CLARIN / CLARINO: An infrastructure supporting Open Science in the Digital Humanities

Koenraad De Smedt

Universitetet i Bergen

desmedt@uib.no

Tromsø

November 30 – December 1, 2017

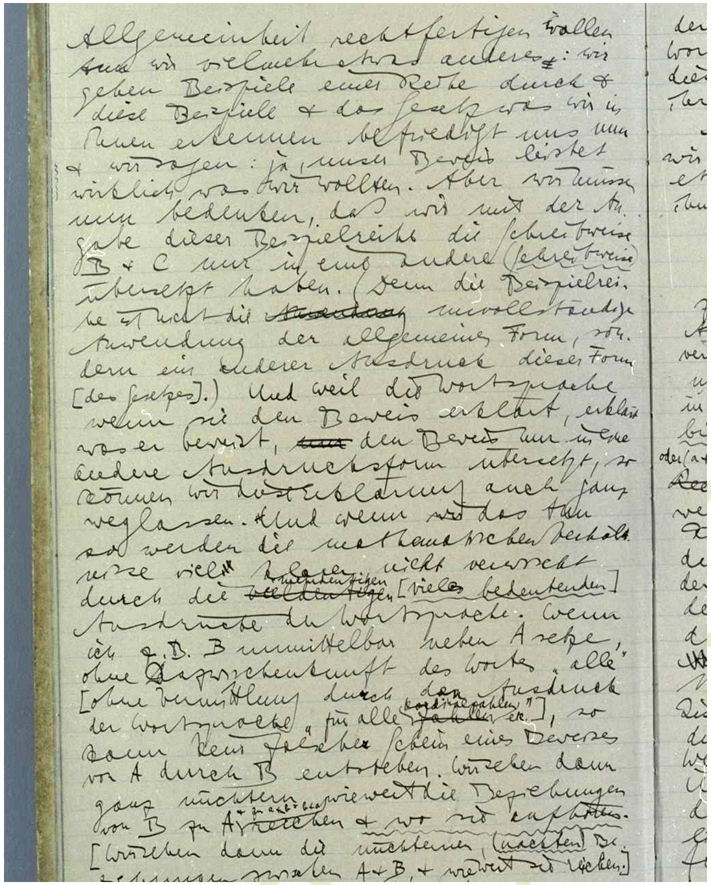
Digital curation at the core of DH

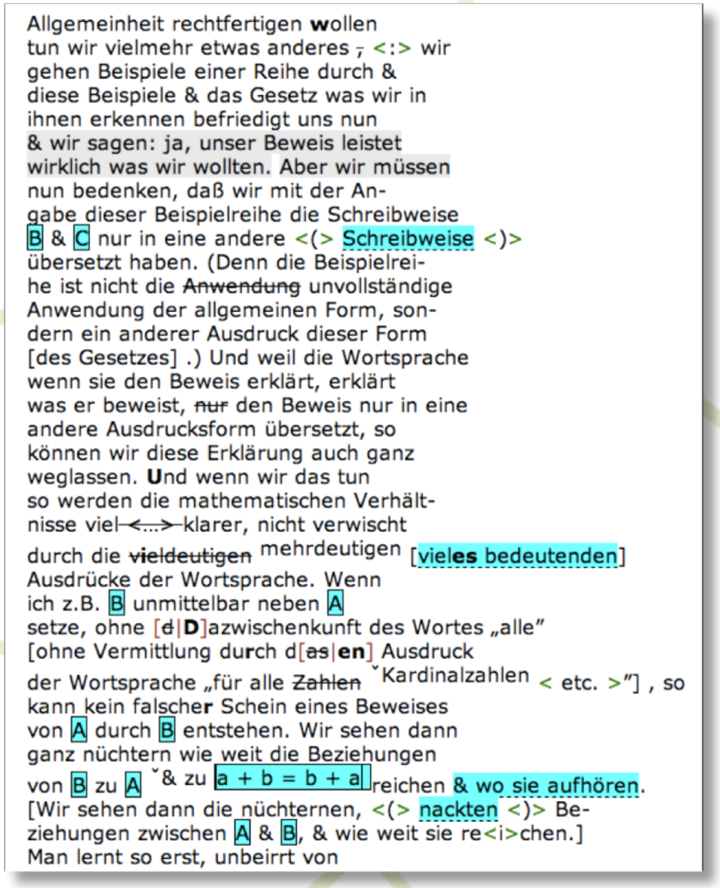
“Curation, analysis, editing, and modeling comprise fundamental activities at the core of DH.”

“The capacity with digital media to create enhanced forms of curation brings humanistic values into play in ways that were difficult to achieve in traditional museum or library settings.”

(Burdick et al. 2012:17–18)

Organized support for digital curation, analysis, editing and modeling involves “platforms, tools, and infrastructures” which “depend upon the basic building blocks of digital activity: digitization, classification, description and metadata, organization, and navigation” (Burdick et al. 2012:17)



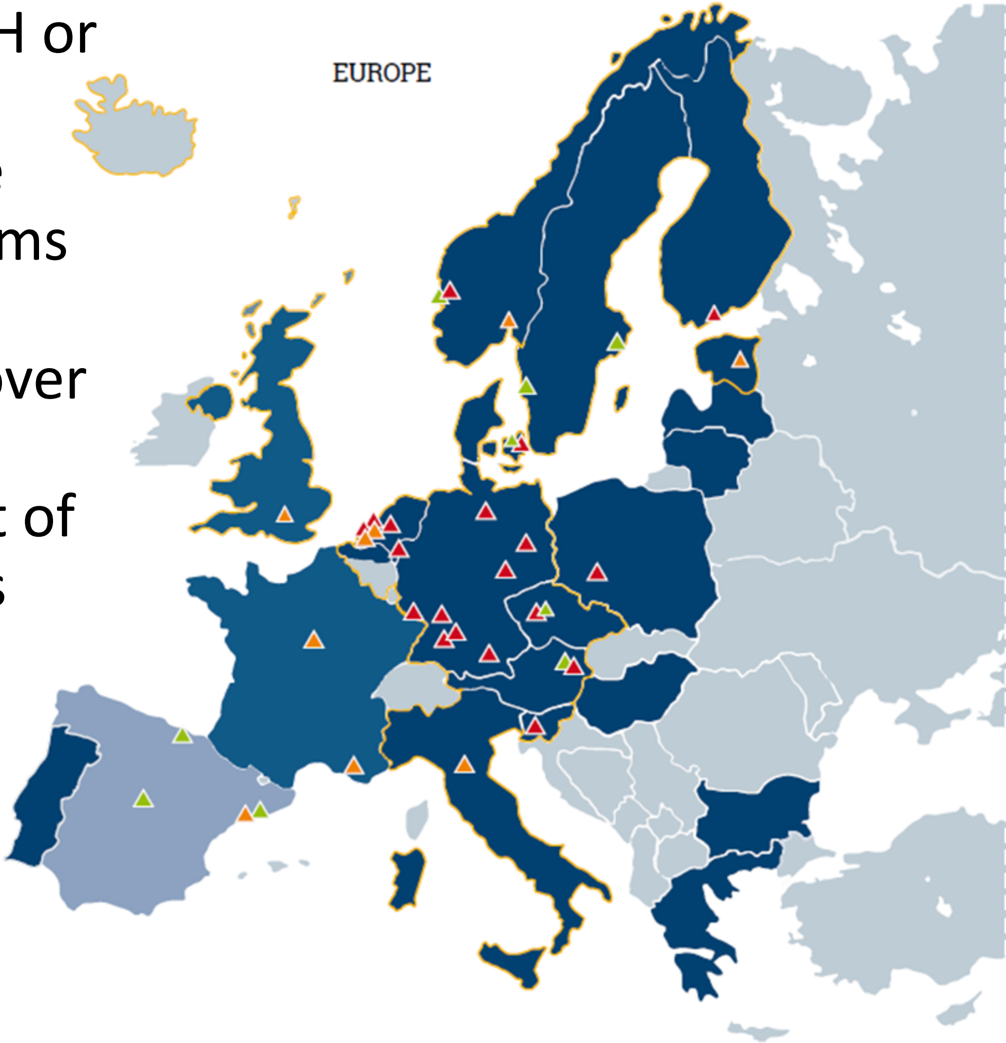


Allgemeinheit rechtfertigen **w**ollen
tun wir vielmehr etwas anderes ~~;~~ <:> wir
gehen Beispiele einer Reihe durch &
diese Beispiele & das Gesetz was wir in
ihnen erkennen befriedigt uns nun
& wir sagen: ja, unser Beweis leistet
wirklich was wir wollten. Aber wir müssen
nun bedenken, daß wir mit der An-
gabe dieser Beispielreihe die Schreibweise
B & C nur in eine andere <(> Schreibweise <)>
übersetzt haben. (Denn die Beispielrei-
he ist nicht die ~~Anwendung~~ unvollständige
Anwendung der allgemeinen Form, son-
dern ein anderer Ausdruck dieser Form
[des Gesetzes] .) Und weil die Wortsprache
wenn sie den Beweis erklärt, erklärt
was er beweist, ~~nur~~ den Beweis nur in eine
andere Ausdrucksform übersetzt, so
können wir diese Erklärung auch ganz
weglassen. **U**nd wenn wir das tun
so werden die mathematischen Verhält-
nisse viel ~~<...>~~ klarer, nicht verwischt
durch die ~~vieldeutigen~~ mehrdeutigen [viel**es** bedeutenden]
Ausdrücke der Wortsprache. Wenn
ich z.B. B unmittelbar neben A
setze, ohne [~~d~~|**D**]azwischenkunft des Wortes „alle”
[ohne Vermittlung durch d[~~as~~|**en**] Ausdruck
der Wortsprache „für alle ~~Zahlen~~ ˇKardinalzahlen < etc. >”] , so
kann kein falsche**r** Schein eines Beweises
von A durch B entstehen. Wir sehen dann
ganz nüchtern wie weit die Beziehungen
von B zu A ˇ& zu a + b = b + a reichen & wo sie aufhören.
[Wir sehen dann die nüchternen, <(> nackten <)> Be-
ziehungen zwischen A & B, & wie weit sie re<i>chen.]
Man lernt so erst, unbeirrt von

CLARIN ERIC: The European research infrastructure consortium for language resources

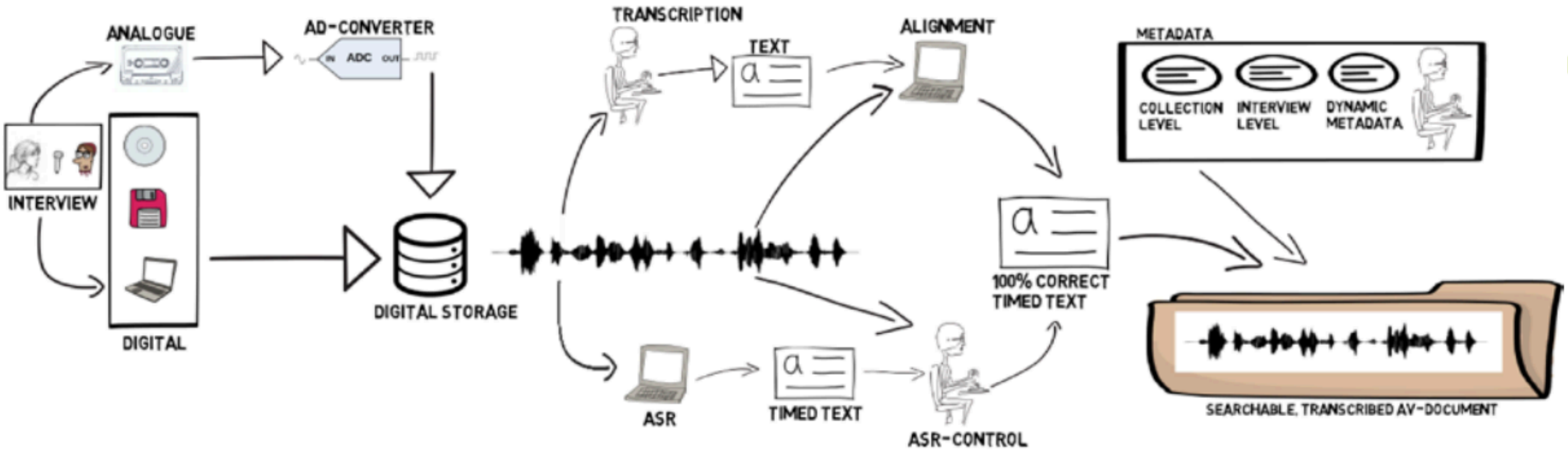
Not forcing a model on the DH or institutionalizing it, but contributing an infrastructure and meeting ground which aims to make “all digital language resources and tools from all over Europe and beyond [...] accessible [...] for the support of researchers in the humanities and social sciences”

(Maegaard et al. 2017:2)



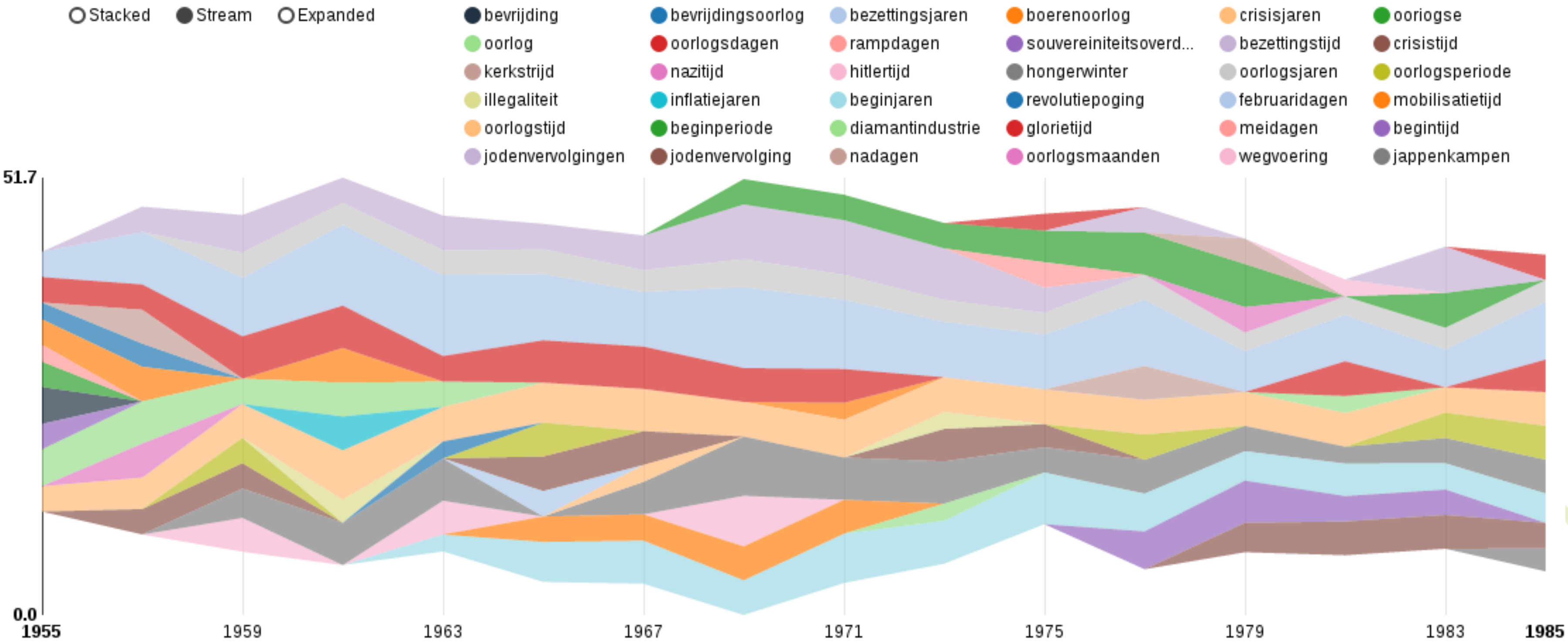
Some research questions

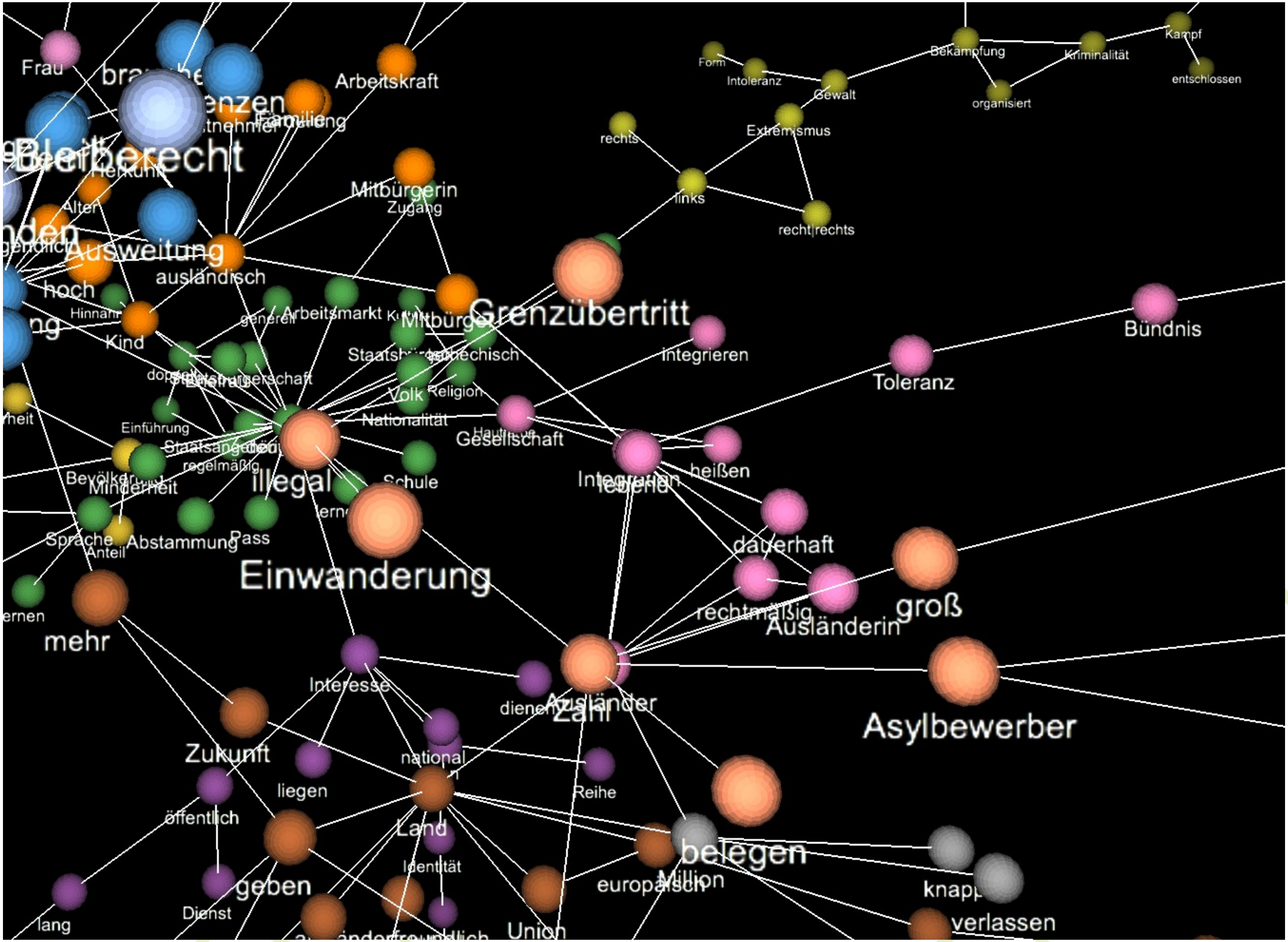
- Who was the real author of the Dutch national anthem?
- How was American consumer culture depicted in the Europe throughout the 20th century?
- How can we make a processing chain for curating and preserving oral history?



Some research questions

- How much polarization is there in social media discourse on climate change?
- Which challenges do language learners face in acquiring grammatical gender in a different language?
- What do historical documents tell us about the relation between gender and work?
- How can we visualize discourse concepts and attitudes by politicians of different parties?
- Which changing concepts are associated with *war* in newspapers?


○ Stacked
● Stream
○ Expanded
bevrijding
bevrijdingsoorlog
bezettingsjaren
boerenoorlog
crisisjaren
ooriogse
oorlog
oorlogsdagen
rampdagen
souvereiniteitsoverd...
bezettingstijd
crisistijd
kerkstrijd
nazitijd
hitlertijd
hongerwinter
oorlogsjaren
oorlogsperiode
illegaliteit
inflatiejaren
beginjaren
revolutiepoging
februaridagen
mobilisatietijd
oorlogstijd
beginperiode
diamantindustrie
glorietijd
meidagen
begintijd
jodenvervolgingen
jodenvervolging
nadagen
oorlogsmaanden
wegvoering
jappenkampen
51.7
0.0
1955
1959
1963
1967
1971
1975
1979
1983
1985


Frau
Arbeitskraft
Familie
Bleiberecht
Herkunft
Alter
Ausweitung
ausländisch
hoch
Hinnahme
Kind
Mitbürgerin
Zugang
Arbeitsmarkt
Grenzübertritt
integrieren
Staatsbürgerschaft
Religion
Volk
Nationalität
Einführung
regelmäßig
Minderheit
Bevölkerung
illegal
Schule
Sprache
Anteil
Abstammung
Pass
Einwanderung
mehr
Gesellschaft
Integration
heißen
dauerhaft
rechtmäßig
Ausländerin
groß
Toleranz
Bündnis
Asylbewerber
Interesse
dienen
Ausländer
Zahl
Zukunft
liegen
national
Reihe
öffentlich
Land
Identität
geben
Dienst
lang
Union
europäisch
Million
belegen
knapp
verlassen
Form
Intoleranz
Gewalt
Extremismus
rechts
links
recht|rechts
Bekämpfung
organisiert
Kriminalität
Kampf
entschlossen

Enhancing access and (re-)usability

“are the barriers to entry that ‘outsiders’ perceive real usability issues, or simply points on DH’s learning curve?”

(Edwards 2012, p. 213)

CLARIN catalog

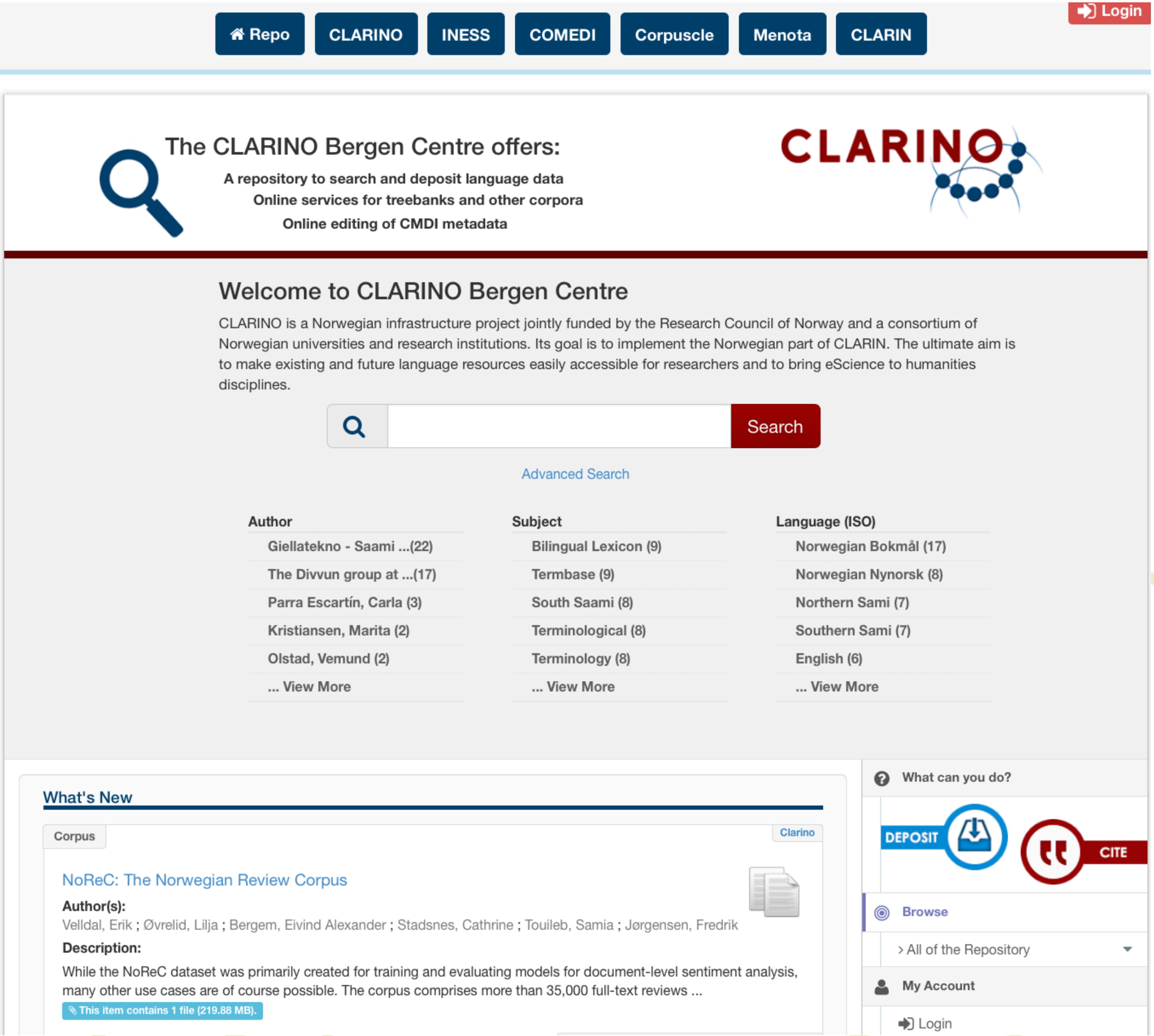
- Virtual Language Observatory (VLO), a registry of Language Resources (LRs) http://vlo.clarin.eu
- 1,600,000 records (including recent addition of Europeana records)
- Faceted search
- Cross-sectorial collaboration with the GLAM sector (Galleries, Libraries, Archives, Museums) planned

FAIR principles

- *Findable:* data must be registered with a persistent ID and items must be collected in a catalog
- *Accessible:* open access protocol (subject to restrictions), clear procedure for authentication and authorization
- *Interoperable:* documented descriptive vocabulary, standards for data and metadata coding
- *Re-usable:* clear licenses, understandable documentation (including provenance), compatibility with community standards and tools

CLARINO: CLARIN in Norway

- Certified centres with repositories
  - UiB/UB (45 downloadable datasets)
  - UiO/Tekstlab (16 downloadable datasets and 2 tools)
  - NB/Språkbanken (42 downloadable datasets)
- Other related centres
  - UiT/Trolling (60 downloadable datasets)
  - UiT/Giellatekno
- National catalog at the National Library of Norway which harvests metadata from other centres



The CLARINO Bergen Centre offers:

A repository to search and deposit language data
Online services for treebanks and other corpora
Online editing of CMDI metadata

Welcome to CLARINO Bergen Centre

CLARINO is a Norwegian infrastructure project jointly funded by the Research Council of Norway and a consortium of Norwegian universities and research institutions. Its goal is to implement the Norwegian part of CLARIN. The ultimate aim is to make existing and future language resources easily accessible for researchers and to bring eScience to humanities disciplines.

Author	Subject	Language (ISO)
Giellatekno - Saami ...(22)	Bilingual Lexicon (9)	Norwegian Bokmål (17)
The Divvun group at ...(17)	Termbase (9)	Norwegian Nynorsk (8)
Parra Escartín, Carla (3)	South Saami (8)	Northern Sami (7)
Kristiansen, Marita (2)	Terminological (8)	Southern Sami (7)
Olstad, Vemund (2)	Terminology (8)	English (6)
... View More	... View More	... View More

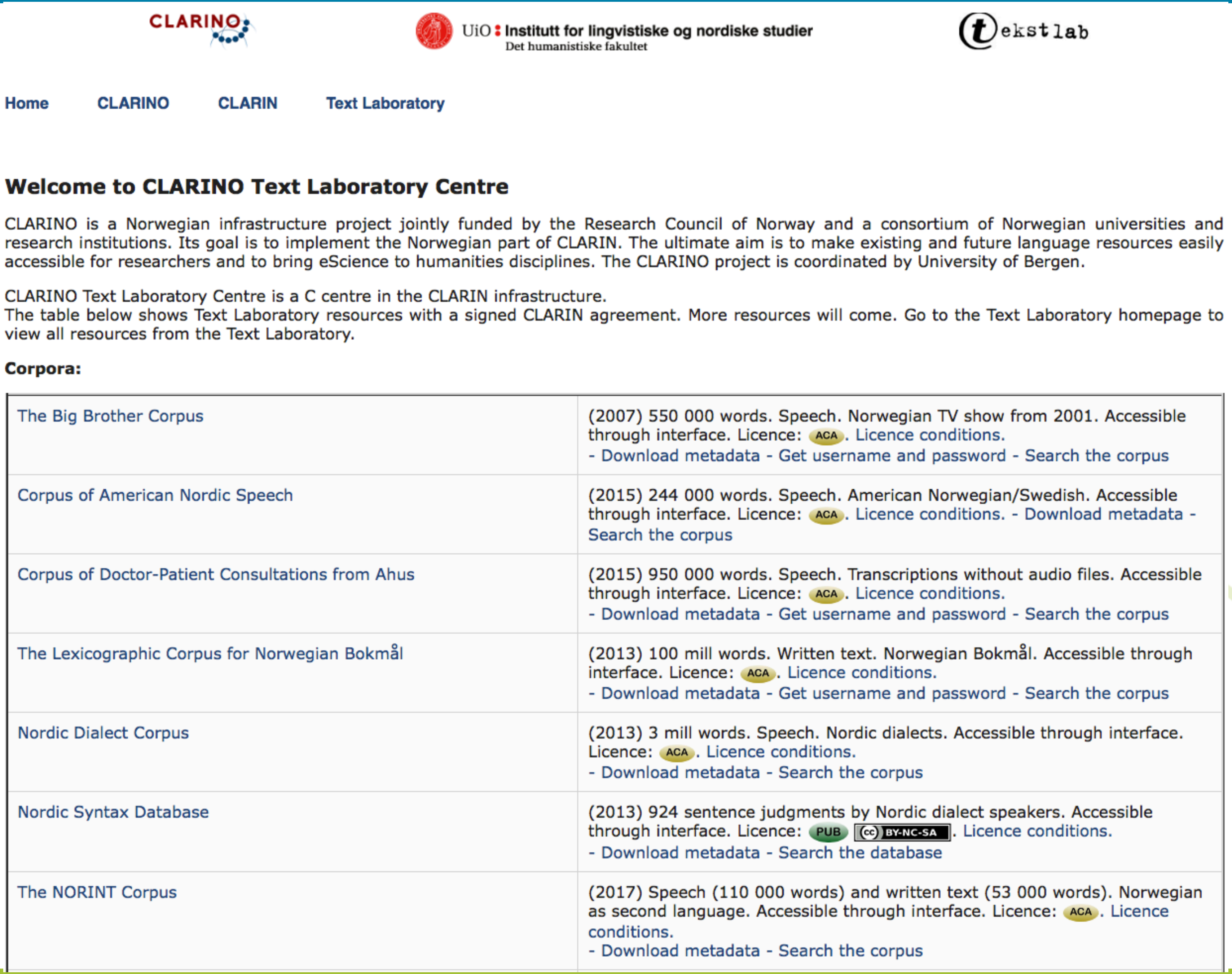
What's New

NoReC: The Norwegian Review Corpus

Author(s):
Velldal, Erik ; Øvrelid, Lilja ; Bergem, Eivind Alexander ; Stadsnes, Cathrine ; Touileb, Samia ; Jørgensen, Fredrik

Description:
While the NoReC dataset was primarily created for training and evaluating models for document-level sentiment analysis, many other use cases are of course possible. The corpus comprises more than 35,000 full-text reviews ...

This item contains 1 file (219.88 MB).


CLARINO
UiO : Institutt for lingvistiske og nordiske studier
Det humanistiske fakultet
tekstlab
Home
CLARINO
CLARIN
Text Laboratory

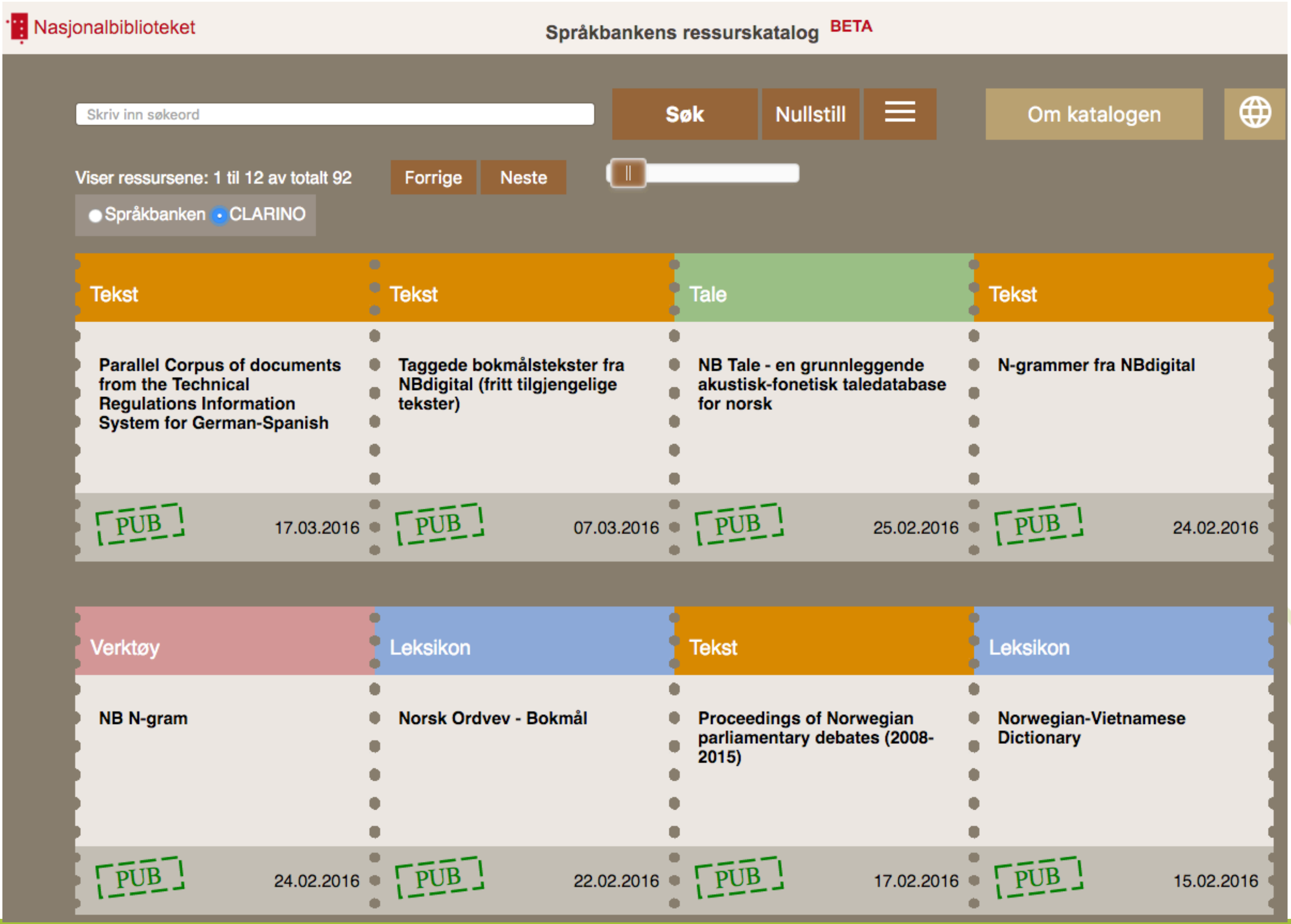
Welcome to CLARINO Text Laboratory Centre

CLARINO is a Norwegian infrastructure project jointly funded by the Research Council of Norway and a consortium of Norwegian universities and research institutions. Its goal is to implement the Norwegian part of CLARIN. The ultimate aim is to make existing and future language resources easily accessible for researchers and to bring eScience to humanities disciplines. The CLARINO project is coordinated by University of Bergen.

CLARINO Text Laboratory Centre is a C centre in the CLARIN infrastructure.
The table below shows Text Laboratory resources with a signed CLARIN agreement. More resources will come. Go to the Text Laboratory homepage to view all resources from the Text Laboratory.

Corpora:

The Big Brother Corpus	(2007) 550 000 words. Speech. Norwegian TV show from 2001. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Get username and password - Search the corpus
Corpus of American Nordic Speech	(2015) 244 000 words. Speech. American Norwegian/Swedish. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Search the corpus
Corpus of Doctor-Patient Consultations from Ahus	(2015) 950 000 words. Speech. Transcriptions without audio files. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Get username and password - Search the corpus
The Lexicographic Corpus for Norwegian Bokmål	(2013) 100 mill words. Written text. Norwegian Bokmål. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Get username and password - Search the corpus
Nordic Dialect Corpus	(2013) 3 mill words. Speech. Nordic dialects. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Search the corpus
Nordic Syntax Database	(2013) 924 sentence judgments by Nordic dialect speakers. Accessible through interface. Licence: PUB (cc) BY-NC-SA. Licence conditions. - Download metadata - Search the database
The NORINT Corpus	(2017) Speech (110 000 words) and written text (53 000 words). Norwegian as second language. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Search the corpus



Språkbankens ressurskatalog BETA

Skriv inn søkeord

Viser ressursene: 1 til 12 av totalt 92

○ Språkbanken ● CLARINO

Tekst

Parallel Corpus of documents from the Technical Regulations Information System for German-Spanish

PUB 17.03.2016

Tekst

Taggede bokmålstekster fra NBdigital (fritt tilgjengelige tekster)

PUB 07.03.2016

Tale

NB Tale - en grunnleggende akustisk-fonetisk taledatabase for norsk

PUB 25.02.2016

Tekst

N-grammer fra NBdigital

PUB 24.02.2016

Verktøy

NB N-gram

PUB 24.02.2016

Leksikon

Norsk Ordvev - Bokmål

PUB 22.02.2016

Tekst

Proceedings of Norwegian parliamentary debates (2008-2015)

PUB 17.02.2016

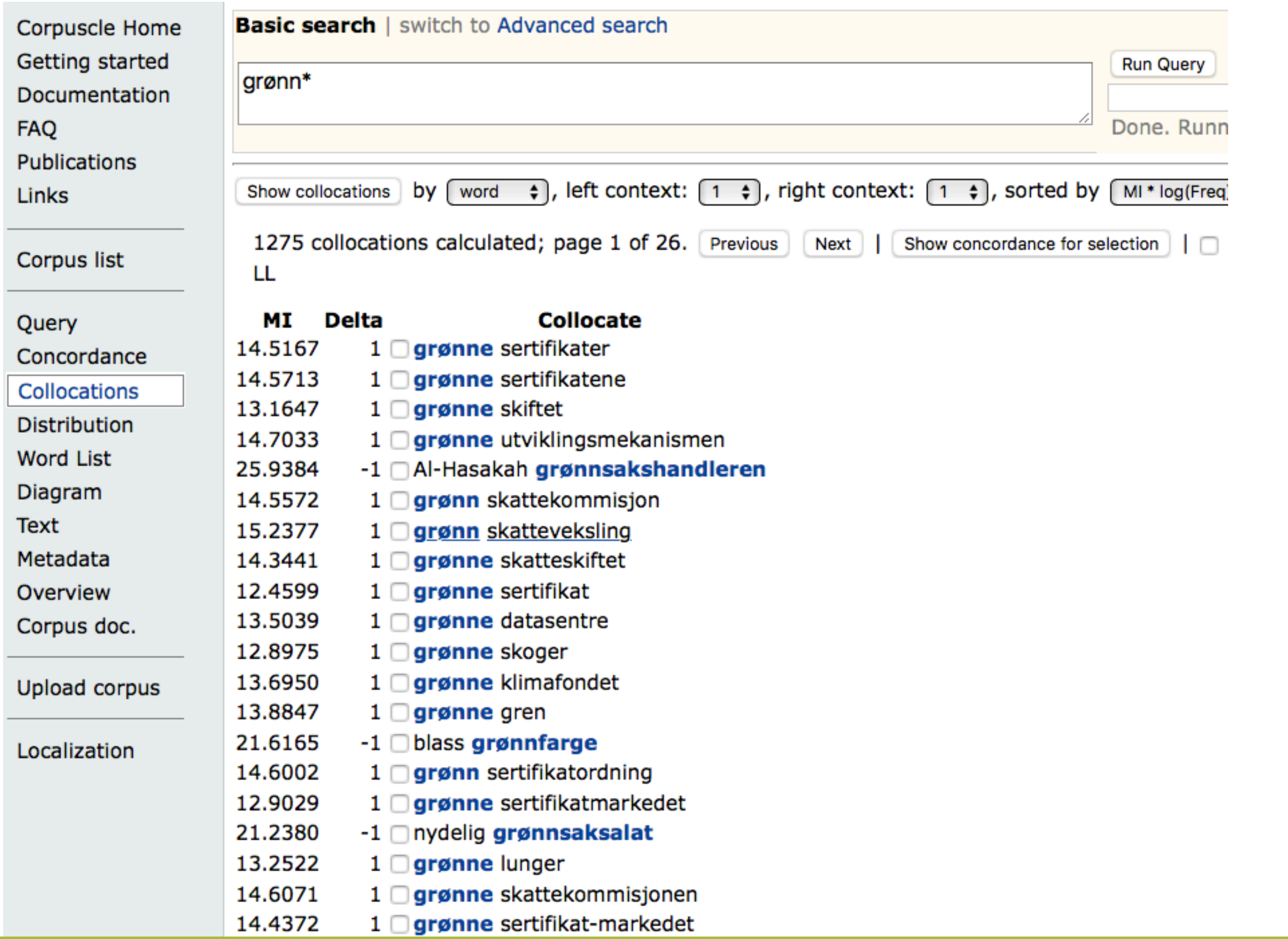
Leksikon

Norwegian-Vietnamese Dictionary

PUB 15.02.2016

CLARINO: More than repositories

- Corpus tools: Glossa (UiO/Text Lab), Corpuscle and INESS (Uni Research Computing/UiB)
- Term Portal (NHH)
- Language Analysis Portal LAP (UiO/IFI)
- Metadata tool COMEDI (Uni Research Computing/UiB)
- Dynamic interactive presentation of digital editions (UiO/EDD)
- Dissemination actions



grønn*

Done. Runn

Show collocations by word, left context: 1, right context: 1, sorted by MI * log(Freq

1275 collocations calculated; page 1 of 26. | □

LL

MI	Delta	Collocate
14.5167	1	□ **grønne** sertifikater
14.5713	1	□ **grønne** sertifikatene
13.1647	1	□ **grønne** skiftet
14.7033	1	□ **grønne** utviklingsmekanismen
25.9384	-1	□ Al-Hasakah **grønnsakshandleren**
14.5572	1	□ **grønn** skattekommisjon
15.2377	1	□ **grønn** skatteveksling
14.3441	1	□ **grønne** skatteskiftet
12.4599	1	□ **grønne** sertifikat
13.5039	1	□ **grønne** datasentre
12.8975	1	□ **grønne** skoger
13.6950	1	□ **grønne** klimafondet
13.8847	1	□ **grønne** gren
21.6165	-1	□ blass **grønnfarge**
14.6002	1	□ **grønn** sertifikatordning
12.9029	1	□ **grønne** sertifikatmarkedet
21.2380	-1	□ nydelig **grønnsaksalat**
13.2522	1	□ **grønne** lunger
14.6071	1	□ **grønne** skattekommisjonen
14.4372	1	□ **grønne** sertifikat-markedet

CLARIN priorities

- Uptake by researchers: outreach to all humanities disciplines (researcher training courses, workshops, etc), service enhancements for consistent user experience
- Technical infrastructure: towards an integrated, interoperable infrastructure (technical centres, services, licenses etc.)
- Knowledge sharing: knowledge centres, mobility grants, video lectures, course registry (with DARIAH)
- Sustainability: extension to new countries, cooperation with GLAM sector, commitments from stakeholders and funders, cooperation with other infrastructures

Links

http://clarin.eu

http://clarino.uib.no