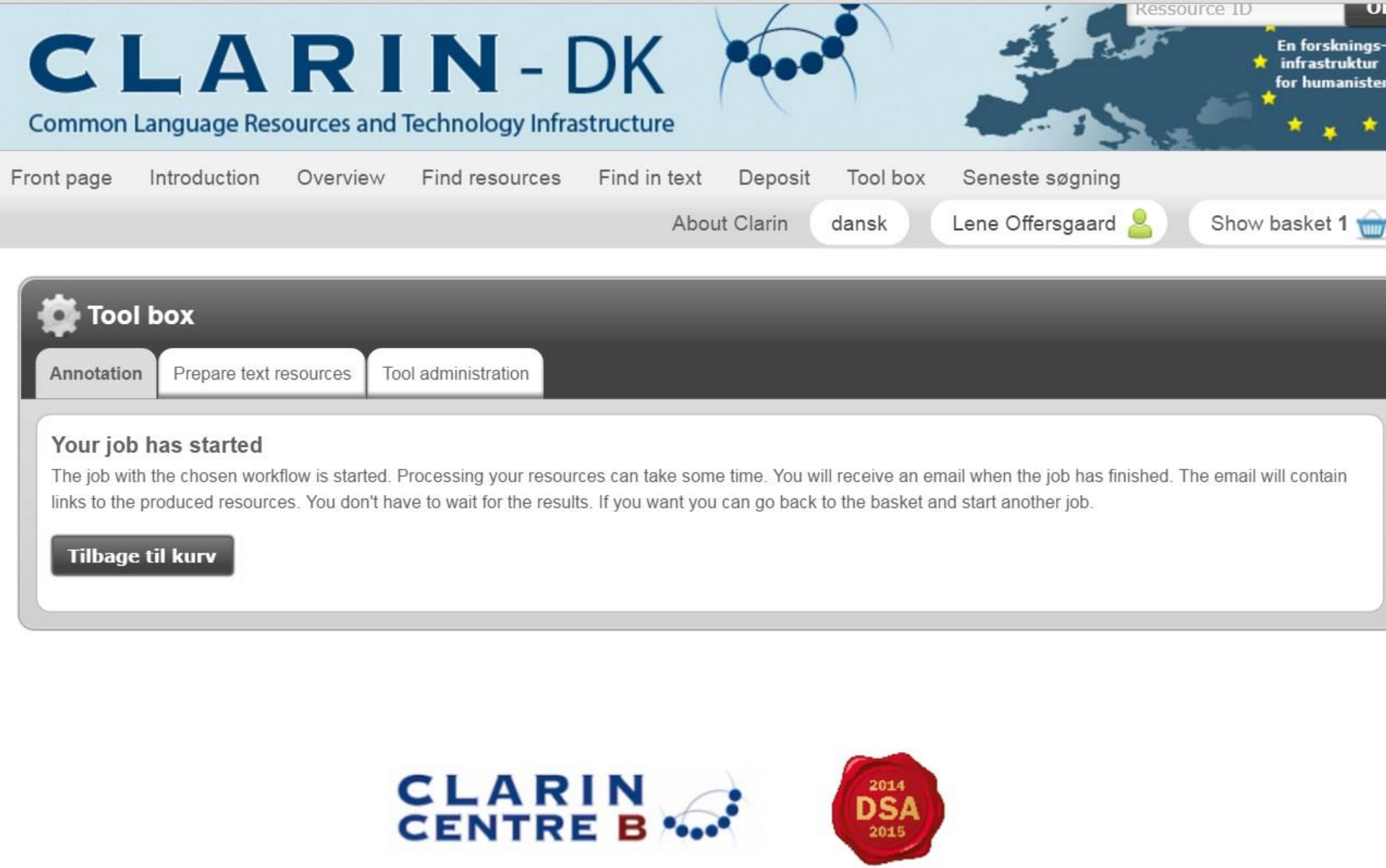
Workflow

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 CENTRE B

2014 DSA 2015

Plans and visions

Personalized annotations:

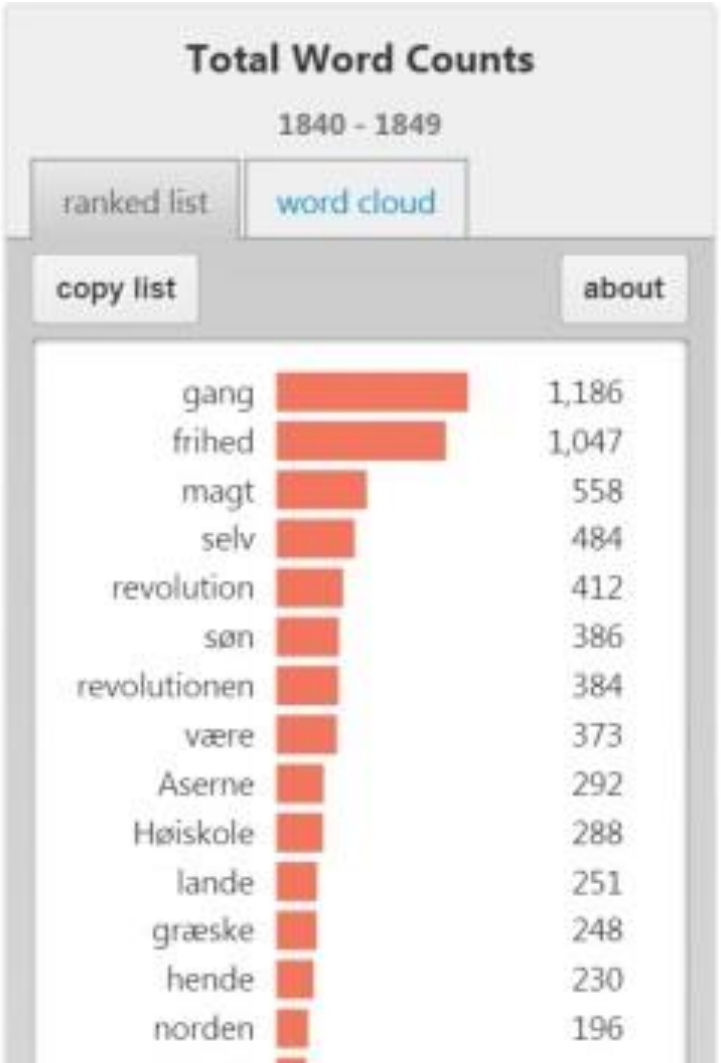
Working with formats which are easy for researchers to add information into

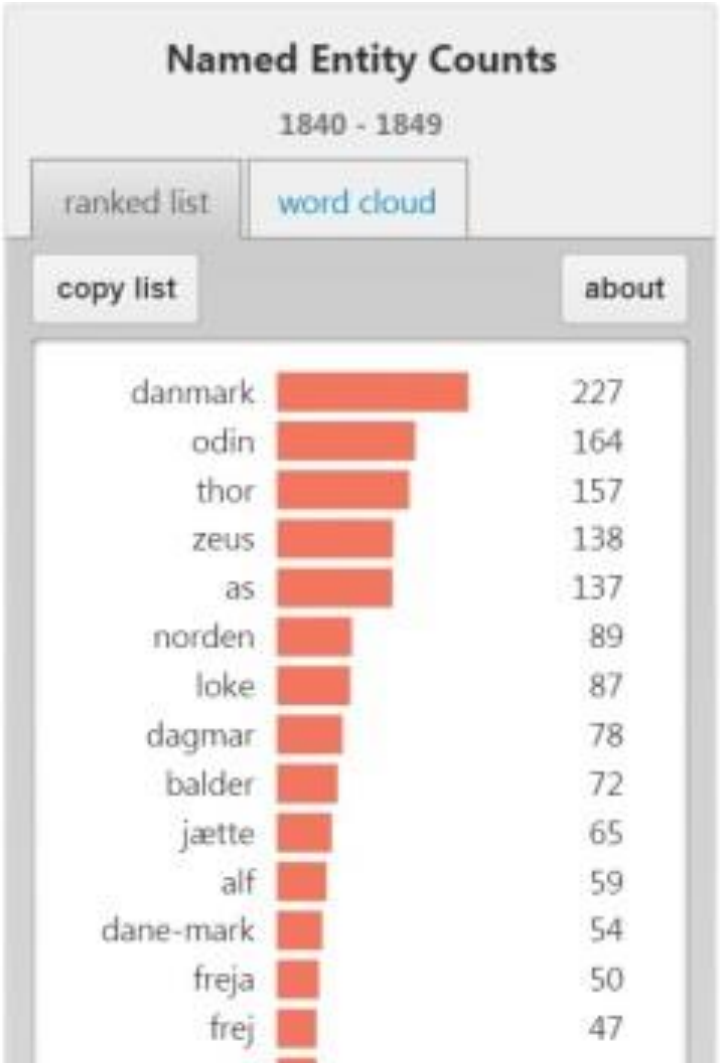
…. and for us to easily configure for other use cases

TEI and JSON

Plans and visions

Timeline selection of data:





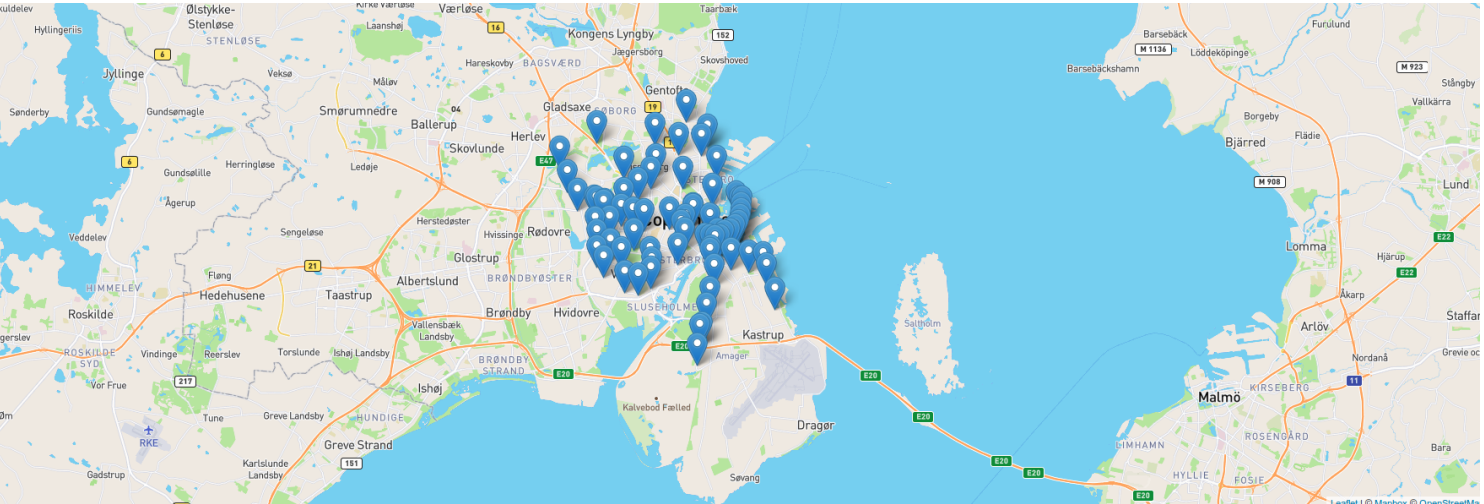
Plans and visions

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 Haltrup Hansen

About Smurf: contact:
sblabs@statsbiblioteket.dk

UNIVERSITY OF COPENHAGEN