UiO : Institutt for lingvistiske og nordiske studier
Det humanistiske fakultet

Korpusressurser for bokmål

Dag Haug

3. februar

tekstlab

Introduksjon

- Jeg er ny – både på Tekstlaben og i verdenen av norske korpora
- Jeg skal si litt om ressurser som mange av dere kjenner mye bedre enn meg
- Men kanskje et interessant utenfrablikk?

Korpora ved Tekstlaben

Ved siden av noen mindre, spesialiserte korpora (elevtekster, norsk som andrespråk, parallelle korpus) har Tekstlaben fire større korpora for bokmål:

- Bokselskap-korpuset
- Oslokorpuset
- Leksikografisk bokmålskorpus
- Norwegian Web as Corpus (NoWaC)

Glossa

(Så godt som) alle korpusene ved Tekstlaben tilgjengeliggjøres gjennom grensesnittet Glossa

- Enkel filtrering på metadata
- Grensesnitt på tre nivåer
  - "Google-aktig"
  - utvidet (med menyer for tagger, lemma o.l.)
  - direkte CQP-søk
- Innlogging og godkjenning av lisenser via forskjellige systemer (FEIDE, eduGAIN, CLARIN)
- Nedlasting av søkeresultater

Bokselskap-korpuset

bokselskap.no er en ebokportal utviklet av Det norske språk- og litteraturselskap, som publiserer norske tekster som har falt i det fri. Tekstene gjøres tilgjengelig for søk i Glossa.

- Bokmålsdelen har 148 tekster med 6679446 tokens
- Litterære tekster fra Petter Dass til Nordahl Grieg
- Under stadig utvidelse (ca. 50 nye tekster i året)
- Metadata: tittel, forfatter, målform, kjønn, sjanger, år for førsteutgave
- Ingen tagging
- Fritt tilgjengelig, også råtekster

Oslo-korpuset

Tekstlaboratoriets første korpus (1999), med skjønnlitteratur, avistekster og sakprosa. Tilgjengelig for søk i det opprinnelige grensesnittet.

- 18.5 millioner ord bokmål
  - 1.7 millioner ord skjønnlitteratur (fra daværende Norsk Tekstarkiv)
  - 9.6 millioner ord avis/ukeblad
  - 6.9 millioner ord sakprosa (lovtekster og NOU)
- Metadata: bare opplysninger om kilde
- Tagget med (en tidlig versjon av) Oslo-Bergen taggeren
- CLARIN ACA (Academic) End-User License +NC +LOC +ND 1.0, men noen av tekstene kan distribueres

Leksikografisk bokmålskorpus (LBK)

LBK er et representativt, vektet korpus laget for leksikalsk utforsking av moderne bokmål ved tidligere avdeling for bokmålsleksikografi.

- 100 millioner ord bokmål fra 1985-2013
  - 45% sakprosa
  - 35% skjønnlitteratur
  - 10% aviser/ukeblad
  - 5% tv-teksting
  - 5% upublisert materiale
- Rike metadata: forfatter, kjønn, alder, emne, oversatt (j/n), år, utgiver, tekstkategori
- Tagget med Oslo-Bergen taggeren
- CLARIN ACA (Academic) End-User License +NC +LOC +ND 1.0, og tekstene kan i følge UiO ikke videredistribueres uten ny avtale med rettighetshaverne

Norwegian Web as a Corpus

Tekstlaboratoriets største korpus bestående av dokumenter fra .no-domenet i perioden nov 2009-jan 2010, lastet ned med tillatelse fra Kulturdepartementet.

- 700 millioner ord bokmål
- Ingen metadata (pga. avtalen med Kulturdepartementet)
- Tagget med Oslo-Bergen taggeren
- Fritt søkbart i Glossa; tekstene kan også lastes ned for forskningsformål i skramblet form (randomisert rekkefølge på setningene)

Gjenbruk av kildedata

- Tekstlaboratoriet har flere ressurser som i prinsippet er aktuelle for et framtidig nasjonalkorpus
- Men lisenssituasjonen gjør det vanskelig
  - eksisterende avtaler legger mange begrensninger
  - mange rettighetshavere, stort arbeid å reforhandle
- Korpusdataene er ofte uløselig knytta til grensesnitt de tilgjengeliggjøres gjennom
- Grensesnittene gir svært god funksjonalitet for mange brukere, samt mulighet for å laste ned dataene for videre prosessering, men har begrensninger

Nye brukere, nye formål?

- Framtidige brukere blir mer avanserte
  - UiO (og sikkert mange andre) tilbyr kurs i statistikk og programmering myntet på lingvister/humanister
- For mange anvendelser er det gunstig å ha tilgang til rådataene
  - preprosessering for lingvistiske søk
  - forskjellige former for distribusjonell analyse (kollokasjoner etc.)
  - maskinlæring (f.eks. språkteknololgi; emnemodellering for litteraturvitere)
  - visualiseringer av dataene
- Særlig viktig ved mer avansert annotasjon (trebanker, koreferanser)
- For tida er det bare Bokselskap-korpuset og NoWaC som har tilgjengelige rådata (men kun for ikke-kommersielle formål og randomisert for NoWaC)