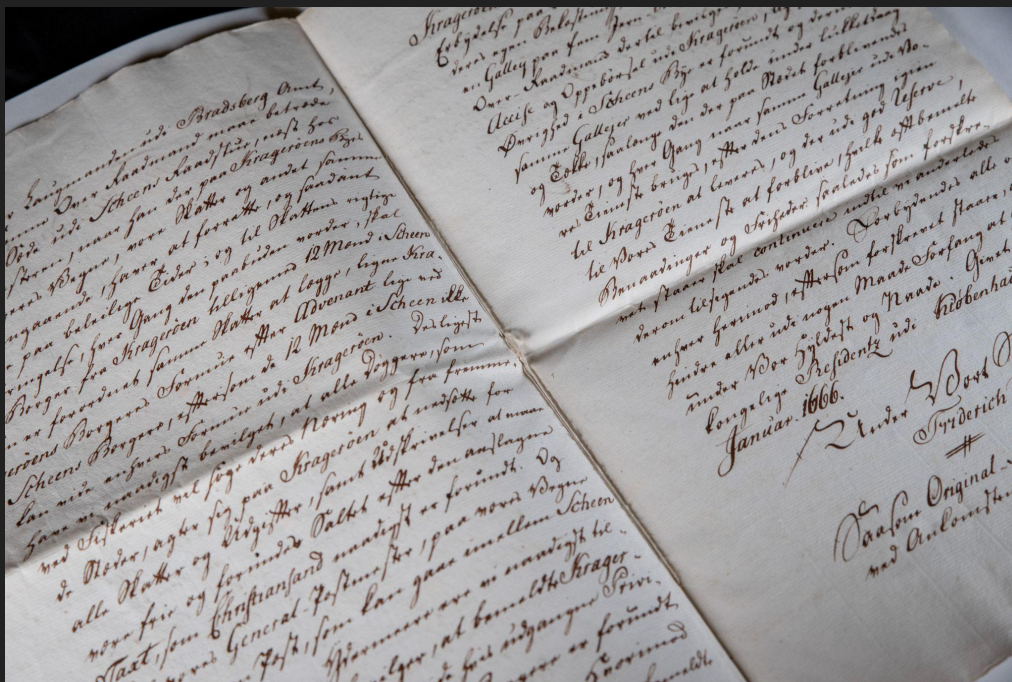


Tilgang til data: Språkteknologi og Nasjonalbiblioteket



Magnus Breder Birkenes
Språkbanken

Språkbanken

- nasjonal infrastruktur for språkteknologi
- motvirke domenetap
- utvikler og deler datasett og andre grunnlagsressurser
- åpne lisenser
- tilgjengelig gjennom vår ressurskatalog, som speiles på data.norge.no og data.europa.eu

Nasjonalbiblioteket | Språkbanken English

Språkbanken **Ressurskatalog**

I samarbeid med CLARINO

Søk i ressurskatalogen ... 🔍

1-10 av 73 treff 1 2 3 ... 8 → Per side ▾ Sist oppdatert ▾

Origin: Språkbanken ✕

Type	Antall
<input checked="" type="radio"/> Alle typer	
<input type="radio"/> Tekst	36
<input type="radio"/> Leksikon	19
<input type="radio"/> Tale	17
<input type="radio"/> Verktøy	5
<input type="radio"/> Video	2

Opphav	Antall
<input type="radio"/> Alle opphav	
<input type="radio"/> CLARINO Bergen	303
<input checked="" type="radio"/> Språkbanken	73
<input type="radio"/> CLARINO Tekstlaboratoriet	25

TEKST 07.09.2022

Stortingsforhandlinger 1814-2000

Dette korpuset inneholder publiserte historiske stortingsforhandlinger fra Stortinget for perioden 1814-2000. De til sammen 2136 bindene ble digitalisert, OCR-lest og prosessert ved ...

Språk: norsk
Opphav: Språkbanken
Lisens: Norwegian Licence for Open Government Data (NLOD)

VERKTØY 04.04.2022

NB N-gram

NB N-gram er ei teneste som gir deg høve til å finne og samanlikne ordfrekvensar, til dømes når og kor ofte ord vert nytta i eit historisk perspektiv. NB N-gram er basert på digitaliserte bøker ...

Språk:
Opphav: Språkbanken
Lisens: Creative_Commons-ZERO (CC-ZERO)

TALE, TEKST 11.03.2022

NST norsk ATG-database (16 kHz) – reorganisert

Denne databasen er laget av Nordisk

TALE, TEKST 11.03.2022

NST dansk ATG-database (16 kHz) – reorganisert

Denne databasen er laget av Nordisk

Internasjonalt samarbeid

Nasjonalbiblioteket deltar i de sentrale europeiske forskningsinfrastrukturene innenfor språkteknologi:

- CLARIN
- DARIAH



The research infrastructure for language as social and cultural data

CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.



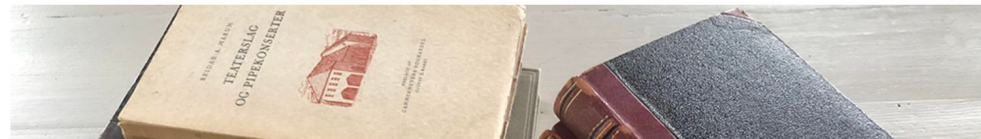
The pan-European infrastructure for arts & humanities scholars

[Learn More About DARIAH](#)

Forskningsprosjekter



NORN - Norske romantiske nasjonalismer ønsker å undersøke hvordan kollektive følelser knyttet til ideen om nasjonen ble utnyttet, aktivert og konstruert gjennom litteratur og teater.



QUEERDOM



Naturhistoriens etterliv

Hvordan er naturkunnskap og humanistisk kunnskap sammenfiltret – historisk og i dag?



Made Abroad

Producing Norwegian World Literature in a Time of Rupture, 1900–50 (MAP)



ParlaMint



FoodLessons: Kulinarisk arv som en ressurs i bygging av "Matnasjonen Norge 2030"

ImagiNation. Mapping the Imagined Geographies of Norwegian Literature from 1814 to 1905

Alternativ tittel: ImagiNasjon. En kartlegging av forestilte geografier i norsk litteratur fra 1814 til 1905

Uferdig fortid

Norge og andre verdenskrig i samtidens estetiske minnekultur

HUGIN-MUNIN - AI Access to Handwritten Norwegian Cultural Heritage

Et av verdens største korpus

- Nasjonalbiblioteket har et av verdens største tilgjengelige korpus
- et svært gunstig utgangspunkt for å utvikle språkteknologi
- presis informasjon om hver tekst (gode metadata)
- 150 milliarder løpende ord som kan brukes til forskning og utvikling



Bygging av språkressurser

Målfrid-prosjektet

- samarbeidsprosjekt mellom Språkrådet og Nasjonalbiblioteket på oppdrag fra Kulturdepartementet
- mål: automatisere målformsrapportering (kravet om minst 25% nynorsk/bokmål i statlig informasjonsmateriale)
- middel: høsting av statlige nettsider, prosessering og automatisk språkdeteksjon av materiale



The screenshot shows the 'Felles datakatalog' (Common Data Catalog) interface. The main heading is 'Målfrid 2021 - Fritt tilgjengelige tekster fra norske statlege nettsider', published on 02.05.2022. The dataset is owned by 'Nasjonalbiblioteket' and has a metadata quality of 57%. It is categorized under 'Allmenn tilgang' (Public access) and 'Vitenskap og teknologi' (Science and technology). The description states that the corpus contains 339 Norwegian state institutions, totaling 4.1 billion words. The distribution section shows the URL 'https://hdl.handle.net/21.11146/69' and download options for PDF and other formats.

Høsting

- verktøy: *wget*
- rekursiv høsting ned til nivå 12 (det dypeste nivået vi fant under en prøveinnhøsting)
- kun tekstdokumenter (HTML, PDF, DOC)
- alle statlige domener med visse unntak
- alle undersider med unntak (f.eks. ikke metadataregistre)
- lagringsformat: WARC

```
WARC/1.0^M
WARC-Type: response^M
WARC-Record-ID: <urn:uuid:a4cf63fa-3854-46b9-9d22-1e8b60e1715f>^M
WARC-WarcInfo-ID: <urn:uuid:b8ae80cf-39c7-4e21-81aa-bffd8dd97fdf>^M
WARC-Concurrent-To: <urn:uuid:cbe5ae2-cd4f-4005-b57c-225076b8d66c>^M
WARC-Target-URI: <https://lanekassen.no/>^M
WARC-Date: 2020-12-15T00:18:53Z^M
WARC-IP-Address: 40.114.181.184^M
WARC-Block-Digest: sha1:5TJ377RL34V0QELV5FPZGVAAN2N5V574^M
WARC-Payload-Digest: sha1:PH7QGNDYX76Y4KM70LRZXTYOIZYGNWGW^M
Content-Type: application/http;msgtype=response^M
Content-Length: 50934^M
^M
HTTP/1.1 200 OK^M
Date: Tue, 15 Dec 2020 00:18:53 GMT^M
Content-Type: text/html; charset=utf-8^M
Content-Length: 49759^M
Connection: keep-alive^M
Cache-Control: private^M
Server: Microsoft-IIS/10.0^M
Content-Security-Policy: default-src 'self';script-src 'self' 'unsafe-inline' 'nonce-45t9w0INlA5tBNDhQZayI9k
js/siteanalyze_6003145.js https://76e1b9fe31ad43799173abdc538c2f99.westeurope.azure.elastic-cloud.com:9243;o
img-src 'self' data: https://i.ytimg.com https://6003145.global.siteimproveanalytics.io https://szsurvey.stt
t-src 'self' https://fonts.gstatic.com;connect-src 'self';base-uri 'self';form-action 'self' https://76e1b9f
f-cloud.com:9243;manifest-src 'none';upgrade-insecure-requests;block-all-mixed-content;report-uri /WebResourc
Strict-Transport-Security: max-age=31536000^M
X-Frame-Options: SameOrigin^M
X-XSS-Protection: 1; mode=block^M
X-Content-Type-Options: nosniff^M
X-Download-Options: noopen^M
Set-Cookie: EPI.StateMarker=true; path=/; secure^M
^M
^M
^M
^M
<!DOCTYPE html>^M
<html lang="nb-NO">^M
<head>^M
```


Prosessering

- HTML: “Boilerplate removal”, Jstext
- PDF: full OCR med Google Cloud Vision API
- WORD/ODT: Python-biblioteket Textract
- alle dokumenter dedupliseres på domenenivå (sjekksom av redusert representasjon)

BANE NOR Hva ser du etter?

nyheter Om oss Slik fungerer jernbanen Prosjekter Kundeportal Leverandør Karriere Norsk jernbaneskole Kundesenter Mer tog

Forside > Om oss > Om Bane NOR

- > Forretningsplan
- > **Om Bane NOR**
- > Nasjonal transportplan
- > Organisasjon og ledelse
- > Eierskap og styring
- > Investor Relations
- > Jernbanereformen
- > Miljø
- > Etiske retningslinjer
- > Varslingskanal
- > Bane NORs historie

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunks- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Søk herSøk

Meny

Forside

Om oss

Om Bane NOR

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen. Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunks- og godsterminalutvikling. Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Skriv ut

Last ned dokumenter

- Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf
- Brochure Bane NOR - We create the railway of the future.pdf

final class	good
cotext-free class	good
heading	False
length (in characters)	157
number of characters within links	0
link density	0.000
number of words	18
number of stopwords	6
stopword density	0.333
html.body.main.div.div.article.div.p	

Språkdeteksjon

- Frekvens av bokstavsekvenser og enkeltord
- Algoritme: Out-of-place rank order (Cavnar/Trenkle 1994)
- Implementasjon: Textcat
- Modeller for bokmål, nynorsk, samiske språk og andre språk (fra UIT/Giellatekno)
- Klassifikasjon på dokument- og avsnittnivå

¶	paragraph_nr		line	lang	tokens
0	0	VEDTEKTER FOR STATENS LÅNEKASSE FOR UTDANNING			6
1	1	Fastsatt av Kunnskapsdepartementet den 15.januar 2016 i medhold av lov 3. juni 2005 nr. 37 om utdanningsstøtte.		nob	17
2	2	1 Formål			2
3	3	Statens lånekasse for utdanning (Lånekassen) skal forvalte utdanningsstøtten i samsvar med bestemmelsene gitt i eller i medhold av utdanningsstøtteleven. Rammene for virksomheten til Lånekassen fastsettes for øvrig av overordnet myndighet.		nob	30
4	4	Utfyllende bestemmelser om virksomhetens formål framgår av Hovedinstruks for Lånekassen.			10
5	5	2 Lånekassens ledelse og oppgaver			5
6	6	Lånekassen er et ordinært forvaltningsorgan som ledes av eget styre og tilsatt administrerende direktør.		nob	14
7	7	Styret er Lånekassens øverste organ og er ansvarlig for den samlede virksomheten. Styret skal blant annet påse at Lånekassen drives i samsvar med regelverket, samt styringssignaler og retningslinjer gitt av overordnede myndigheter. Styret skal sikre at fastsatte mål oppnås og at ressursbruken er effektiv. Styret har også ansvar for at det foreligger vurderinger som identifiserer de viktigste risikofaktorene knyttet til virksomheten og at det foreligger planer for oppfølging av disse faktorene.		nob	71
8	8	Styret skal fastsette Lånekassens strategi og sørge for at denne blir realisert. Styret skal behandle forslag fra Lånekassen om vesentlige endringer i utdanningsstøtteordningen, samt øvrige viktige saker for Lånekassens virksomhet.		nob	30
9	9	Styret ansetter administrerende direktør på åremål. Betingelsene for ansettelsesforholdet fastsettes av Kunnskapsdepartementet.		nob	12

Ressurser for taleteknologi

TALE 30.11.2021

Stortingskorpuset

Dette er versjon 1.1 av Stortingskorpuset (engelsk forkorting NPSC). Korpuset er utvikla ved Språkbanken på Nasjonalbiblioteket. NPSC er sett saman av lydopptak av møte i Stortinget, ortografisk ...

Språk: norsk
Opphav: Språkbanken
Lisens: Creative_Commons-ZERO (CC-ZERO)

TALE, TEKST 15.12.2022

Norsk talestyringskorpuser

Norsk talestyringskorpuser (engelsk forkorting NVCC) er eit tekst- og talekorpuser som består av skrivne og innlesne setningar (spørjingar). Dette er spørjingar ein typisk nyttar til å styre t.d. ...

Språk: norsk
Opphav: Språkbanken
Lisens: Creative_Commons-ZERO (CC-ZERO)

LEKSIKON 13.03.2023

NB Uttale

NB Uttale er et uttaleleksikon for bokmål laget av Språkbanken. Leksikonet består av 785 000 ord som er fonemisk transkribert til fem dialekter, dvs. dialektområder; østnorsk (Agder, Innlandet, ...

Språk: norsk
Opphav: Språkbanken
Lisens: Creative_Commons-ZERO (CC-ZERO)

**Trenings-
data
talegjen-
kjenning**

**Testsett
talegjen-
kjenning**

**Stortings-
korpuset+**

Stortingskorpuset

- 140 timer transkribert tale
- åpent tilgjengelig
- 267 talere fra hele landet
- bokmål og nynorsk



Stortingskorpuset+

- Basert på ParlaSpeech-HR
- Ekstrahert fra referater
- Ikke ord-til-ord
- 5000 timer
- 600 timer nynorsk



NB Whisper

- Whisper: Flerspråklig talegjenkjenning fra OpenAI (2022)
- Utviklet fra videoer og undertekst hentet på nettet
- Fungerer godt på bokmål
- NB-Whisper (2024): forbedret versjon av Whisper på norsk utviklet av AI-laben med data fra Språkbanken og NRK



Norsk talestyringskorpus

- Talestyrte mobilassistenter
- 10 000 spørringer
- Dialektal tale



- spæll sangen fra byn'sjla – spill sangen fra begynnelsen
- æ vil vessta kor mytji klokka e – jeg vil vite hvor mye klokka er
- korsen dato e det – hvilken dato er det
- sle tå høgtalaren – slå av høyttaleren
- kest e Nasjonalbiblioteket – hvor er Nasjonalbiblioteket
- spel songen atte ittepå – spill sangen igjen etterpå
- du lyt åpne meldingi – du må åpne meldingen

NB Uttale

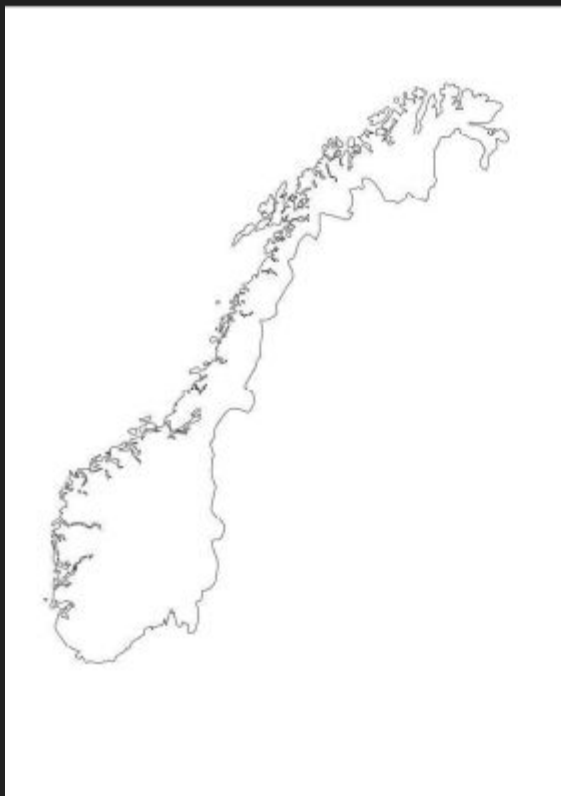
- SCRIBE
- Uttaleleksikon bokmål
- 800 000 ord
- 30 000 nyord
- Regulære uttrykk
- Nofabet

leser, L EE1 S AXo R → L AE1 S
sover, S OA1 V AXo R → S OE V

```
# Change stressed vowel and remove present tense suffix in
# L EE1 S AX0 R --> L AE1 S
# S OA1 V AX0 R --> S OE1 V
dialect_irreg_verbs_prs_n = {
  'name': 'dialect_irreg_verbs_prs_n',
  'areas': ['n_spoken'],
  'rules': [
    {
      'pattern': r'\bE(E|H))([13]) ([LST]) AX0 R$',
      'replacement': r'AE\2\3 \4',
      'constraints': [
        {
          "field": 'wordform',
          "pattern": r'(dett|les|selg|sett|sprett)er$',
          "is_regex": True
        },
        {
          "field": 'pos',
          "pattern": r'VB',
          "is_regex": False
        }
      ],
    },
    {
      'pattern': r'\bOA([13]) V AX0 R$',
      'replacement': r'OE\1 V',
      'constraints': [
        {
          "field": 'wordform',
          "pattern": r'sover$',
          "is_regex": True
        },
        {
          "field": 'pos',
          "pattern": r'VB',
          "is_regex": False
        }
      ],
    }
  ],
}
```


Testsett for talegjenkjenning

- 10 timer
- NRK-materiale
- Transkripsjon og normvariasjon
- Fem dialektområder
 - Østnorsk
 - Sørvestnorsk
 - Vestnorsk
 - Trøndersk
 - Nordnorsk
- Kjønn
- Alder



NYNORSKORBOKA |

byggje I, bygge I

byggja, bygga

VERB

VIS BØYING +



Mimir-prosjektet

Mimir-prosjektet

- behov for store mengder data i store språkmodeller
- Oppdrag fra Regjeringen (KUD) i desember 2023: “Departementet ber med dette Nasjonalbiblioteket sette i gang et koordinert forsknings-/utviklingsprosjekt for om mulig å undersøke verdien av opphavsrettslig beskyttet materiale i trening av norske generative språkmodeller.”
- partnere: NorwAI (NTNU), Language Technology Group (UiO), Sigma2



Mål

1. Undersøke verdien av opphavsrettslig beskyttet materiale i trening av norske generative språkmodeller
2. Vurdere grunnlaget for en eventuell kompensasjonsordning for norske rettighetshavere
 - a. i utgangspunktet litteratur
 - b. aviser tas med i tillegg
3. Utarbeide forslag til en kompensasjonsordning

Korpus

- **mimir base**
 - åpent datasett for fri tilgang uten begrensninger
 - aviser og bøker i det fri eller etter avtale
 - publikasjoner fra det offentlige
 - tilrettelagte data fra Språkbanken, f.eks. fra Web (Målfrid)
- **mimir extended**
 - internt datasett
 - inneholder mimir-base og i tillegg bøker og aviser under opphavsrett
- **varianter**
 - mimir base + rettighetsbelagt skjønnlitteratur
 - mimir base + rettighetsbelagte aviser

Status	Initialization	Data	Name
✓	From scratch	mimir-base	mimir-mistral-7b-base-scratch
✓	From scratch	mimir-extended	mimir-mistral-7b-extended-scratch
✓	Pre-existing	mimir-base	mimir-mistral-7b-base
✓	Pre-existing	mimir-extended	mimir-mistral-7b-extended
✓	mimir-mistral-7b-base-scratch	mimir-fiction	mimir-7b-fiction
✓	mimir-mistral-7b-base-scratch	mimir-nonfiction	mimir-7b-nonfiction
✓	mimir-mistral-7b-base-scratch	mimir-factual	mimir-7b-factual
✓	mimir-mistral-7b-base-scratch	mimir-newspapers	mimir-7b-newspapers
✓	mimir-mistral-7b-base-scratch	mimir-books	mimir-7b-books
✓	mimir-mistral-7b-base-scratch	mimir-rightholders	mimir-7b-rightholders
✓	mimir-mistral-7b-base-scratch	mimir-untranslated-withnewspapers	mimir-7b-untranslated-withnewspapers
✓	mimir-mistral-7b-base-scratch	mimir-untranslated	mimir-7b-untranslated
✓	mimir-mistral-7b-base-scratch	mimir-translated	mimir-7b-translated

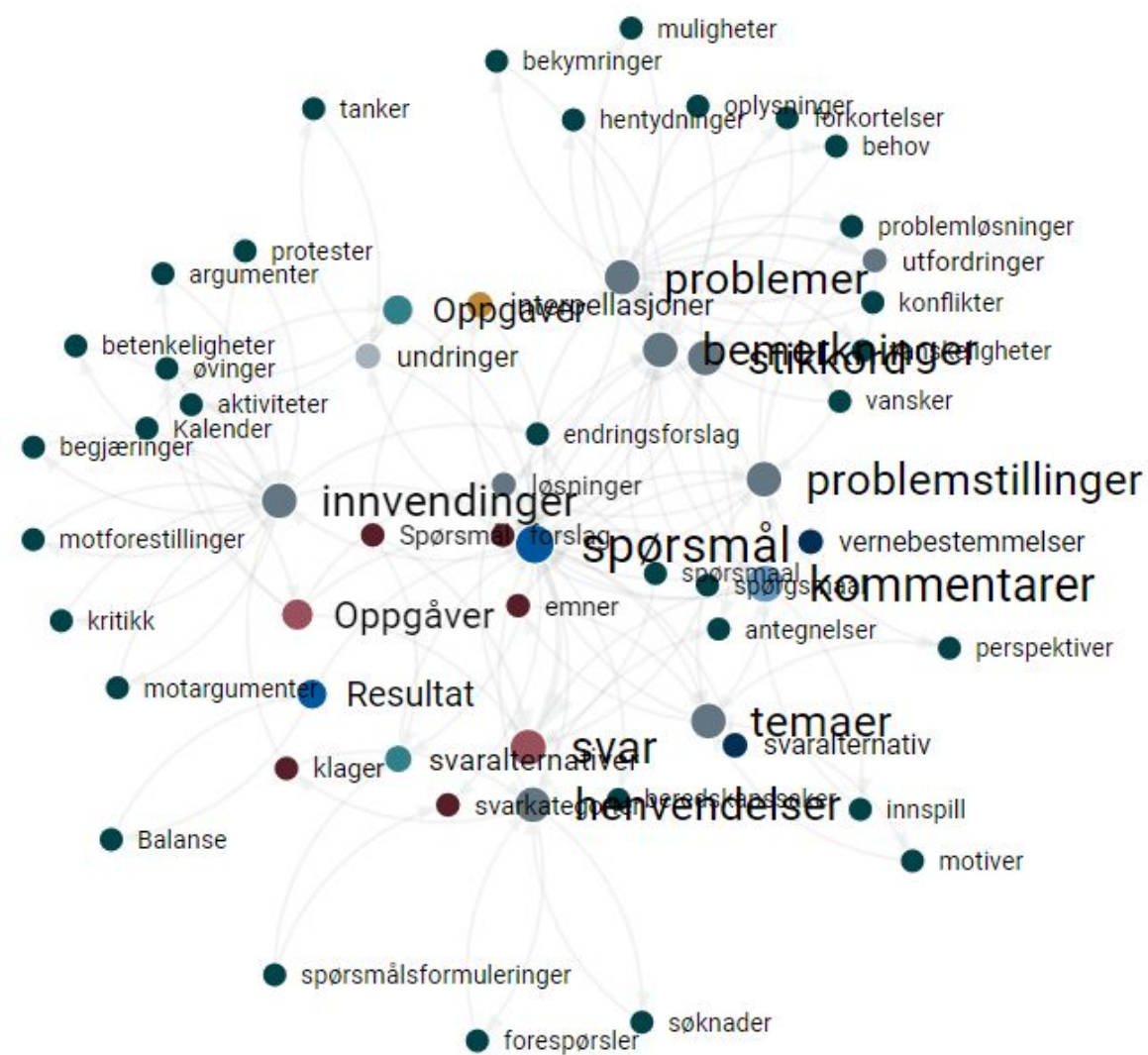
Evaluering

- sentimentanalyse
- rettferdighet og sannferdighet
- leseforståelse
- kunnskap om verden
- resonneringsevne
- norsk syntaks
- sammendrag
- oversettelse
- variasjon og lesbarhet

Aggregated Scores per Model Skill

- Sentiment Analysis
- Fairness & Truthfulness
- Reading Comprehension
- World Knowledge
- Commonsense Reasoning
- Norwegian Language
- Summarization
- Translation
- Variation & Readability





nb.no/sprakbanken



Nasjonalbiblioteket