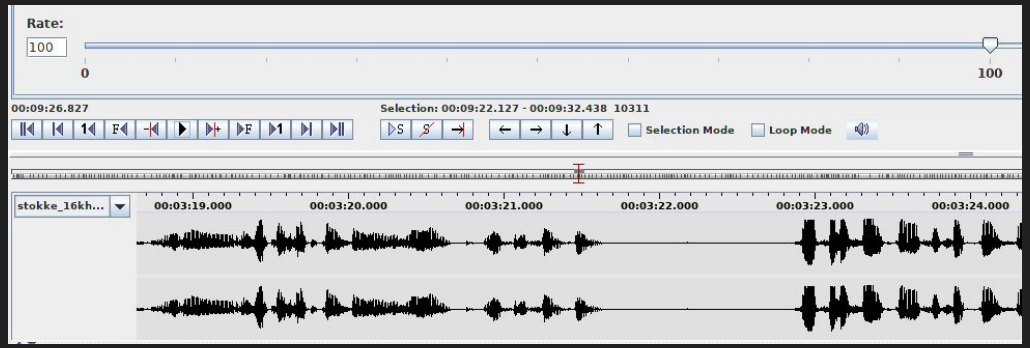
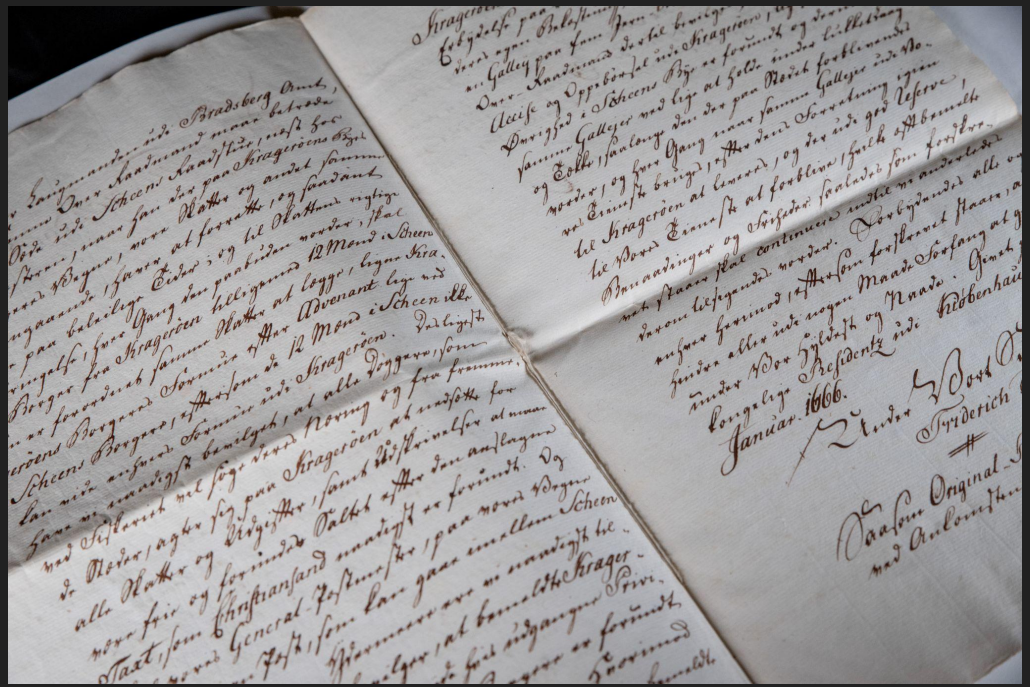




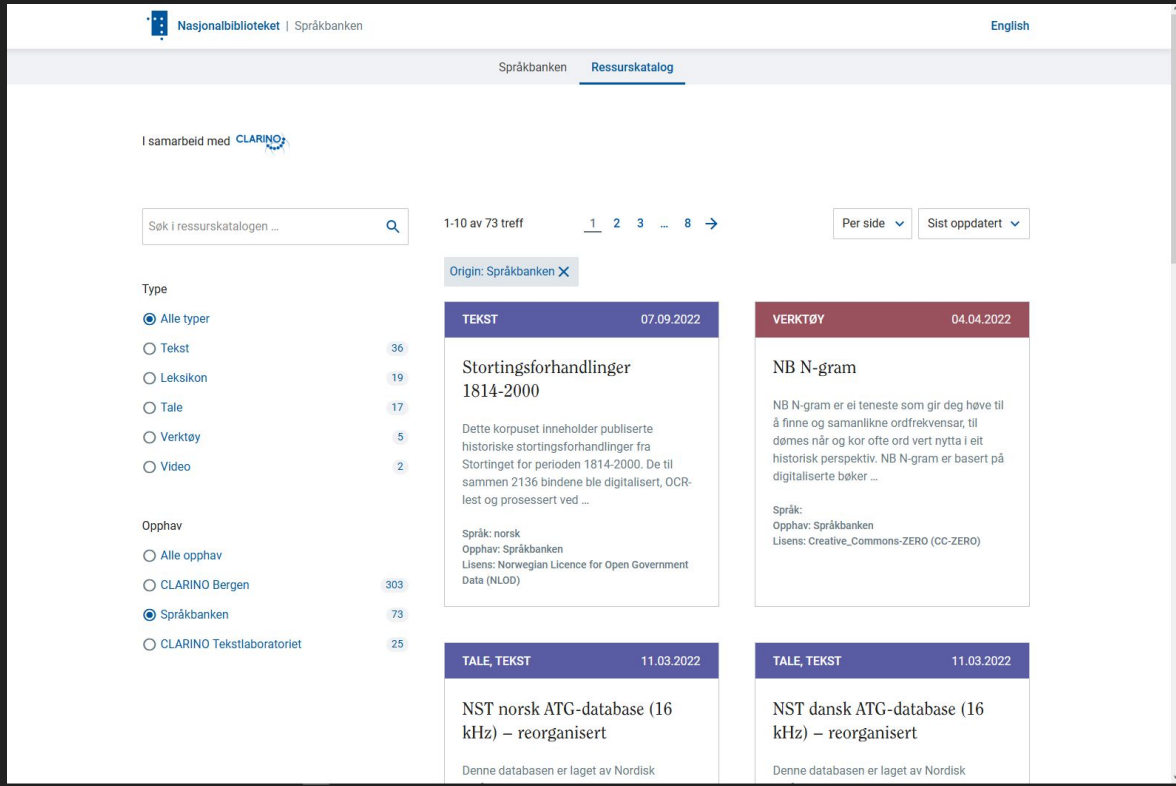
Tilgang til data: Språkteknologi og Nasjonalbiblioteket

Magnus Breder Birkenes
Språkbanken

Nasjonalbiblioteket

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



Internasjonalt samarbeid

Nasjonalbiblioteket deltar i de sentrale europeiske forskningsinfrastrukturene innenfor språkteknologi:

- CLARIN
- DARIAH

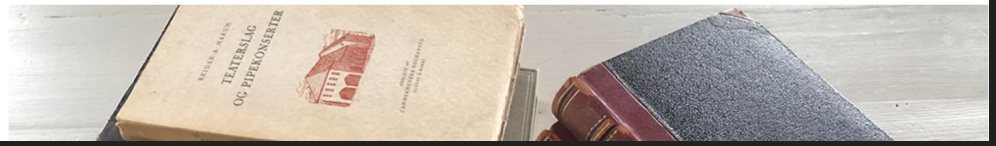
Forskningsprosjekter

QUEERDOM

CLARINO

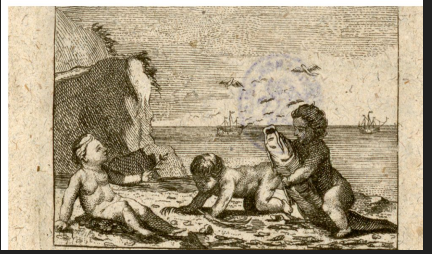
SCRIBE
Machine transcription of
Norwegian conversational speech

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



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

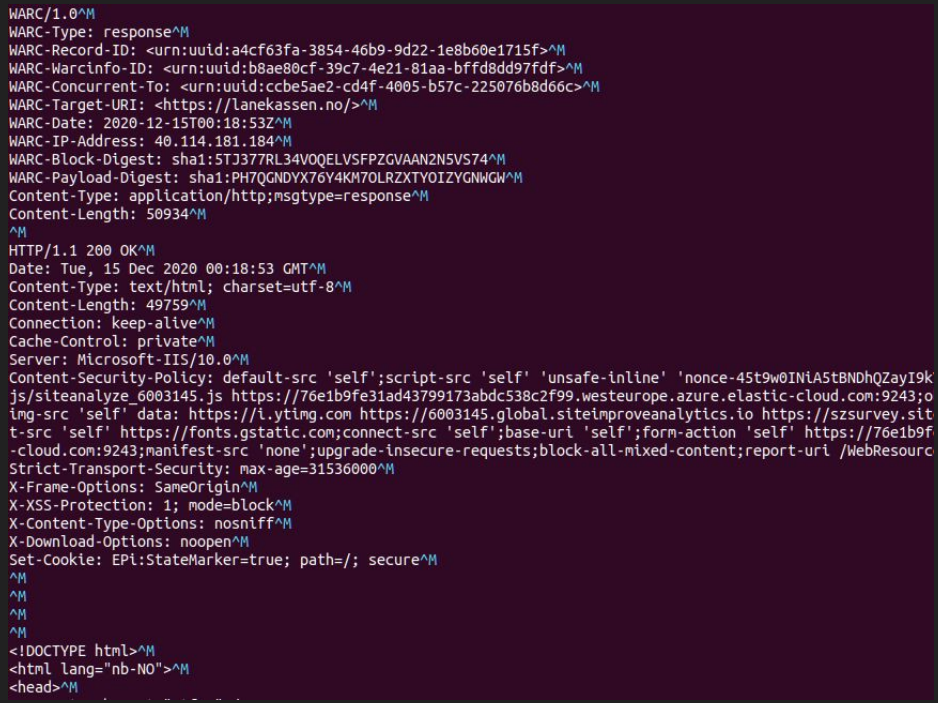
Bygging av språkressurser

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

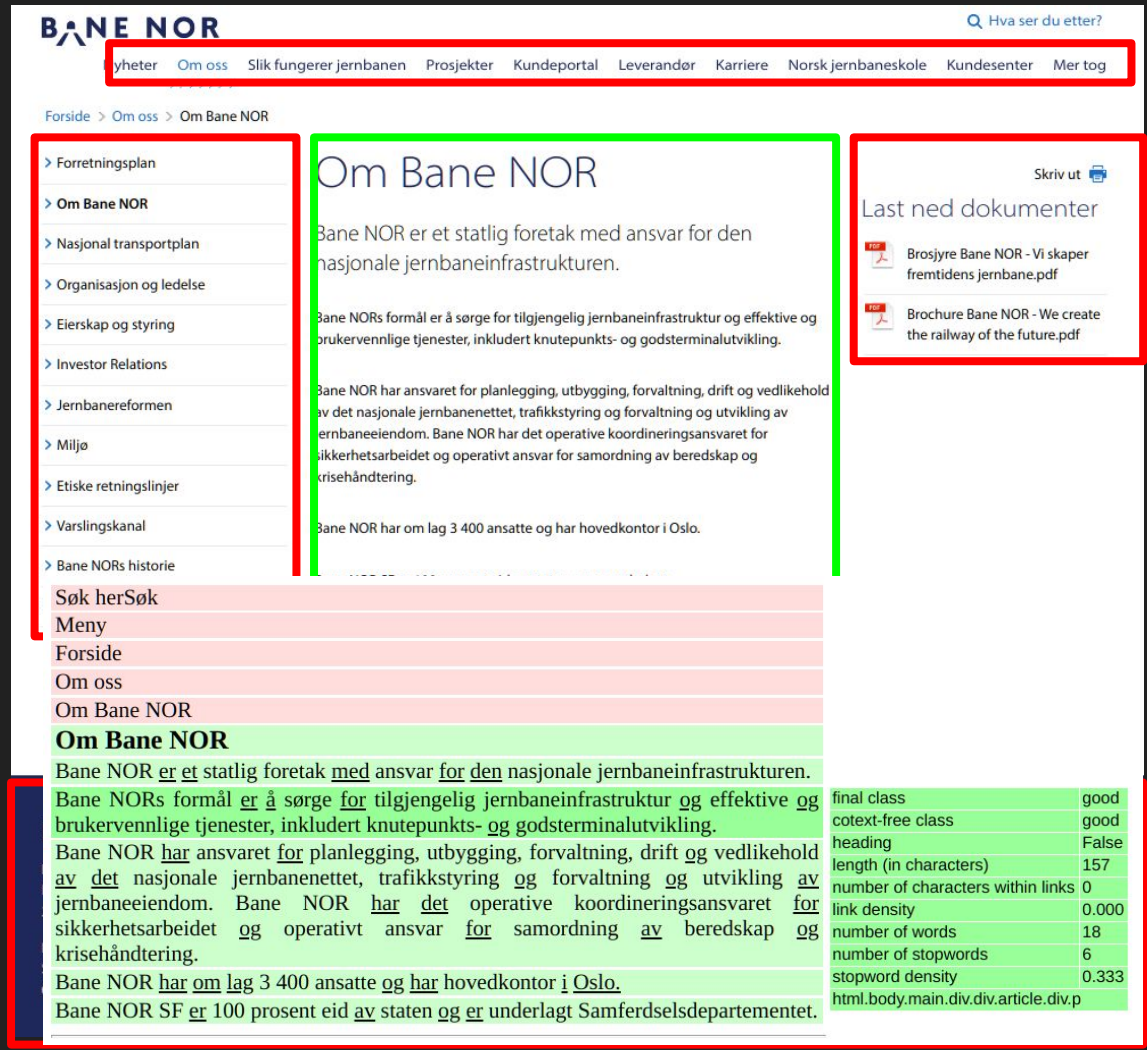
Høsting

- 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



Prosessering

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



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 utdanningsstøtteloven. 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

Trenings-data talegjen-kjenning

Testsett talegjen-kjenning

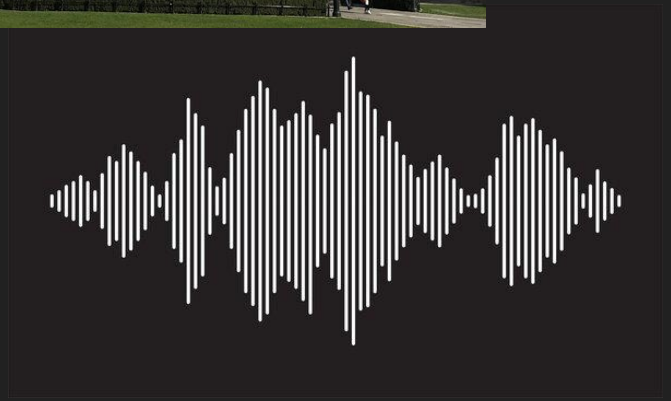
Stortings-korpuset+

Stortingskorpuset

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

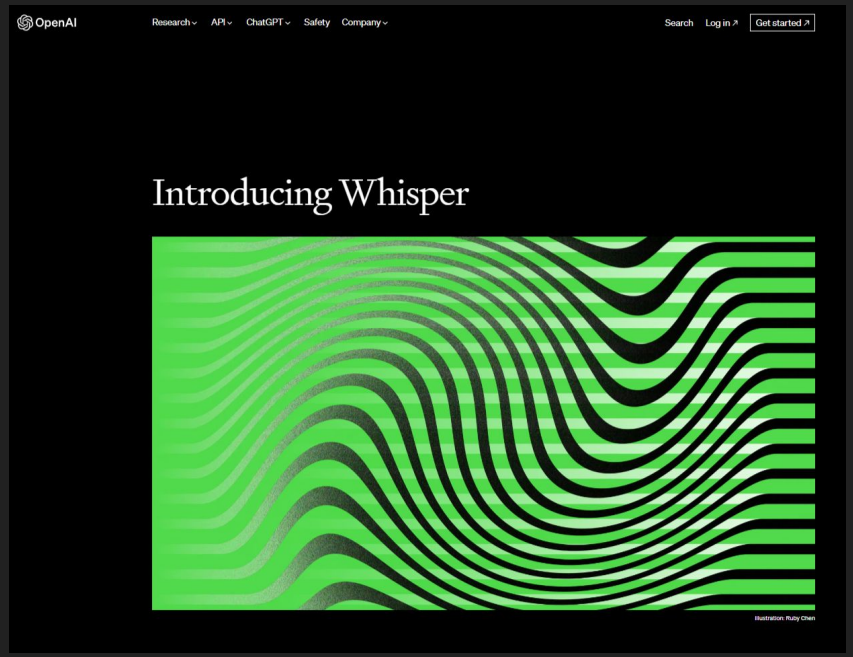
Stortingskorpuset+

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



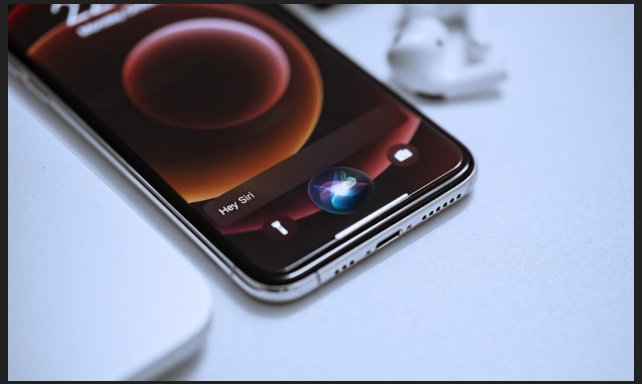
NB Whisper

- 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

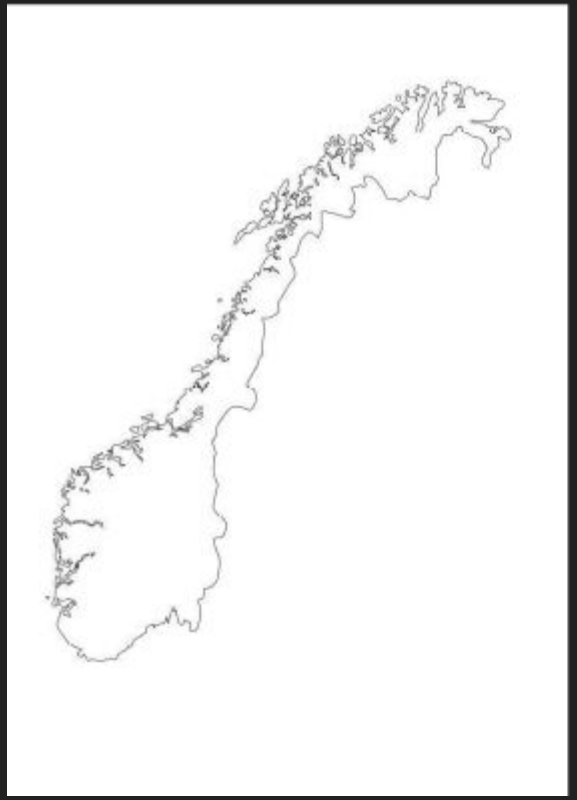
- SCRIBE
- Uttaleleksikon bokmål
- 800 000 ord
- 30 000 nyord
- Regulære uttrykk
- Nofabet

leser, L EE1 S AX0 R → L AE1 S
sover, S OA1 V AX0 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



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

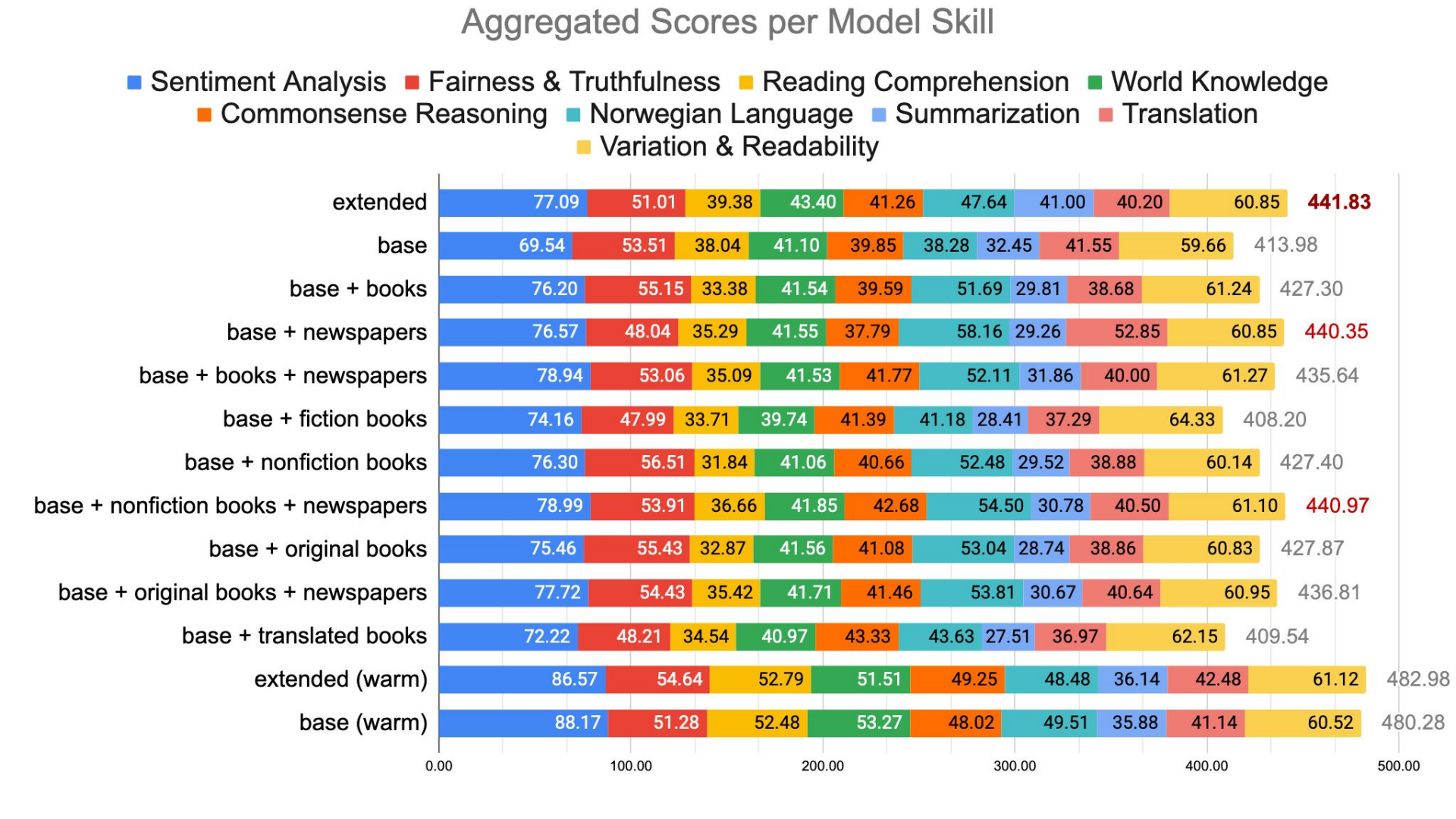
Korpus

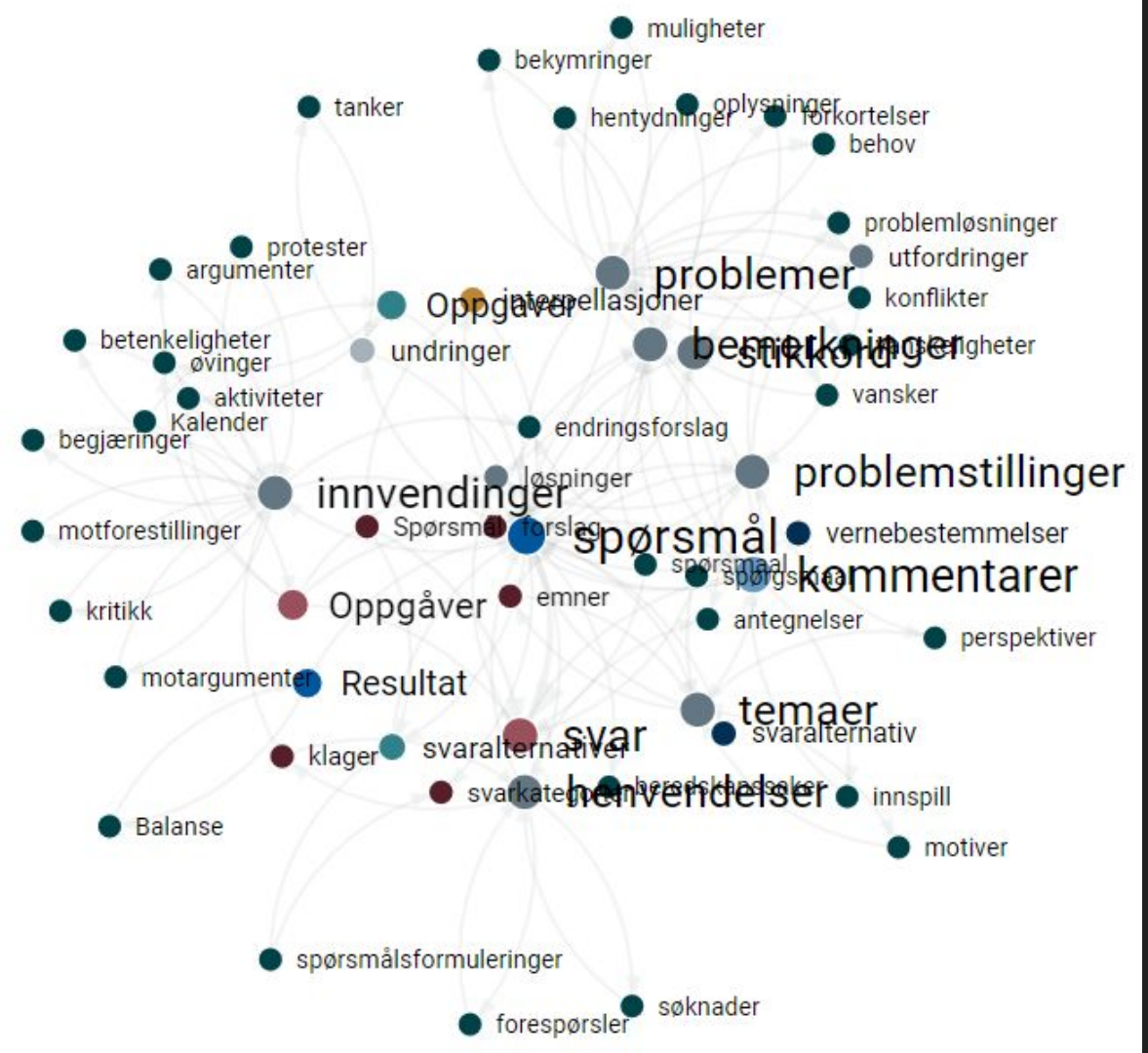
- mimir base
  - åpent datasett for fri tilgang uten begrensninger
  - aviser og bøker i det fri eller etter avtale
  - publikasjoner fra det offentlige
  - tilrettelagte data fra Språkbanken, f.eks. fra Web (Målfrid)
- mimir extended
  - internt datasett
  - inneholder mimir-base og i tillegg bøker 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
extended 77.09 51.01 39.38 43.40 41.26 47.64 41.00 40.20 60.85 441.83
base 69.54 53.51 38.04 41.10 39.85 38.28 32.45 41.55 59.66 413.98
base + books 76.20 55.15 33.38 41.54 39.59 51.69 29.81 38.68 61.24 427.30
base + newspapers 76.57 48.04 35.29 41.55 37.79 58.16 29.26 52.85 60.85 440.35
base + books + newspapers 78.94 53.06 35.09 41.53 41.77 52.11 31.86 40.00 61.27 435.64
base + fiction books 74.16 47.99 33.71 39.74 41.39 41.18 28.41 37.29 64.33 408.20
base + nonfiction books 76.30 56.51 31.84 41.06 40.66 52.48 29.52 38.88 60.14 427.40
base + nonfiction books + newspapers 78.99 53.91 36.66 41.85 42.68 54.50 30.78 40.50 61.10 440.97
base + original books 75.46 55.43 32.87 41.56 41.08 53.04 28.74 38.86 60.83 427.87
base + original books + newspapers 77.72 54.43 35.42 41.71 41.46 53.81 30.67 40.64 60.95 436.81
base + translated books 72.22 48.21 34.54 40.97 43.33 43.63 27.51 36.97 62.15 409.54
extended (warm) 86.57 54.64 52.79 51.51 49.25 48.48 36.14 42.48 61.12 482.98
base (warm) 88.17 51.28 52.48 53.27 48.02 49.51 35.88 41.14 60.52 480.28
0.00
100.00
200.00
300.00
400.00
500.00



nb.no/sprakbanken

Nasjonalbiblioteket