LexicoNordica-symposium 31

Hvilket datamateriale bygger nordiske ordbøger på? Skævheder, udfordringer og løsninger

15.-17. februar 2024
Voksenåsen, Oslo – www.voksenaasen.no

Temabeskrivelse og Abstracts

Hvilket datamateriale bygger nordiske ordbøger på?
Skævheder, udfordringer og løsninger
Temabeskrivelse og spørgsmål

LexicoNordica-redaktionen
Henrik Hovmark & Terje Svardal
Ømålsordbogen, Institut for Nordiske Studier og Sprogvidenskab, Københavns Universitet &
Språksamlingane, Universitetet i Bergen
hovmark@hum.ku.dk & terje.svardal@uib.no

Leksikografiske resurser spiller en central rolle for nordisk sprogforståelse både mellem talere af de forskellige sprog og for gensidig forståelse mellem forskellige etniske og sociale grupper. De tilgrundliggende tekstkorpusser kan imidlertid være skævt sammensatte, med for ringe repræsentation af fx talesprog, unges sprog og minoriteters sprog. Der er dermed risiko for at resurserne ikke i tilstrækkelig grad inkluderer og tilgodeser den sprogbrug, og dermed de opfattelser og værdier, som fx kendetegner forskellige mindretal.

Tekstkorpusser indsamles i stigende grad også automatisk af maskiner. Det giver risiko for en for stor repræsentation af maskinoversat, dvs. uoriginal tekst i tekstkorpusser, med risiko for fejl i de leksikografiske resurser.

Nordisk Forening for Leksikografi indkalder hermed forslag til foredrag ved det 31. LexicoNordica-symposium, der kan belyse hvordan der arbejdes med at kvalitetssikre det datamateriale som ligger til grund for nordiske ordbøger. Vi er interesserede i foredrag inden for følgende delemner:

- Komplettering af det eksisterende datamateriale som ligger til grund for nordiske ordbøger: Hvordan sikres det at materialet er så afbalanceret og repræsentativt som muligt, sprogligt og kulturelt?
- Håndtering af megakorpussernes automatisk indsamlede nettekster: Hvordan sikrer man sig imod maskingenereret sprog og indflydelsen fra chatbots o.l.
- Indsamling af materiale som er repræsentativt for minoriteters sprog, talesprog, ungdomssprog, m.fl. Eksempler på datamateriale og evt. alternative måder at indsamle materiale på?
- Eksempler på anvendergenererede resurser som kilder: muligheder og udfordringer?

***

Representativeness and biases in Icelandic corpora

Steinþór Steingrímsson & Einar Freyr Sigurðsson
Árni Magnússon-institutet för isländska studier, Reykjavík
steinthor.steingrimsson@arnastofnun.is & einar.freyr.sigurdsson@arnastofnun.is

In recent years there has been substantial growth in available language resources for use in Icelandic language technology and linguistic research. The Icelandic Gigaword Corpus (IGC) is the largest of these resources. Its latest version comprises approximately 2.5 billion words. A corpus like the IGC can be useful for lexicographers in dictionary work – and it has in fact been used within a number of projects, e.g., adding new words to the Database of Icelandic Morphology. The data contained in these language resources stem from different sources and represent different registers and genres. It can be argued that a certain dataset is in some way representative of a certain type of Icelandic, due to its origins and how it was collected. Nonetheless, all language data is inherently biased to some extent, as collection methods, availability of texts and recordings, and the views of the collectors will always affect the process and its results. For the same reasons, it can also be problematic to define how a corpus is a "balanced" representation of the language of a given period, language community or register. The proportions of data in many recent large corpora projects are affected by the availability of electronic texts, leading to a skewing of the corpus data towards domains that are well represented online. This is quite prominent in the available Icelandic data sets.

When language data are used for research, the researcher must be aware of these limitations. While a large corpus that strives to be "balanced" in some way may be wideranging, due to the nature of available texts, it will in all probability be biased towards conservative language use and the language of the majority of speakers. A corpus of written language may also be missing vital information on spoken language or the language of young people or other groups less well represented in published texts.

In our study we focus on representation and biases in a broad sense and try to answer the following questions:

1. How can we use existing corpora to detect that something is not representative of, e.g., a certain register or that something only represents the language of a subset of speakers?

2. How can we use existing corpora to find biases in our data, such as gender biases?

3. What kind of metadata is needed to facilitate research on biases and representativeness?

***

Fra “sandheden om sproget” til et opgør med stereotyper: korpusleksikografiens udfordringer i 2024

Sanni Nimb
Det Danske Sprog- og Litteraturselskab, København
sn@dsl.dk

I ældre danske ordbøger skinner subjektive holdninger til køn, etnicitet og seksuel orientering ofte igennem i betydningsbeskrivelserne. Med korpusleksikografiens fremkomst fik man mulighed for at bevæge sig væk fra den introspektive metode. Man så det i stedet som leksikografiens opgave at formidle objektive, statistisk baserede oplysninger i deskriptive ordbøger, fx i Den Danske Ordbog (Hjorth & Kristensen m.fl. (2003-2005)), uanset en erkendelse af at de statistiske resultater ofte afspejlede skævheder i opfattelsen af fx køn og minoriteter (jf. Scheuer (1995): "Hans Hustru, hendes Bryster").

I de senere år, hvor korpora også danner grundlag for AI og træning af sprogmodeller, er der opstået et andet syn på hvordan sådanne skævheder bør håndteres. Firmaer der udvikler sprogteknologiske produkter baseret på statistiske analyser af store tekstmængder, ser det som en samfundspligt at undlade at reproducere stereotyper og endda også visse ord. Leksikografer der redigerer korpusbaserede ordbøger, må nødvendigvis forholde sig til dette. Medmindre det i det leksikografiske arbejde er muligt at opbygge tekstkorpora der inkluderer en så stor variation af tekster at stereotype mønstre kan undgås, er man nødsaget til at anvende den samme introspektive metode som man ellers havde forladt, når man beskriver kontroversielle lemmaer. Hvordan gøres dette på en måde hvor man stadig sikrer det videnskabelige grundlag, og hvor man samtidig tager højde for den store variation i sprogbrug og holdninger til ordene som man finder blandt sprogbrugerne og måske også blandt redaktørerne? I foredraget vil jeg give eksempler på stereotype beskrivelser af ord i ældre danske ordbøger og skitsere en metode der kan sikre en mere neutral beskrivelse i moderne ordbøger. Metoden bygger på 1) emnebaseret indkredsning af det ordforråd der bør behandles introspektivt, 2) leksikalske oplysningstyper om kontroversielle ord beskrevet i Huyssteen & Tiberius (2023): ”Towards a lexical database of Dutch taboo language” samt 3) opmærkning af de leksikalske data foretaget af flere redaktører.

***

Talspråk och ordboksresurser

Helga Hilmisdóttir
Árni Magnússon-institutet för isländska studier, Reykjavík
helga.hilmisdottir@arnastofnun.is

Talspråk har länge intresserat isländska ordboksredaktörer och i Blöndals (1920–1924) isländsk-danska ordbok förekommer ett ganska stort antal ord och uttryckssätt som fångats upp i talspråk, t.ex. dialektala ord, nya lånord, svordomar, tilltal och interjektioner. Vid Árni Magnússon institutet för isländska studier förvaras också ett stort arkiv över ordförrådet i det talade språket. Materialet i arkivet samlades in under nittonhundratalets senare hälft och består bl.a. av dialektala ord och uttryckssätt. Materialet har inte bearbetats i en ordbok men en del av de dialektala orden togs upp i *Íslensk orðabók* (1962, 2002). I den nya webbordboken *Íslensk nútímamálsorðabók* har man däremot lagt mycket lite vikt på lokala företeelser. Eventuellt kan detta förklaras med att de korpusar som arbetet bygger på huvudsakligen består av redigerat skriftspråk där det inte finns mycket plats för dialektala ord och uttryck. Eventuellt kunde en stor talspråkskorpus med vardagligt språk lyfta fram fler ord och ordbetydelser och ge en mer nyanserad bild av språket som det ser ut idag.

I detta föredrag kommer jag att diskutera hur talspråk representeras i moderna isländska ordböcker och hur det eventuellt kunde ges en större roll genom användningen av samtalskorpusar. I och med den digitala revolutionen har förutsättningarna för talspråksinsamling ändrats radikalt. Det är nu enklare än tidigare att spela in och transkribera samtal som sedan kan användas som resurs till lexikografiskt arbete (se t.ex. Hilmisdóttir 2021; *Samtalsorðabók*). I föredraget kommer jat att svara på följande frågor: 1) Varför borde vi använda talspråkskorpusar i lexikografiskt arbete?, 2) Vilka frågor skulle användningen av talspråk väcka inom isländsk lexikografi och isländsk språkpolitik? 3) Vad måste man ta i beaktande när man planerar insamling av talspråksmaterial? 4) Vad kan vi som lexikografer förvänta oss att få ut av en välsammansatt talspråkskorpus?

Blöndal, Sigfús. 1920–1924. *Íslensk-dönsk orðabók*. Reykjavík: Íslensk-danskur orðabókarsjóður <blondal.arnastofnun.is>.
Hilmisdóttir, Helga. 2021. Talspråkskorpusar, diskurspartiklar och lexikografi. *LexicoNordica* 28:79–100.
Íslensk nútímamálsorðabók. Þórdís Úlfarsdóttir och Halldóra Jónsdóttir (red.) Árni Magnússon-institutet för isländska studier <islenskordabok.arnastofnun.is>.
Íslensk orðabók handa skólum og almenningi. 1962. Árni Böðvarsson (red.). Reykjavík. Bókaútgáfa Menningarsjóðs.
Íslensk orðabók. 2002. Mörður Árnason (red.). Reykjavík. Edda.
Samtalsorðabók. Helga Hilmisdóttir (red.). Árni Magnússon-institutet för isländska studier <samtalsordabok.arnastofnun.is>.

***

Jagten på hverdagssproget - Brugen af tekster fra internetfora i arbejdet med Den Danske Ordbog

Kirsten Lundholm Appel, Jonas Jensen & Nathalie Hau Sørensen
Det Danske Sprog- og Litteraturselskab, København
ka@dsl.dk & nats@dsl.dk

Lemmaselektion er en central, men tidskrævende, del af det leksikografiske arbejde. Under den løbende opdatering og udvidelse af Den Danske Ordbog (DDO), en korpusbaseret monolingval ordbog med mere end 100.000 lemmaer, har et stadig mere ensidigt korpus gjort det særligt vanskeligt systematisk at fremsøge én bestemt type lemmaer: Dem der er udbredt i hverdagssproget, men som grundet begrænset nyhedsværdi relativt sjældent optræder i korpussets mange journalistiske tekster. I denne undersøgelse præsenteres en metode til automatisk opdagelse af netop disse lemmaer, bl.a. ved hjælp af et nyt korpus med tekster fra chatfora på internettet, der formodes at have en højere koncentration af hidtil oversete ord fra hverdagssproget end det eksisterende korpus. Metoden kræver videreudvikling af et eksisterende værktøj til automatisk fremsøgning af lemmakandidater samt en fremgangsmåde for både kvantitativ og kvalitativ evaluering af de fremsøgte lemmaer.

Værktøjet som metoden hviler på er allerede prøvekørt med succes på det eksisterende korpus (Sørensen et al., 2023). Skønt dette tæller godt 1,1 milliarder løbende ord, lider det under én væsentlig svaghed: Siden årtusindeskiftet er det næsten udelukkende blevet udbygget med avisartikler og andre journalistiske tekster, og det har gradvis gjort korpusset mere ensidigt. Som led i denne undersøgelse har redaktionen derfor forsøgsmæssigt sammensat et korpus bestående af andre teksttyper, specifikt tekster indsamlet fra forskellige chatfora på internettet.

Som nævnt er vores tese at dette nye internetkorpus vil indeholde flere hverdagsord end det eksisterende, og ord af denne type forventes desuden at optræde med større frekvens i internetkorpusset. Før det omtalte værktøj til automatisk opdagelse af lemmakandidater kan finde oplagte lemmaer i det nye korpus, må søge- og udvælgelseskriterierne dog finindstilles. Værktøjet bygger på en vægtet score af en række kriterier, som hver især estimerer et ords potentiale som lemmakandidat – fx ved at finde lignende ord i ordbogen gennem en sprogmodel. Både kriterierne og vægtningen heraf er udviklet på baggrund af gode lemmakandidaters karakteristika i journalistiske tekster. Algoritmen må derfor videreudvikles til at tage højde for forskelle mellem teksttyperne. Vi forventer fx at et diskussionsforum på internettet indeholder flere fejlstavninger end avistekster forfattet af professionelle journalister. Når det tilpassede værktøj efter justering og test har gennemført den automatiske udvælgelse, vil to eller flere redaktionsmedlemmer vurdere et udvalg af de fremsøgte lemmakandidater, dels med henblik på at afdække deres egnethed (og dermed metodens succesrate), dels for at fjerne eventuelle resterende fejllemmatiseringer, vildfarne proprier og andre uønskede indslag fra den automatisk genererede liste.

Gennem disse undersøgelser håber redaktionen at lette arbejdet med lemmaselektion og øge muligheden for automatisk udvælgelse af gode, men hidtil oversete, lemmakandidater der tilhører dén del af hverdagssproget som sjældent optræder i avistekster og lignende.

Sørensen, N. H., Sørensen, N. H., Appel, K. L. and Nimb, S. (2023). Trawling the corpus for the overlooked lemmas. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography*. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023. Brno: Lexical Computing CZ s.r.o., pp. 392-409.

***

Finns det skevheter och utmaningar i de material som enspråkiga finländska ordböcker bygger på?

Caroline Sandström & Tarja Heinonen
Institutet för de inhemska språken, Helsingfors
caroline.sandstrom@sprakinstitutet.fi & tarja.heinonen@sprakinstitutet.fi

På Institutet för de inhemska språken redigeras fyra stora enspråkiga ordböcker. Den finska och svenska dialektordboken *Suomen murteiden sanakirja* och *Ordbok över Finlands svenska folkmål*, en ordbok över äldre finskt skriftspråk *Vanhan kirjasuomen sanakirja* och en nufinsk deskriptiv ordbok *Kielitoimiston sanakirja*. De tre förstnämnda kan alla kategoriseras som s.k. flergenerations ordböcker och redigeringen av dem pågår fortfarande. Den nufinska ordboken *Kielitoimiston sanakirja* omfattar ca 100 000 uppslagsord. Den är publicerad både som tryckt ordbok och digitalt och uppdateras kontinuerligt på nätet.

I vårt bidrag kommer vi i samråd med redaktörskolleger från de fyra ordböckerna att göra en jämförelse som fokuserar på vilka skevheter och utmaningar det finns i de material ordböckerna bygger på. Vi strävar efter att ge en översikt och ta fram skillnader och likheter i de problem som redaktörerna möter, men fokuserar också på hur redaktionerna har löst och hanterat de skevheter och utmaningar som finns i ordböckernas datamaterial. Eftersom vi själva arbetar med *Ordbok över Finlands svenska folkmål* (Sandström), respektive *Kielitoimiston sanakirja* (Heinonen) kommer vi främst att ta fram exempel på problem och hur de har lösts ur dessa två ordböcker.

Ordböckerna publiceras på Institutet för de inhemska språkens webbplats:
https://kaino.kotus.fi/fo/ (finlandssvenska dialektordboken)
https://www.kielitoimistonsanakirja.fi/#/ (nufinska ordboken)
https://kaino.kotus.fi/sms/ (finska dialektordboken)
https://kaino.kotus.fi/vks/ (ordbok över äldre finskt skriftspråk)

***

Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text

Markus Forsberg & Louise Holmer
Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet
louise.holmer@svenska.gu.se

När *Svenska Akademiens ordlista* (SAOL 13) publicerades år 2006 var textmaterialet som ordlistan baserades på enormt, med den tidens mått mätt. Korpusmaterialet då utgjordes av ungefär 200 miljoner ord, hämtade framför allt från svensk tidningstext. När SAOL 14 kom ut år 2015 var det tillgängliga korpusmaterialet betydligt större och uppgick till många miljarder ord. Även den samtida definitionsordboken Svensk ordbok utgiven av Svenska Akademien (SO 2021) bygger på detta stora textmaterial.

Det textmaterial som ligger till grund för ordböckerna är Språkbanken Texts textsamlingar som exempelvis görs tillgängligt genom forskningsverktyget Korp (Borin et al. 2012). Trots sin relativa storlek har samlingarna utvecklat en slagsida åt olika texttyper genom åren; först åt tidningstext och därefter åt olika former av allmänt tillgängliga resurser som bloggar, twittertexter, diverse nätforum m.m. Materialet med tidningstexter har dessutom lidit av att vara relativt ojämnt fördelat över både tid och geografisk hemvist. Att få tillgång till den typ av material som tidningstexter utgör har under åren från 2010 och framåt dessutom visat sig svårare än tidigare, framför allt av rättighetsskäl.

Sedan 2021 finns dock en betydande positiv utveckling i fråga om både nya material och nya språkteknologiska metoder, tack vare ett samarbete mellan Språkbanken Text vid Göteborgs universitet, där ordboksredaktionen numera ingår, och KB-labb vid Kungliga Biblioteket i Stockholm (Kungliga Biblioteket 2023). I fråga om material har samarbetet lett fram till följande nya datasamlingar, som är samlade i Kubord-läget i Korp:

- Årgångar av morgontidningarna Dagens Nyheter, Svenska Dagbladet och Göteborgs-Posten från år 2010 (2013 för Göteborgs-Postens del) till år 2021
- Årgångar av kvällstidningarna Aftonbladet och Expressen från år 2010 till år 2021
- Årgångar av de regionalt täckande Sydsvenskan och Östgötakorrespondenten från år 2010 till år 2021

Dessa nya material uppgår i nuläget sammanlagt till ca 5 miljarder ord, men som av upphovsrättsliga skäl är begränsade till enskilda ord som berikats med språkteknologisk analys — exempelvis ordklasstaggning, syntaxanalys och lemmatisering — tillsammans med detaljerade källhänvisningar. Trots begränsningen, så möjliggör denna datasamling diverse olika nyordsundersökningar i och med att varje årgång ord kan jämföras med den föregående. Vidare ingår både morgonpressen och kvällspressen, vilket inte har varit fallet tidigare, liksom en större geografisk representation av texter än förut.

Utöver dessa förädlade ordsamlingar så samarbetar vi för närvarande med att hitta andra typer av representationer av upphovsrättsligt begränsade material i KB:s samlingar, som möjliggör obegränsad tillgång, till exempel ordvektorer (Forsberg och Sköldberg 2022).

Sammanfattningsvis innebär samarbetet utökade möjligheter att balansera Språkbanken Texts samlingar och utveckla fler vetenskapliga metoder för analys av ordens semantiska och kombinatoriska egenskaper. I föredraget presenteras både materialen och metoderna närmare, liksom det lexikografiska arbetet.

Borin, Lars, Markus Forsberg och Johan Roxendal. 2012. ”Korp – the corpus infrastructure of Språkbanken”. I: *Proceedings of LREC 2012. Istanbul: ELRA*, volume Accepted, 474-478.

Forsberg, Markus och Emma Sköldberg. 2022. ”Ordvektorer i lexikografiskt arbete”. I: Volodina, Elena et al. (red.), *Live and Learn. Festschrift in honor of Lars Borin*. Göteborg: Institutionen för svenska, flerspråkighet och språkteknologi, 37–41.

KB-labb: Kungliga Biblioteket 2023. Tillgänglig: https://www.kb.se/samverkan-och-utveckling/kb-labb.html

Korps Kubordläge: https://spraakbanken.gu.se/korp/?mode=kubord

Projektet Svenska Akademiens samtidsordböcker: https://spraakbanken.gu.se/projekt/svenska-akademiens-samtidsordbocker

Språkbanken Text: http://spraakbanken.gu.se

Svenska.se 2023. Svenska Akademiens ordboksportal. Tillgänglig: https://svenska.se/

***

Ordbog over moderne islandsk: udvikling og tilføjelser

Ellert Þór Jóhannsson & Þórdís Úlfarsdóttir
Árni Magnússon-institutet för isländska studier, Reykjavík
etj@hi.is & thordis.ulfarsdottir@arnastofnun.is

Íslensk nútímamálsorðabók (Ordbog over moderne islandsk = OMI) er et leksikografisk projekt, der har til formål at beskrive ordforrådet i det nutidige islandske sprog. Denne ordbog er forankret i ISLEX, en flersproget islandsk-nordisk ordbog (se Úlfarsdóttir 2013), og er blevet yderligere bearbejdet og suppleret med nye korpusdata og andre specifikke initiativer (se nærmere Jónsdóttir & Úlfarsdóttir 2019).

I dette foredrag ville vi gennemgå forskellige tilføjelser til ordbogen og motivationen bag dem. Vi forklarer, hvordan projektet har udviklet sig og hvordan det berører forskellige problemstillinger, der er relevante for moderne leksikografi. Dette inkluderer, hvordan korpusdata opdateres, og hvilke foranstaltninger der skal træffes for at bevare ordbogens nøjagtighed, repræsentation og relevans i forhold til det aktuelle aktive sprog. Vi fokuserer på fire strategier for at opretholde ordbogens materiale afbalanceret og repræsentativt:

1) **Nye korpusdata:** Integreringen af nye korpusdata er afgørende for at holde OMI opdateret og for at bevare dens aktualitet i forhold til hvordan sproget bliver brugt. Indførelsen af friske sprogdata gør at OMI afspejler sprogsamfundet så omfattende som muligt og hjælper med at sikre ordbogens relevans.
2) **Specialiseret excerpering:** Inkludering af specialiseret ordforråd, (fx termer relateret til vulkanudbrud eller helsekost), viser en bevidsthed om behovet for at omfatte forskellige domæner inden for det islandske sprog, som måske ikke er repræsenteret i tilgængelige korpusdata. Dette har betydning for spørgsmålet om datasammensætning, og at ordbogen ikke kun repræsenterer almindeligt dagligdags sprog, men også specialiseret terminologi, der er vigtig for specifikke emner. Her er det dog nødvendigt at vælge omhyggeligt hvilken af sådan termer er relevante for almindelige brugere.
3) **Brugerbidrag:** Brugerne bidrager ofte med tilføjelser og forslag til ordbogen, hvilket bringer den tættere på de personer, der anvender materialet. En kritisk gennemgang og evt. ændringer i materialet pga. brugerhenvendelser sikrer, at ordbogen er inkluderende og tættere på det aktuelle sprog, der anvendes af forskellige grupper. Dette afspejles også i den løbende diskussion om brugergenererede ressourcer som værdifulde bidrag til leksikografi.
4) **Redaktionelle forbedringer:** Redaktionelle tilføjelser, såsom substantiver til tilsvarende verber eller omvendt, illustrerer behovet for, at leksikografer tilpasser eksisterende data. Denne form for redaktionelt arbejde mindsker risikoen for, at ordbogen muligvis overser nogle vigtige og relevante ord og gør den bedre egnet til at fange ændringer i ordforrådet.

OMI stræber efter at registrere og belyse det nutidige islandske ordforråd og forblive et værdifuldt referenceværktøj for alle, der er interesseret i det islandske sprog. Ved at se nærmere på nogle af de udfordringer ved at forbedre ordbogens data og redegøre for forskellige strategier for at sikre ordbogens nøjagtighed, repræsentation og relevans, understreger vi OMI's rolle i det stadigt skiftende sproglige landskab, samtidig med at vi illustrerer, hvordan ordbogen bedst kan afspejle den aktuelle sprogbrug.

ISLEX-orðabókin. Úlfarsdóttir, Þórdís (red.). Stofnun Árna Magnússonar í íslenskum fræðum. https://islex.arnastofnun.is/is/ (september 2023).

Íslensk nútímamálsorðabók. Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir (red.). Stofnun Árna Magnússonar í íslenskum fræðum. https://islenskordabok.arnastofnun.is/ (september 2023).

Jónsdóttir, Halldóra, & Þórdís Úlfarsdóttir (2019). Íslensk nútímamálsorðabók. *Orð og tunga* 21:1-25. https://doi.org/10.33112/ordogtunga.21.2.

Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir (2020). Omdannelsen af en flersproget til en monolingval ordbog. *Nordiska studier i lexikografi 15. Rapport från 15 konferensen om lexikografi i Norden. Helsingfors 4—7 juni 2019:*175—186. Helsingfors: Skrifter udgivet af Nordisk Forening for Leksikografi.

Risamálheildin. Stofnun Árna Magnússonar í íslenskum fræðum. http://malheildir.arnastofnun.is/ (september 2023).

Steingrímsson, Steinþór, Sigrún Helgadóttir & Eiríkur Rögnvaldsson (2018). An Icelandic Gigaword Corpus. *Nordiske Studier i Leksikografi. Rapport fra 14. Konference om Leksikografi i Norden, Reykjavík 30. maj – 2. juni 2018*:246–254. Reykjavík: Skrifter udgivet af Nordisk Forening for Leksikografi.

Úlfarsdóttir, Þórdís (2013). ISLEX – norræn margmála orðabók. *Orð og tunga 15*:41-72.

***

Vegen vidare for dialektordboka Norsk Ordbok

Gyri Smørdal Losnegaard
Institutt for lingvistiske, litterære og estetiske studier, Universitetet i Bergen
Gyri.Losnegaard@uib.no

I 2015 vart tolvte og siste bind av det dokumentariske ordbokverket *Norsk Ordbok* ferdigstilt. Året etter vart ordboka og samlingane ho byggjer på, overført til Universitetet i Bergen (UiB), og frå 2018 har det vore løyvd offentlege midlar til å bygge opp eit fagmiljø i leksikografi der. *Norsk Ordbok*, som er både ei historisk ordbok over det nynorske skriftspråket og ei ordbok over dei norske dialektane, er i dag under revisjon i prosjektet Norsk Ordbok a-h (NO-AH), finansiert av Kultur- og likestillingsdepartementet. Eit hovudmål i prosjektet er å lage ei komplett og moderne nettutgåve av ordboka, som frå no av skal vere ei rein nettordbok. I dette føredraget vil eg prøve å identifisere dei viktigaste utfordringane knytt til innsamling og organisering av nytt kjeldemateriale i det pågåande revisjonsarbeidet, med hovudvekt på talemålsdokumentasjon.

Norsk Ordbok byggjer på eit variert og samansett kjeldegrunnlag, ein naturleg konsekvens av at verket dokumenterer både skrift og tale. Den opphavelege målsettinga var at ordboka skulle arbeide saman alt som fram til då «var bokført av norske ord», og at ho skulle bygge på alt av nytt tilfang frå den nynorske litteraturen og dei norske dialektane. Med andre ord – alt skulle med.

I redigeringsarbeidet har desse kjeldene vore dei mest sentrale fram til i dag:

- setelarkivet til Norsk Ordbok
- Nynorskkorpuset
- nettbiblioteket til Nasjonalbiblioteket
- Norsk målføresynopsis
- databasen Ordbokshotellet

Eit viktig spørsmål når ein no skal planlegge framtida til *Norsk Ordbok* som ei rein nettordbok, er kva ei moderne og oppdatert dialektordbok kan og bør vere. Kva skal ordboka dokumentere, og kva kjelder skal ein bruke? Det er ikkje lenger realistisk at alt skal med. Ei fullstendig kartlegging av norske dialektar er ikkje mogeleg å gjennomføre, og enno mindre om ein skal legge diakrone omsyn til grunn. Der ein før skulle dokumentere det "tradisjonelle" i dialektane, er det dermed eit spørsmål kva ein skal ta mål av seg å dokumentere i dag. Det vil krevje klare avgrensingar og tydelege retningslinjer. I dag kommuniserer vi i hovudsak digitalt, og digitaliseringa av samfunnet opnar nye mogelegheiter med omsyn til innsamling av nytt talemålsmateriale. Det fører med seg ei rekke spørsmål: kva vil det å be brukarar om bidrag medføre av arbeid i form av sortering og redaksjonelle vurderingar? Kva vil dette krevje av tilrettelegging og kvalitetssikring? Når vi ikkje lenger har eit fast nettverk av godkjende informantar, skal ein byggje opp eit nytt, eller skal ein ta i mot målføreopplysningar direkte frå ordbokbrukarane? Bør ein stille visse krav til bidragsytarar? Kan vi utnytte data frå eksisterande dialektforsking? Kan ein sjå føre seg ei

nyinnsamling knytt til målføresynopsisen, og kan dette eventuelt gjennomførast ved bruk av ny og brukarstyrt teknologi?

***

Fra ordinnsamling, oversettelser og arkiver til kvensk ordbok

Anna-Kaisa Räisänen, Aili Eriksen, Trond Trosterud, Thomas Brevik Kjærstad & Tobias Kvalness
Kvensk Institutt, Nasjonalt senter for kvensk språk og kultur, Børselv, Nordnorge
anna-kaisa.raisanen@kvenskinstitutt.no & aili.eriksen@kvenskinstitutt.no

Kvensk er et truet språk. Det har bare unntaksvis blitt overført innad i kvenske familier siden 1960-tallet. I dag finns det under 5000 språkbrukere, og de som har kvensk som morsmål tilhører aldersgruppen over 60 år. For å hindre at kvensk blir et utdødd språk, har både språkforskere, språkbrukere og kvenske institusjoner jobbet for å revitalisere det kvenske språket.

En sentral del av revitaliseringsarbeidet har vært å utvikle tilstrekkelige ressurser for språkundervisning og dermed har etablering av kvensk skriftspråk gått hånd i hånd med produksjon av læremateriell både til grunnskole og til universiteter. Samtidig har Kvensk institutt gjort kvensk synlig også på andre samfunnsområder med oversettelser, spesielt til offentlig forvaltning og til andre kvenske institusjoner.

En sentral ressurs til lærere, elever, studenter og oversettere er tospråklig kvensk digital ordbok. Når Kvensk institutt og Giellatekno satt i gang ordboksarbeid i 2014, tok de utgangspunkt i eksisterende ordlister med til sammen 7500 ord. Tidligere er disse ordlistene blitt brukt i kvenskundervisninga på universitetet i Tromsø og i skoler i Porsanger. Kvensk institutt og Kvensk språkting startet også ordinnsamling fra ulike kvenske dialektområder i 2014. Kvenske oversettere utvikler nytt ordforråd for kvensk. Ordinnsamlingen og utviklingen av nytt ordforråd har vært tett knyttet til leksikografisk arbeid ved instituttet. Parallelt har Kvensk institutt og Giellatekno ved UiT utformet kvensk tekstkorpus (Korp). Den er basert på de skriftlige ressurser som eksisterer på kvensk, og det suppleres vedvarende med nye tekster. Med eksisterende datamateriale har kvensk ordbok blitt utvidet fra 7500 ord til 15 000 kvensk oppslagsord.

I de siste årene har kvensk leksikografi vært tett knyttet til språkteknologi, og utviklinga av ordbøkene utnytter språkteknologiske ressurser som er innpasset i ordbøkene. Ved bruk av språkteknologiske verktøy kan man effektivt undersøke i hvor stor grad ordforrådet i det kvenske tekstkorpuset i Korp er dekket av ordboka.

I dag er kvensk tekstkorpus i stor grad basert på oversatte tekster. Det finnes mye arkivmateriale i ulike dialektarkiver som ikke er brukt systematisk i leksikografisk arbeid med kvensk ordbok. Den store utfordringa i kvensk leksikografi er på den ene siden å kunne bruke eksisterende datamaterialer fra dialektarkivene, og på den andre siden å kunne sørge for at ordboksarbeidet kan bidra på en best mulig måte til revitaliseringa av kvensk språk.

***

Språkteknologi för att samla in texter och analysera språket i Korp – hur gör man på meänkieli?

Rickard Domeij, Jacob Larsson, Lina Lejdebro Enwald, Elina Kangas, Magnus Ahltorp & Gunnar Eriksson
Språkrådet, Institutet för språk och folkminnen (Isof), Sverige
Jacob.Larsson@isof.se, Elina.Kangas@isof.se & Magnus.Ahltorp@isof.se

Språkrådet i Sverige har till uppgift att stödja utvecklingen av språkteknologi för språken i Sverige. Det gör vi i första hand genom Nationella språkbanken[1], som vi under namnet Språkbanken Sam driver tillsammans med Språkbanken Text vid Göteborgs universitet och Språkbanken Tal vid Kungliga Tekniska Högskolan i Stockholm. Nationella språkbankens syfte är att skapa möjligheter att forska i digitala text- och talmaterial med hjälp av verktyg och metoder i gränslandet mellan språkteknologi och AI.

Ett centralt verktyg i det arbetet är Korp[2] som är Språkbankens korpusverktyg med vilket man kan söka i stora mängder svensk text från bland annat dagstidningar, skönlitteratur och sociala medier. Med Korp kan språkforskaren effektivt analysera stora textmängder för att se hur ord används över tid i olika kontexter och texttyper. För Språkrådets språkvårdare i svenska är det en ovärderlig källa till empirisk kunskap om språket och utgör underlag för att bedriva språkvård, konstruera lexikon och andra nödvändiga språkresurser.

För Språkrådets språkvårdare i meänkieli är situationen en annan, liksom för flera andra av de nationella minoritetsspråken. Där saknas grundläggande språkteknologiska verktyg och textresurser för att analysera språket i Korp och utveckla stavningskontroll, elektroniska ordböcker och andra viktiga verktyg som krävs för språkanvändning i digitala sammanhang. Sådana resurser är inte bara nödvändiga för språkforskningen, utan också för språkets användning, normering och överlevnad i stort (Domeij et al 2022, Mattson & Ahltorp 2023).

Därför har Språkrådet börjat ta fram grundläggande språkteknologi för meänkieli i samarbete med språkvetare/språkvårdare i meänkieli och Universitetet i Tromsö (UiT), Norges arktiska universitet som en del i ett större arbete med språkteknologi för de små språken i Norden (Domeij et al 2023). Vi har skapat ett elektroniskt lexikon (Ordbok meänkieli-svenska)[3] som också utgör grund till den pågående utvecklingen av en språkmodell som både kan användas för stavningskontroll och för textanalys av meänkieli i Korp. Vi har också börjat titta på metoder för att automatiskt identifiera och samla in texter på meänkieli på webben (t.ex. Skeppstedt et al 2020).

[1] Arbetet med utveckling av språkteknologi för meänkieli möjliggörs av Vetenskapsrådet genom forskningsinfrastrukturen Nationella språkbanken (2017-00626). https://sprakbanken.se/
[2] https://spraakbanken.gu.se/verktyg/korp
[3] Ordboken har tagits fram av Meän Akateemi och Giellatekno vid UiT med finansiering av Isof som ansvarar för drift och utveckling av gränssnitt. https://www.isof.se/nationella-minoritetssprak/meankieli/sprakhjalp/ordbok-meankieli-svenska

I presentationen beskriver vi hur vi går tillväga för att identifiera, samla in, analysera, bearbeta och söka i texter på meänkieli för att förstå språket och språkanvändningen, uppdatera lexikon och bedriva språkvård. Vi beskriver också de särskilda utmaningar som små språk med knappa textresurser och pågående normeringsprocess ställs inför vid skapandet av en systematiskt insamlad korpus och de språkteknologiska verktyg som krävs för att utforska den.

Domeij, Rickard, Kristine Eide, Peter Juel Henrichsen, Per Langgård, Sjur Moshagen Nørstebø (2023): *Initiative to promote language technology and CALL for the Nordic minority languages*. Abstract till presentation på Eurocall 2023, Reykjavik.

Domeij, Rickard, Ola Karlsson, Trond Trosterud & Sjur Nørstebø Moshagen (2019): Enhancing information accessibility and digital literacy for minorities using language technology: The example of Sami and other national minority languages in Sweden. I: *Perspectives on Indigenous Writing and Education* [red] Kirk P. H. Sullivan and Coppélie Cocq, Leiden, The Netherlands: Brill, 2019.

Mattson, Marie & Magnus Ahltorp (2023): Lexikografiska resursers betydelse i utvecklingen av språkteknologiska verktyg för minoritetsspråk. I: *Lexiconordica 2023*, under utgivning.

Nørstebø Moshagen, Sjur, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen & Per Langgård (2022): *Report on the Nordic Minority Languages*. European Language Equality Consortium.

Skeppstedt, Maria, Elina Kangas, Peter Ljunglöf, Magnus Ahltorp, Gunnar Eriksson, & Rickard Domeij (2020): *Plans for using texts from public authorities for creating a partly parallel Meänkieli corpus.* Parallel Corpora as Digital Resources and Their Applications.

***