UNIVERSITETET I BERGEN

Universitetsbiblioteket

Teknisk og fagleg revisjon av Nynorskkorpuset ved Språksamlingane

Peder Gammeltoft, Rune Kyrkjebø og Paul Meurer

3. februar 2021

Nynorskkorpuset som del av Språksamlingane

- Dei nordiske språka er under press frå engelsk
- Nødvendig å sikre norsk som sjølvstendig språk gjennom tilgengeleggjering av språkressurer
  - Norge har ei utfordring med to målformer
  - Nynorsk har ei stor utfordring med færre språkressursar
- Oppgåve for Språksamlingane å prøve å betre dette
  - Fokus på nynorsk, bl.a. korpus som ein sentral nynorskressurs
  - Korpus i heile perioden for nynorsk – ca. 150 år!
  - Dokumentasjon av nynorsk frå skjønlitteratur, religionsutøving, aviser, lærebøker og tidsskrift m.m.

Organiseringa av Nynorskkorpuset

- Språksamlingane
  - Administrativ handsaming
    - (Rune kjem tilbake til det om litt)
  - Teknisk drift
    - Samarbeid mellom Universitetsbiblioteket i Bergen og IT-avdelinga ved UiB
    - Står for korpusimport/-eksport

Organiseringa og drifta av Språksamlingane

- Språksamlingane
  - Mål: Nasjonal ressurs for norsk språk, moderne og historisk
  - Innhald:
    - **Leksikografi**
      - *Nynorskorpus*
    - **Namnegransking**
    - **Terminologi**
    - Norrønt
    - Målføre
  - Årleg ramme: 6 millioner p.a. + 4 millioner, Termportalen

- Leiing, overordna
  - Styringsgruppe, relevante UiB interessentar
  - Fagråd, fagleg og forskningsmessig forankring i heile landet

Administrativ handsaming av Nynorskkorpuset (Rune)

- Korpuset var høgt på lista over samlingsressursane ved ILN, UiO (hausten 2014)
- Teknisk overføring frå UiO til UiB hausten 2017
  - no2014.uio.no var sett opp som no2014.uib.no
- Avtalematerialet frå prosjektet Norsk Ordbok sitt papirarkiv om korpuset overført 2016 til UiB
  - Avtalane med teksteigarane
  - Redaktørarkiv m.m.
  - Kjeldebibliotek
- Overføringa var del av heile samlingsflyttinga
  - Ingen særskild avtale mellom UiO og UiB om korpuset
- Leksikografisk Bokmålskorpus er ikkje overført til Språksamlingane

Avtalane mellom teksteigarar og UiO/Prosjektet Norsk Ordbok 2014

- Nokre hovudpunkt:
  - Mål: fullføre den leksikografiske kartlegginga og skildringa av ordtilfanget i dei norske dialektane og det nynorske skriftmålet
  - Korpuset skal berre brukast til forskingsformål
  - Tilslag i korpuset er fritt tilgjengelege i kontekst med lengd innanfor sitatretten
  - Tekstane inngår i korpuset på ubestemd tid eller permanent
- UiB har drive korpuset vidare på desse vilkåra
- Det er gjort førearbeid for ein ny runde med avtalar mellom UiB og alle teksteigarar

Uttrekket av tekst nyare enn år 2000 – UiB gjekk inn som avtalepart

- Språkrådet leia ein ny avtalerunde som galdt eit uttrekk av nyare tekst i omkalfatra form til fri deling
  - Hovudmål: språkteknologisk bruk, også kommersiell
  - Denne runden vart gjennomført 2017, 2018 med UiB *de facto* gått inn i staden for UiO i avtalane med teksteigarane
  - Alle slik delte tekstar har ny avtale
- Samlaget og UiB har ny avtale om tillegg av tekst
- Allkunne.no og UiB har ny avtale om tillegg av tekst

Clarino og verktøya INESS og Corpuscle

- Nynorskkorpuset og Leksikografisk Bokmålskorpus i Clarino
  - Deponering av begge i Clarino-verktøyet INESS var gjort av ILN, UiO i 2015 i avtale med LLE, UiB
- Vilkår
  - For Nynorskkorpuset er det originale tekstmaterialet tilgjengeleg for INESS si forsking
  - Søkjeresultat for Nynorskkorpuset kan gjerast tilgjengelege via portal eller nedlastbare i forskingsformat, ikkje leseformat
  - Søkjeresultat for Leksikografisk Bokmålskorpus kan gjerast tilgjengelege som analyse av enkeltsetningar

Vidare arbeid

- Behovet for korpusarbeid var i praksis mogeleg å møte for UiB med hjelp av Clarino-ressursane og den tilhøyrande teknisk/lingvistiske kompetansen

- Aktualisert av Revisjonsprosjektet og prosjektet NO-AH

- Vidare drift og utvikling blir i UB og IT-avdelinga sitt tekniske fagmiljø som òg arbeider teknisk med andre prosjekt, blant andre:
  - Clarino
  - SAMLA
  - Menota, Wittgensteinarkivet, Ludvig Holbergs skrifter

Teknisk inntak av nye tekstar i Nynorskkorpuset (Paul)

Nynorskkorpuset

Nynorskkorpuset hittil:

- 107 mil. ord (inkl. tegnsetting)
- opprinnelig grammatisk tagget med Daniel Ridings' Brill-tagger (tekstene fra etter 1938)
- søkbar på websiden til Norsk ordbok og i Oracle/Delphi

Etter flytting til Bergen:

- retagget med Oslo-Bergen-taggeren
- søkbar i verktøyet Korpuskel(-Leks)
- utbygging med **nye tekster**

Nynorskkorpuset: inntak av nye tekster

Samlaget leverer sine nyutgivelser hvert år for innlemming i NNK.

år	antall tekster	importert
2016	122	
2017	83	64
2018	118	
2019	91	

Format: **PDF**

- *Fordel*: tekstene er tilnærmet feilfrie, ingen OCR-feil!
- *Ulempe*: vanskelig å konvertere til ren tekst med basal markup, hver tekst har sine særegenheter

Nye tekster: Preprosessering

Resultatet av preprosesseringen skal være:

- ren tekst med markdown
- delt inn i avsnitt (og setninger)
- koding av grovstrukturen (overskrifter, diktvers, mm.)

Halvautomatisk prosess som bruker regulære uttrykk for gjenkjenning av strukturelle elementer som må omformes eller skal fjernes

Første steg:

- Konvertering av PDF-dokumentet til ren tekst med Linux-programmet **pdftotext**

løfta til å komme nest etter kongen i rang – ein parallell til 1. Moseboks forteljing om Josef i Egypt.

Kong Xerxes kan ifølgje persisk lov ikkje trekke attende dekretet han har sendt ut. Derimot sender han ut eit nytt dekret, der han gir ordre om

^L at alle åtaka skal skje ein spesifikk dag, og at jødane i kvar by har lov til å stå saman og forsvare livet sitt. «Dei kunne utrydda, drepa og gjere ende på kvar væpna styrke i alle folk og provinsar som gjekk til åtak på dei, også barn og kvinner.»229

ESTER-BOKA HAR trass det alvorlige emnet ein del feststemte detaljar. Den første festen som kong Xerxes held for stormennene og provinsleiarane sine, varer i 180

Dette er for bibellesarar i dag ein brutal forsvarstanke. Men dekretet set mot i jødane, og når dagen for angrepet kjem, er det ingen som kan stå seg imot dei: «Jødane slo alle fiendane sine med sverd; dei drap dei og gjorde ende på dei. Dei gjorde som dei ville med alle som hata dei.»230 Reformatoren Martin Luther likte ikkje at

meter til mål. Frå augekroken såg han at keepe-
ren stod litt for langt ute frå målet. Ein tanke
skaut igjennom hovudet hans: han måtte sende
ballen i ein boge. Men i same sekund ombestemte
han seg. Trenaren hadde sagt fleire gonger at han
ikkje måtte prøve å gjere alt så vakkert heile tida.
Han måtte vere meir målbevisst og bruke kvar
einaste sjanse han fekk, til å skyte på mål.

Stein trippa kvikt for å få foten ved sida av bal-
len og skaut så hardt han kunne mot mål. Ein
strak og låg ball flaug av garde og trefte bakken
fem meter før mål. Det våte graset gav ballen ekstra

fart. Keeperen var sjanselaus. Ballen susa forbi
stonga og inn i mål.

Nye tekster: Preprosessering

Videre prosess*:*

- Fjerning av topp- og bunntekst
- Tegnkonvertering (spesialkoding med høye unicode-verdier; ligaturer fl, fi, anførselstegn)
- Retting av margen
- Fjerning av tabeller, bildetekster og annet rusk
- Fjerning av fotnoter (både nummer og fotnoteteksten)
- *Eller:* flytte fotnoteteksten til slutten av teksten (nyttig når det er lange fotnoter som ikke bør forkastes)
- Heuristikker for gjenkjenning av avsnitt, sidetall, overskrifter og løpeoverskrifter

Nye tekster: Preprosessering

Regulære uttrykk for alt dette når det er avvik fra heuristikken

Eksempel:

```
(merge-lines "projects:samlaget;2017;Viljen til liv.txt"
    :title-regex "KAPITTEL|DEL|^^L*(\\w|\\s){5,65}$"
    :footnote-regex "^[1-9][0-9]?[0-9]? [A-ZØÆÅ].{20}"
    :title-line-2 t
    :indent-is-para nil)
```

Nye tekster: Preprosessering

Sette sammen delte ord (ved linjeskift)

Regler:

- Bindestrek beholdes foran storbokstav (Karl-Ove) og bak tall (1970-åra)
- Bindestrek får mellomrom etter seg i koordinerte fraser: - og, - eller, mm.
- Ellers slettes bindestrek

Nye tekster: Preprosessering

Uhåndterlige problemer:

- Tospaltesats
- Sperret tekst

  No går bølgjene høge, og frå T å r e s j ø e n høyrest hul-

Feil i PDF-konverteringen:

- Manglende ligaturer
- Manglende tall

Nye tekster: Import i korpuset

Språkmarkering

Rundt 1,5 % av setningene er bokmål og andre språk.

Foreløpig setter jeg «**nno**» på nynorsksetninger og «**nob**» på alt annet, men dette må raffineres.

Splitting i setninger

Morfologisk tagging (lemma, morfosyntaktiske trekk) med OBT

Nye tekster: Inkorporering i korpuset

Nytt korpus med tekstene fra 2017, 3.5 mil. ord. (De andre årgangene er under bearbeiding.)

Indekserte attributter (med forbedrete søkemuligheter):

- ord, lemma, morfologi
- referansekode (som peker til bibliografilisten)
- sidetall
- målform
- Utgivelsesdato eller -år
- sjanger

  skjønnlitteratur prosa, litteratur for born og ungdom, dikt (med bunde versemål), folkeminne, saklitteratur, faglitteratur om (serl norsk) språk, …

- Planlagt: tittel, forfatter, oversetter, deres kjønn (slik det er gjort i trebankene)

Veien videre

På sikt skal også hovedkorpuset få samme fasong (utvidet sett med attributter)

Annet materiale som etter hvert skal/bør inkluderes:

- Allkunne.no, NRK?, Atekst?
- Mer?

En **nyimplementering** av korpussystemets brukergrensesnitt er på vei (i regi av Clarino+ og Revisjonsprosjektet) med veldefinert API og klart skille mellom korpustjener og brukergrensesnitt

Ønskelig:

- forbedring av nynorsk-delen av Oslo-Bergen-taggeren
- mulighet for tagging av eldre tekst

peder.gammeltoft@uib.no
rune.kyrkjebo@uib.no
paul.meurer@uib.no

Nynorskkorpuset
Universitetsbiblioteket
Språksamlingane

UNIVERSITETET I BERGEN
Universitetsbiblioteket