Innbyding til digitalt seminar om korpusutvikling

Nynorskkorpuset (NNK) som ressurs for språkforsking, ordbokarbeid og språkteknologi

Stad og tid: Zoom, onsdag 3. februar 2021 kl. 09.00 til 16.00

Høgskulen i Volda samarbeider med Universitetet i Bergen om prosjektet *Norsk ordbok a–h revidert (NO-AH)*, digitalisering og revisjon av alfabetstrekket a–h. Prosjektet vart sett i gang i 2019 med løyving over statsbudsjettet. Utvikling av korpus og språkteknologiske ressursar er viktige føresetnader for prosjektet. Nynorskkorpuset spelar her ei sentral rolle og er derfor utgangspunktet for seminaret.

Målet for seminaret er formulert slik:

- Drøfte korleis Nynorskkorpuset kan bli utvida, balansert og tilrettelagt som ein sentral ressurs for redigering av Norsk Ordbok, for andre språkvitskaplege studium og for utvikling av språkteknologi for nynorsk.
- Drøfte om og eventuelt korleis Nynorskkorpuset kan inngå som delressurs saman med korpusressursar for bokmål i utviklinga av eit felles nasjonalkorpus for norsk.

På seminaret tek vi sikte på å drøfte fagleg-tekniske sider som utviding, balansering, inndeling, tagging og forbetra søkjemoglegheiter for Nynorskkorpuset, også sett i lys av korpusressursar for bokmål. Vi ønskjer å få synspunkt og behovsanalysar på vidareutvikling og bruk sett frå framståande forskarar på feltet.

Alle som er interesserte i dette forskingsfeltet, er hjarteleg velkomne til å ta del på heile eller delar av seminaret!

Program

09.00–09.15 Opning og innleiing

- Rektor Johann Roppen
- Komitéen: Johan Myking og Stig Helset

09.15–10.00

Oddrun Grønvik (UiB/NO-AH) og Christian-Emil Smith Ore (Universitetet i Oslo):

Nynorskkorpuset – historikk og status

Nynorskkorpuset vart skapt for å vera kjelde til vitskapleg, diakron og synkron, undersøking av ordtilfanget i nynorsk, primært for arbeidet med Norsk Ordbok (NO2014-prosjektet). Det vidare føremålet er å vera ei kunnskapskjelde for ålmenta. Nynorskkorpuset er difor fritt tilgjengeleg og søkbart for alle, og har sidan oppstarten i 2003 vore nytta som resurs i

språkforsking, ordbokarbeid og språkteknologi, men og i mange andre samanhengar. Nynorsk-korpuset er no på om lag 105 mill tokens, og femner vidt i tid, sjanger og stil. Normalspråkleg tekst har vore prioritert, unnateke tida før 1940 . Meir enn 85 % av teksten er frå perioden 1975 - 2015. Nynorskkorpuset er eit monitorkorpus, og det ligg tekstreservar i Språksamlingane som etter tidlegare plan skal inn i Nynorskkorpuset. Dette innlegget vil presentera Nynorskkorpuset slik det er i høve til innhald og funksjonalitet. Spesielt vil vi sjå på samansetning og strukturelle prinsipp, bibliografiske metadata og koplinga mot Metaordboka.

10.00–10.30

Margunn Rauset og Gyri Smørdal Losnegaard (UiB/Revisjonsprosjektet og NO-AH):

Nynorskkorpuset og utviklingsbehov for leksikografisk arbeid

I innlegget kjem vi til å sjå på Nynorskkorpuset i samanheng med dei andre ressursane i korpussamlinga Korpuskel-leks. Dette verktøyet let oss søkje i inntil tolv korpus samstundes, og det blei implementert i 2018 for å stette dei leksikografiske behova ved UiB. I dag inneheld nynorskdelen av Korpuskel-leks rundt 180 millionar ord, medan bokmålsdelen inneheld 2,6 milliardar ord. Så kva gjer vi med at nynorskressursane utgjer under 6,5 % av det samla korpusmaterialet? Skal omsynet til balanse mellom ulike sjangrar og ulike historiske periodar vege mykje tyngre enn det overordna behovet for meir nynorskmateriale? Kan det vere ulike leksikografiske behov for Revisjonsprosjektet og NO-AH? Er det betre å leggje til fleire spesialressursar som netthausta korpus (à la noWaC og HaBiT), SoMe-korpus, elevtekstkorpus osv., og la det vere opp til å redaktørane og ordbokprosjekta å vurdere kva korpus ein søkjer i i Korpuskel-leks, heller enn å byggje Nynorskkorpuset størst mogleg?

10.30–10.45 Pause

10.45–11.15

Peder Gammeltoft, Rune Kyrkjebø og Paul Meurer (UiB/Språksamlingane):

Teknisk og fagleg revisjon av Nynorskkorpuset ved Språksamlingane

Språksamlingane er ein nasjonal ressurs for dokumentasjonen av norsk språk, historisk og moderne. Nynorskkorpuset er svært viktig for Språksamlingane som ein dokumentasjon av bruken av nynorsk over dei siste 150 åra i skjønlitteratur og religionsutøving, aviser, lærebøker og tidsskrift m.m. Universitetsbiblioteket i Bergen i samarbeid med IT-avdelinga ved Universitetet i Bergen har i dag den administrative og tekniske drifta av nynorskkorpuset, og står for import til og eksport frå korpuset. Etter ein innleiande merknad om den organiseringa vil innlegget vårt gå inn på korleis tekstar blir handsama og tekne inn i nynorskkorpuset og i trebankinfrastrukturen INESS, med nye tekstar frå Samlaget som døme. Det dreiar seg om digitalt fødde tekstar som Universitetsbiblioteket mottek i PDF-format, som blir konverterte slik at ein til slutt har enkeltsetningar i reint tekstformat. Dei konverterte dokumenta er eit godt utgangspunkt for parsing i Oslo-Bergen-taggaren og så import i nynorskkorpuset. Dei er også eit godt utgangspunkt for syntaktisk parsing med NorGram og import i INESS som trebank.

11.15–11.45

Victoria Rosén (UiB/INESS):

Søk i NorGramBank for leksikografiske formål

Ved Universitetet i Bergen har infrastrukturprosjektet INESS i løpet av det siste tiåret utviklet det største syntaktisk annoterte korpuset for norsk, NorGramBank, samt et grensesnitt for søk. NorGramBank er en viktig informasjonskilde for leksikografisk arbeid. Det er enkelt å få frem bruksmønstre for ulike ord; for eksempel kan man finne alle argumentrammer for et bestemt verb, med frekvenser for de ulike rammene, sammen med korpuseksemplene. Innlegget vil gi en innføring i søkemulighetene av interesse for leksikografisk arbeid.

11.45–12.45: Lunsj

12.45–13.15

Per Magnus Finnanger Sandsmark (Nynorsk kultursentrum):

Allkunne i tal og prinsipp

Den 1. januar 2021 tek Vestland fylkeskommune over Fylkesleksikon for Sogn og Fjordane og Store norske leksikon over ansvaret for å drive leksikon på nynorsk. Over 10 år har Nynorsk kultursentrum leia arbeidet med å etablere eit moderne nynorsk oppslagsverk på internett: Allkunne. I innlegget vert innhaldet presentert med utgangspunkt i mengd, tema og opphav. Hovuddelane i oppslagsverket Allkunne ved omleggingstidspunktet var 15 251 redigerte og omsette artiklar frå Caplex, 6 984 artiklar frå Fylkesleksikon for Sogn og Fjordane, stort sett utarbeidd av NRK Sogn og Fjordane, om lag 2 500 nyskrivne artiklar for oppslagsverket og 1 500 artiklar frå eksisterande verk. Redigeringsprinsipp, språkval, temabreidde og bruk vert presentert for kvar kategori.

13.15–13.45

Dag Trygve Truslew Haug (Universitetet i Oslo/Tekstlaboratoriet):

Korpusressurser for bokmål

I dette innlegget vil jeg presentere bokmålsressursene på Tekstlab, særlig Leksikografisk Bokmålskorpus, NoWaC og Bokselskap-korpuset. Jeg vil gå gjennom tekstutvalgene som er representert, tilgjengelig annotasjon og metadata, søkemuligheter og lisenssituasjonen for korpusene.

13.45–14.00: Pause

14.00–14.30

Lars G. Johnsen (Nasjonalbiblioteket/Språkbanken):

Norske korpus som ressurs i språkteknologi for norsk språk

Med utgangspunkt i digitaliserte tekster fra Nasjonalbiblioteket og data fra Norsk Ordbank, vil jeg se på hvordan språklige register og variasjon kan systematiseres og kvantifiseres. Registre kan benyttes i språkgjenkjenning, for eksempel, å skille mellom nynorsk og bokmål, som et alternativ til n-grammer basert på bokstaver. Den kvantitative tilnærmingen gir også en

mulighet til å snakke om hvor nær forskjellige varianter er, innenfor og på tvers av målformer, i tillegg til å kunne gi et mål på det språklige spennet innenfor en målform - er variasjonen i nynorsk større enn i bokmål? Samtidig blir vi i stand til å si noe om utbredelsen av registre, og utvikling av under-normer innenfor de to målformene.

14.30–15.00
Tor Arne Haugen (Høgskulen i Volda):

Nynorskkorpuset og Leksikografisk bokmålskorpus som grunnlag for Norsk nasjonalkorpus bokmål og nynorsk?

For norsk har ein to leksikografiske skriftspråkskorpus som på mange måtar fungerer som referansekorpus for kvart sitt skriftspråk: *Leksikografisk bokmålskorpus* og *Nynorskkorpuset* knytt til *Norsk Ordbok*. Begge korpusa kom til som grunnlag for leksikografisk arbeid, men dei er svært viktige ressursar også for normering generelt og for språkvitskapleg forsking. Felles for begge korpusa er at dei så langt er tilført nytt tilfang fram til om lag 2013, og dei har dermed, i alle fall til ein viss grad, fungert som monitorkorpus for dei to skriftspråka. Spørsmålet er om tida no er inne for å samordne desse ressursane, gjere dei meir parallelle og utvikle eit norsk nasjonalkorpus etter felles prinsipp for begge skriftspråka.

15.00–16.00
Oppsummering og avsluttande diskusjon