Norsk nasjonalkorpus for bokmål og nynorsk?

Tor Arne Haugen

Seminar om korpusutvikling, 3. feb. 2021

HØGSKULEN I VOLDA

To referansekorpus

- Leksikografisk bokmålskorpus
- Nynorskkorpuset
- Er det ønskeleg med meir einsarta referansekorpus av denne typen?
  - Norsk nasjonalkorpus for bokmål og nynorsk?

Disposisjon

- Ideen om nasjonalkorpus
- Nokre ønske frå eit brukarperspektiv
- Nokre spørsmål for vegen vidare

Ideen om nasjonalkorpus

- British National Corpus (BNC)
- 100 millionar ord
- Allment, balansert korpus over britisk engelsk
- Nasjonalkorpus = allment, balansert referansekorpus

Ideen om nasjonalkorpus

- British National Corpus
- Statisk, balansert samplingskorpus, representerer britisk engelsk på slutten av 1900-talet
- Subkorpus og variasjon

Ideen om nasjonalkorpus

From being a sample of the whole of language, the BNC was rapidly repositioned as a repository of language variety. This was in retrospect a sensible repositioning; a more divers collection of materials than the BNC is hard to imagine. Handling this diversity effectively however requires a clearer and better agreed taxonomy of text types than currently exists, and better access facilities for subcorpora. (Burnard 2002: 13)

- Både meir bruk og bruk av andre grupper enn ein hadde sett føre seg (Burnard 2002)

Nokre ønske frå eit brukarperspektiv

- Meir einsarta korpus som grunnlag for utforskinga av bokmål og nynorsk
- Felles brukargrensesnitt tilpassa ulike brukarar og behov
  - Er det mogleg å utnytte potensialet for bruk av korpus i språklæring?
- Betre grunnlag for kontrastive studium

Nokre ønske frå eit brukarperspektiv

- Rike moglegheiter for avgrensing gjennom metadata
  - Felles metadata sentralt for samanlikning
- Ein må vite nøyaktiv kva ein søkjer i
  - Storleiken på utval må gå tydeleg fram
  - Frekvensinformasjon
- Same moglegheiter for vidare prosessering og statistikk over søkeresultat
  - Moglegheiter for utrekning av kollokasjonsstyrke

HØGSKULEN I VOLDA

Vegen vidare?

- Er Leksikografisk bokmålskorpus og Nynorskkorpuset gode utgangspunkt, eller må ein byrje på nytt?
  - Svært ulike tidsspenn i materialet
  - Krevjande situasjon kring rettigheiter til tekstmateriale

- Balanserte, statiske samplingskorpus eller monitorkorpus?
  - Kan desse tilnærmingane kombinerast, som i COCA (Corpus of Contemporary American English)?

Vegen vidare?

- Felles prinsipp for teksttypar og vekting?
- Felles prinsipp for tilføring av tilfang?
- Harmonisering av tidsspenn på lengre sikt?
- Felles brukargrensesnitt og søkemoglegheiter?

HØGSKULEN I VOLDA

Sluttord

Meir einsarta referansekorpus = endå større verdi

Litteratur

Burnard, Lou. 2002. Where did we go wrong? A retrospective look at the British National Corpus. *Teaching and Learning by Doing Corpus Analysis*, 51–70. doi.org/10.1163/9789004334236_007 (8 November, 2019).

Knudsen, Rune Lain & Ruth E Vatvedt Fjeld. 2013. LBK2013: A balanced, annotated national corpus for Norwegian Bokmål. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013* (NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings), vol. 88, 12–20. Linköping. http://www.ep.liu.se/ecp/088/003/ecp1388003.pdf.