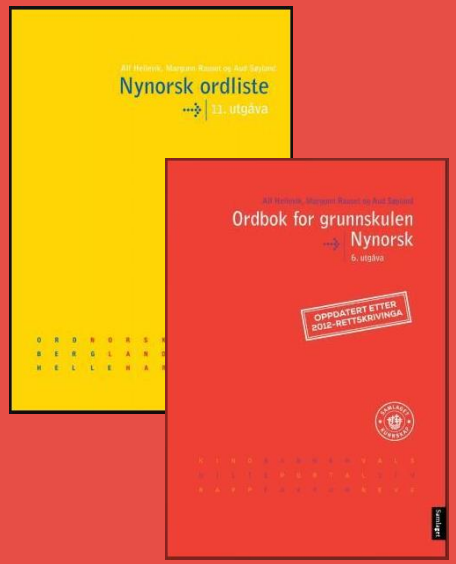
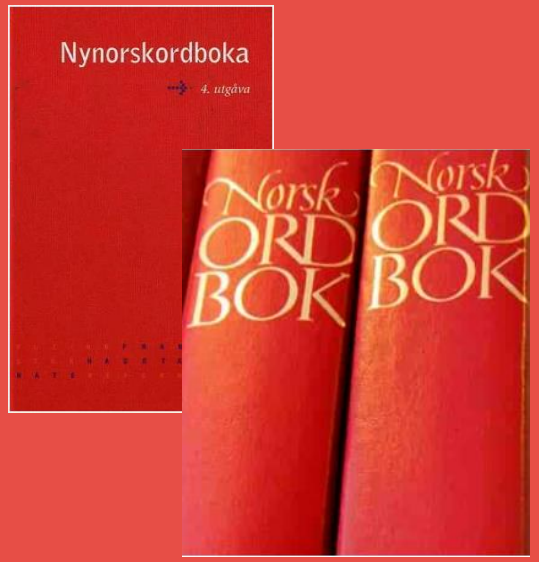
Nynorskkorpuset og utviklingsbehov for leksikografisk arbeid

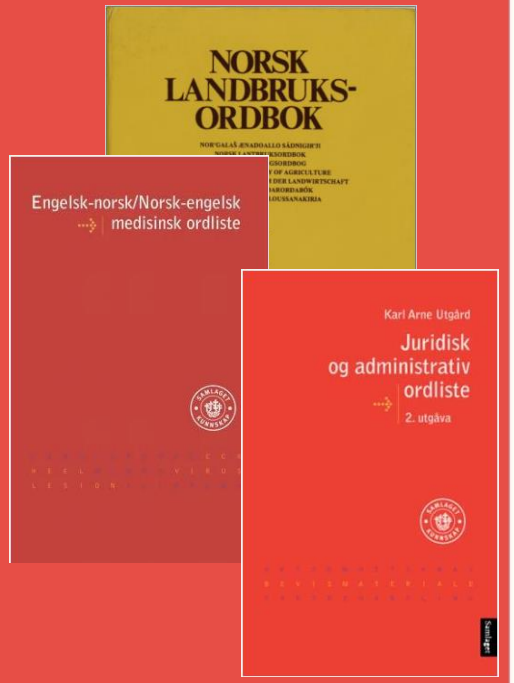
Margunn Rauset (Revisjonsprosjektet)
og Gyri Smørdal Losnegaard (NO-AH)

UNIVERSITETET I BERGEN

Nynorsk leksikografi er så mangt







Korpus	Skriftspråk	Mill. ord	Tidsrom	Sjanger	Søkjeattributt	Tilgang
Nynorskkorpus	nynorsk	107,8	1866–2012	tekst: blanda (sakprosa og skjønnlitteratur)	ordform, lemma, ordklasse	avgrensa
Nynorskkorpus (2017)	nynorsk	2,8	2017	tekst: blanda (sakprosa og skjønnlitteratur)	ordform, lemma, ordklasse	avgrensa
NBs frie tekstar (nn)	nynorsk	46,02	1850–2010	tekst: blanda (sakprosa og skjønnlitteratur)	ordform	open
Aviskorpus (nn)	nynorsk	21,01	1998–2020	tekst: sakprosa (avis)	ordform	open
Dialektendring	nynorsk	4,5	1960–2020	tale: dialektopptak	ordform, lemma, ordklasse	avgrensa
Industristad	nynorsk	1,78	1948–2013	tale: dialektopptak	ordform, lemma, ordklasse	avgrensa
Talesøk	nynorsk	1,56	ikkje spesifisert	tale: dialektopptak	ordform, lemma, ordklasse	avgrensa
Talk of Norway	bokmål, nynorsk	63,8	1999–2016	tekst: sakprosa	ordform, lemma, ord klasse	open
Forskning.no	bokmål, nynorsk	21,47	1998–2017	tekst: sakprosa (avis)	ordform, lemma, ordklasse	avgrensa
Leksikografisk bokmålskorpus	bokmål	102,29	1985–2013	tekst: blanda (sakprosa og skjønnlitteratur)	ordform, lemma, ord klasse	avgrensa
Aviskorpus (bm)	bokmål	1952,67	1998–2020	tekst: sakprosa (avis)	ordform	open
NBs frie tekster (bm)	bokmål	516,39	1765–2013	tekst: blanda (sakprosa og skjønnlitteratur)	ordform	open

Kalde fakta om Korpuskel leks

Ord i nynorske korpus:

- til saman 185 406 787
- 7,8 millionar av er desse er frå talemålskorpus
- utgjer 6,52 % av den samla korpusressursen

Ord i bokmålske korpus:

- til saman 2 656 622 701
- nokre få av desse 2,66 milliardane ord er på nn.
- utgjer 93,48 % av den samla korpusressursen

Til samalikning utgjer nynorskdelen av noTenTen 6,6 % i SketchEngine

Leksikografiske behov for NO-AH

Kva for artiklar finn ein ikkje i NO?

- Nyare ordtilfang i alfabetstrekket *a–h*
- Ord som ikkje svarer til redaksjonelle krav
- Lakuner som kjem av manglar i kjeldetilfanget

Norsk Ordbok byggjer på eit variert kjeldemateriale

- Norsk Ordbok har materialbasert ordutval *(Grønvik 2020)*
- Ein kan dele inn materialgrunnlaget i tre hovudtypar *(Vikør 2018, s. 31–33)*:
  - -prenta tekstar i bok- eller tidsskriftform
  - -setlar
  - -korpus

Ordutval

- Krav til ord som skal få artikkel i Norsk Ordbok:

  autentisitet, frekvens, stabilitet i form og bruk over tid,

  attkjenneleg ordstruktur,

  godt belagt i skriftlege og munnlege kjelder
- Ord som ikkje skal handsamast: *ordbokord, tilfeldige spontanlagingar, nokre typar samansetningar, namn*
- Utvalskriterie for einskilde typar av ord: *fagord, eksotika, enkelte lånord og nemningar...*

(Vikør 2018, Grønvik m.fl. 2016, Grønvik 2020)

Prenta bøker og tidsskrifter

Stort og representativt utval av den nynorske skriftproduksjonen
- hovudvekt på kanoniske forfattarskap og på original tekst (ikkje omsett)
- skjønlitteratur rikare representert enn sakprosa, eldre tid meir enn nyare tid

Skeivskap redaksjonen i dei siste åra arbeidde for å rette opp

Særskild undergruppe: ordbokskjeldene
- Aasen, Ross m.fl.
- ordbøker og -samlingar frå målføre
- eldre bygdemålsordsamlingar (frå før 1814)

Vikør (2018, s. 31-33)

Setlar

- Papirsetlar med opplysningar om ord: *skrivemåte, tyding og bruk, grammatisk informasjon, heimfesting, datering...*
- I det tidlege arbeidet med NO og før ein hadde digitale arbeidsmåtar og hjelpemiddel: setlar skrivne av medarbeidarar eller av redaktørane sjølve
- I nyare tid: elektroniske setlar, med ulike kjelder som grunnlag

Korpus

«I dag er et godt og dekkende korpus overtatt for seddelsamlingene og ansees for å være en betingelse for å kunne redigere en ordbok av høy kvalitet. **Seddelsamlinger gir informasjon om den partikulære, mens korpus om den generelle ordbruken.** Når en arbeider diakront eller med dialektorienterte ordbøker vil det ofte være nødvendig med både et korpus og en stor seddelsamling siden de kompletterer hverandre. NO2014 bruker derfor begge deler, og prosjektet har et korpus på 100 millioner ord.»

Notat av Christian-Emil Ore, vedlegg til Kvitbok om Norsk Ordbok (2012)

Behov for å få kartlagt kva som trengst for å få betre balanse

- Område som kan nemnast:
  - Eldre tekst
  - Informantbidrag (setlar og samlingar)

- Med oversyn over hol og skeivskap kan ein:
  - Ta særlege omsyn i det redaksjonelle arbeidet
  - Arbeide målretta med å tette hol i kjeldetilfanget (der dette lar seg gjere)

Leksikografiske behov for Revisjonsprosjektet

Kva treng Revisjonsprosjektet?

- *Bokmålsordboka* og *Nynorskordboka* skal vere «kjelde til kvalitetssikra og oppdatert informasjon om normert rettskriving og om vanleg språkbruk for skriftmåla nynorsk og bokmål.» (Samarbeidsavtalen mellom Språkrådet og UiB)
- stikkord: allmennspråk, oppdatert språkbruk, skriftspråka
- korpusbehov: at ressursane er store nok (som dei ikkje er i dag), og at vi har tilgang til dei i Korpuskel leks

Korleis kjem vi dit?

- Nynorskkorpuset må samarbeide med store tekstprodusentar (Allkunne, Samlaget, NRK?, Atekst?)
- Kan vi utfordre tradisjonen med redaksjonsstyrte medium?
- Nynorskdelen av Aviskorpuset må gripast fatt i.
- Kan nynorskprosenten bli høgare med betre skanningar i Nasjonalbiblioteket?
- Ser andre korpus verdien av å bli ein del av Korpusel leks-pakken, slik at dei blir ein del av grunnlaget for dokumentasjon av norsk?

Leksikografiske vs. allmenne behov

- Truleg ønskjer den gjengse språkbrukaren og språkvitaren eit så stort Nynorskkorpus som mogleg.
- Norske ordbokprosjekt har tradisjonelt bygt på publiserte tekstar som har vore gjennom ei form for kvalitetskontroll.
- Behov for konsistens på tvers av eit prosjekt og balanserte kjelder.
- Ulike ordbokprosjekt = ulike spesialkorpus i Korpuskel leks som ein kan velje å ta med eller ikkje?
  - netthausta korpus (à la noWaC, HaBiT og noTenTen17)
  - SoMe-korpus
  - elevtekstkorpus
  - lærebokkorpus osv.

Oppsummering

- Skal omsynet til balanse mellom ulike sjangrar og ulike historiske periodar vege tyngre enn det overordna behovet for meir nynorskmateriale?
- Kan det vere ulike leksikografiske behov for NO-AH og Revisjonsprosjektet?
- Er det betre å leggje til fleire spesialressursar 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?

Nynorskkorpuset kan ikkje gjere jobben åleine!

Kjelder

Grønvik, Oddrun (2020). *Materialbasert lemmaseleksjon – generelle omsyn og døme* (PPT og opptak). Undervisningsmateriale NO-AH/UiB, internt digitalt arkiv.

Grønvik, Oddrun, Helge Gundersen, Lars S. Vikør, Laurits Killingbergtrø og Dagfinn Worren (2016). *Redigeringshandbok for Norsk Ordbok 2014*.
http://no2014.uio.no/eNo/tekst/redigeringshandboka/redigeringshandboka.pdf

Ore, Christian-Emil. 2011. *Notat om leksikografi ved ILN etter 2015*. I Kvitbok om Norsk Ordbok etter 2014 (2012).
http://no2014.uib.no/eNo/tekst/kvitboker/kvitbok_for_prosjekt_post_2014.pdf

Vikør, Lars S. (2018). *Inn i Norsk Ordbok. Brukarrettleiing og dokumentasjon.* Det Norske Samlaget, Oslo.

uib.no