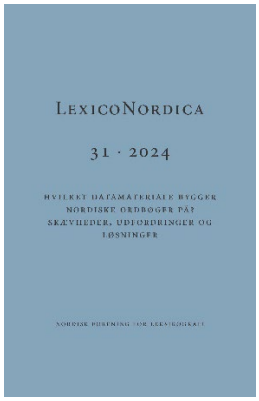
LexicoNordica

Titel:	Hvilket datamateriale bygger nordiske ordbøger på? Skævheder, udfordringer og løsninger
Forfatter:	Henrik Hovmark & Terje Svardal
Kilde:	LexicoNordica 31, 2024, s. 7-16
URL:	https://tidsskrift.dk/lexn/issue/archive



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

Henrik Hovmark & Terje Svardal

Det er med stor fornøjelse at Nordisk Forening for Leksikografi (NFL) hermed kan præsentere endnu et bind af tidsskriftet *LexicoNordica*, det 31. i rækken. Som sædvanlig består hovedparten af årets nummer af en tematisk del med i alt ni artikler der på forskellig vis belyser emnet: *Hvilket datamateriale bygger nordiske ordbøger på? Skævheder, udfordringer og løsninger*. De ni artikler baserer sig på foredrag holdt ved det 31. *LexicoNordica*-symposium 15.-17. februar 2024 med samme emne, og de inddrager hver især i forskelligt omfang de mange erfaringsudvekslinger som fandt sted i løbet af symposiet. Symposiet blev ligesom forrige år afholdt på Voksenåsen, Norges nationalgave til Sverige, lidt uden for Oslo. Det er med stor glæde at Nordisk Forening for Leksikografi på denne måde atter har fået mulighed for at forankre sine aktiviteter i et nordisk miljø. Nummeret indeholder desuden en anmeldelse og et mindeord, samt en orientering fra bestyrelsen for Nordisk Forening for Leksikografi og til slut redaktionelle anvisninger.

Redaktionen for *LexicoNordica* ønskede med årets symposium at rette et særskilt og kritisk blik på det datamateriale som nordiske ordbøger og andre resurser bygger på. Leksikografiske resurser spiller som udgangspunkt en central rolle for nordisk sprogforståelse, både mellem talere af de forskellige sprog og for gensidig forståelse mellem forskellige etniske og sociale grupper. Men en resurse er aldrig bedre end det grundlag som resursen bygger på. I de senere år er der kommet et langt større og mere kritisk fokus på de skævheder og fælder som kan gemme sig i leksikografisk kil-

dematerialе, også i de store tekstkorpusser som ellers på afgørende punkter revolutionerede det leksikografiske arbejde og gav langt mere velfunderet og direkte, empirisk adgang til større mængder af autentisk sprogbrug. Men tekstkorpusser (og andre samlinger af kildemateriale) kan være skævt sammensatte, med for ringe repræsentation af fx talesprog/hverdagssprog, bestemte tekstgenrer eller unges sprog. Der er dermed risiko for at resurserne ikke i tilstrækkelig grad afspejler det omgivende samfund og den fulde, aktuelle sprogbrug. Der er mere specifikt også risiko for at resurserne ikke i tilstrækkelig grad inkluderer og tilgodeser den sprogbrug – og dermed indirekte også de opfattelser og værdier – som fx kendetegner forskellige mindretal.

Med den stigende kildekritiske tilgang til korpusser og brugen af dem i leksikografisk praksis er man generelt blevet mere bevidst om at et korpus, uanset størrelse og sammensætning, altid vil have en bestemt profil eller bestemte karakteristika. Typisk vil nogle genrer, teksttyper eller samfundsdomæner være bedre repræsenteret end andre. Leksikografen er i sit arbejde således nødt til at være bevidst om disse karakteristika for at undgå ukritisk at gøre den leksikografiske fremstilling skæv eller i modstrid med samfundsrelaterede hensyn og udviklinger. Man kan ikke nødvendigvis uden videre kopiere et statistisk output fra korpus over i ordbogen. Det er ligeledes mere end nogensinde vigtigt at være bevidst om hvorvidt det datamateriale man har til rådighed, i tilstrækkelig grad er repræsentativt i forhold til den specifikke funktion som en given ordbog eller leksikalsk resurse har eller skal have.

I og med at et korpus rummer store mængder af autentisk sprog og sprogbrug, vil det også afspejle forskellige holdninger og værdier som er på spil i (sprog)samfundet. Dette er et af hovedpointerne ved at have et korpus, nemlig muligheden for at have empirisk belæg på hvordan sproget rent faktisk udfolder sig i brug og kontekst. Det kan imidlertid give udfordringer i en verden hvor man bliver mere og mere bevidst om forskellige samfundsgruppers ret-

tigheder, og om at sprog og sprogbrug i mange tilfælde er bærere af stereotyper og kan være kontroversielle. Denne udvikling har fx haft konsekvenser for arbejdet med de store almensproglige ordbøger som ikke kun bruges bredt i befolkningerne, men som også citeres ofte og opfattes som "officielle" af mange mennesker. Det forventes at disse ordbøger er i pagt med tiden og tager hensyn til holdninger og værdier og ikke ukritisk viderebringer kontroversiel sprogbrug. Dette understreger behovet for at være refleksiv og kritisk som leksikograf i forhold til brugen af korpusser, ikke kun hvad angår definitioner, men også fx eksempelmateriale.

Netop dette forhold står centralt hos Einar Freyr Sigurðsson og Steinþór Steingrímsson som i detaljer viser hvordan to store islandske korpusser gemmer på en række skævheder og stereotype forestillinger med hensyn til fx køn (mænd og kvinder). Det interessante og afgørende resultat af forfatternes undersøgelse er at disse skævheder ofte ligger mere eller mindre implicit eller skjult – de viser sig ikke nødvendigvis ved simple, automatiske statistiske undersøgelser, tværtimod kan disse give et falsk billede. Der påvises også tydelige forandringer over tid i den måde sproget bruges på. Undersøgelsen vidner generelt om at sprogteknologisk kompetence på højt niveau er til stor nytte inden for leksikografien, fx i form af undersøgelser af word embeddings, men forfatterne peger også på at en øget og mere bevidst brug af metadata kan være en vej frem mod en mere præcis og nuanceret udnyttelse af det væld af informationer som de store tekstkorpora faktisk rummer.

Erkendelsen af at korpusser bør have en passende spredning med hensyn til genrer, teksttyper og domæner er indlysende, og den er da heller ikke af ny dato. Ikke desto mindre er det blevet sværere og sværere at opfylde denne målsætning. Korpusser består i stigende grad eller næsten udelukkende af avistekster, idet andre genrer og teksttyper ofte ikke umiddelbart kan stilles til rådighed som følge af lovgivning og regler vedrørende ophavsrettigheder og persondataoplysninger. Dette kendetegner fx det store

korpus som afløste det oprindelige og langt mere repræsentativt sammensatte, men mindre korpus ved *Den Danske Ordbog*. Kirsten Appel, Nathalie Hau Sørensen og Jonas Jensen beskriver i den forbindelse hvordan man for at råde bod på denne skævhed målrettet har etableret et alternativt korpus baseret på tekster fra internettet (chatfora) med henblik på at få et datagrundlag som i højere grad giver et indblik i hverdagssproget. Man har samtidig udviklet et værktøj (Findor) som automatisk kan fremfinde kandidater til nye lemmaer i ordbogen. Forfatterne beskriver hvordan det nye værktøj, som er under stadig udvikling, ikke kun baserer sig på frekvens, men foretager en lang række filtreringer af datamaterialet, med lovende resultater.

Den Danske Ordbog er ikke en talesprogsordbog, men brugen af talesprogligt farvede chatfora er begrundet i at denne type datamateriale giver adgang til og indblik i hverdagssprog som ikke vil optræde i avistekster. Helga Hilmisdóttir redegør for hvordan man også i islandsk kontekst arbejder med talesprogsmateriale som supplement til de store korpusser. I dette tilfælde er man gået skridtet videre og arbejder for tiden med deciderede talesprogskorpusser. Forfatteren viser hvordan arbejdet med især ét talesprogligt specialkorpus har bidraget med vigtig, specifik information om almensproget: anglicismer, pragmatiske funktioner af adverbier og diskursive funktioner af ord og kollokationer. Derudover introducerer forfatteren en alternativ måde at præsentere talesproglige træk på, i form af en særlig portal med transskriberede samtaler der inddrager lydklip. Også ved Språkbanken Text i Göteborg gøres der aktuelt en stor indsats for at supplere eksisterende korpusser. Markus Forsberg og Louise Holmer gør rede for hvordan man siden 2021 har arbejdet bevidst med at skabe et mere afbalanceret korpus med hensyn til såvel tid og genrer, men også med hensyn til geografi/sted for sikre større repræsentation af regionalt sprog. Repræsentationen af skønlitteratur er også blevet styrket, ligesom forskellige, mindre specialkorpusser er blevet

tilføjet. Samlet set er datagrundlaget for såvel *Svenska Akademiens ordlista* (SAOL) og *Svensk ordbok utgiven av Svenska Akademien* (SO) blevet langt bedre med disse initiativer.

I den bedste af alle verdener indsamler man et repræsentativt korpus af tilstrækkelig størrelse og med fyldige metadataoplysninger som herefter kan danne solidt grundlag for en eller flere leksikografiske resurser. Men som det allerede er fremgået, ser virkeligheden sjældent ud på den måde. Artiklerne i dette bind af *LexicoNordica* giver mange eksempler på hvordan ordbogsprojekter og -redaktører håndterer forskellige udfordringer knyttet til det leksikografiske grundlagsmateriale. Tarja Riitta Heinonen og Caroline Sandström fremdrager fx et udbredt forhold fra den brogede virkelighed, nemlig at ordbøger ofte er længerevarende projekter, og at vitale dele af grundlagsmaterialet kan være af ældre dato som ikke findes som tekstkorpusser, men som samlinger af excerperet datamateriale. Selve ordbøgerne kan også beskrive ældre sprogtrin og sprogbrug. Disse træk gør sig således på forskellig vis gældende for de to store dialektordbøger og for ordbogen over ældre skriftsprog i Finland. Her kan det være langt vanskeligere at supplere sit grundlagsmateriale. For historiske beskrivelser af skriftsprog er muligheden for at opbygge et repræsentativt korpus begrænset af de tekster der overhovedet er overleveret. Og hvad angår dialektordbøger, altså talesprogsordbøger, er ældre sprogtrin ofte forsvundet og kan ikke længere dokumenteres. Forfatterne gør dog samtidig opmærksom på et forhold som ofte er underbelyst, nemlig at det kontinuerlige arbejde med denne type udfordrende og sparsomme datamateriale har fremelsket en meget detaljeret kildekritisk sans og praksis hos redaktørerne, hvor også omstændigheder omkring indsamlingen af data spiller en stor rolle. Samtidig skal ordbøgerne fungere i en nutidig sammenhæng, og her dukker mulige skævheder i forhold til fx beskrivelsen af køn og folkeslag op.

Tarja Riitta Heinonen og Caroline Sandstrøm berører også

et andet aspekt fra virkeligheden, nemlig begrænsede resurser. I arbejdet med den finske almensproglige ordbog *Kielitoimiston sanakirja* (Språkbyråns ordbok), som i øvrigt bygger på to ældre ordbøger og en ældre seddelsamling, har det ikke været muligt at etablere et eget, repræsentativt korpus over nutidsfinsk, hvilket ellers ville være overordentlig ønskværdigt i betragtning af at ordbogen både har deskriptiv og præskriptiv funktion. Redaktørerne er i stedet nødt til at bruge andre, eksisterende korpusser og komplettere grundlagsmaterialet mere målrettet inden for bestemte domæner. De manglende resurser, kombineret med det ældre grundlagsmateriale og dets præg af bestemte redaktørers personlige interesser, gør imidlertid også at visse domæner er langt bedre dækket ind end andre.

Målrettet komplettering af bestemte samfundsdomæners aktuelle sprogbrug fremhæves også som en vigtig metode i Ellert Þór Jóhannsson og Þórdís Úlfarsdóttirs beskrivelse af arbejdet med den nye islandske samtidsordbog, *Íslensk nútímamálsorðabók* (Ordbog over Moderne Islandsk). Forfatterne identificerer fire processer i suppleringen af datagrundlaget og dermed optagelsen af nye lemmaer: korpusdata, selektiv excerpering inden for bestemte domæner, feedback fra brugere og særlige redaktionelle tilføjelser (fx ord med produktive suffikser eller neologismer). Forfatternes undersøgelse minder om at grundlagsmateriale og arbejdet med det er mangesidigt og inddrager ganske forskellige kildetyper. Tekstkorpusser vil typisk spille en meget central rolle, men i praksis bliver de ofte suppleret af andre kilder og input.

En anden ting som fremgår af Jóhannson og Úlfarsdóttirs beskrivelse, er hvor vigtig en rolle den menneskelige faktor spiller. Uanset hvilken datatype der er tale om – tekstkorpusser, seddelsamlinger, brugerrespons – vil det næsten altid være nødvendigt med en grad af menneskelig redaktionel mellemkomst hvis man skal undgå skævheder og holde en tilfredsstillende, høj kvalitet. Sanni Nimb har særskilt fokus på dette forhold i sin artikel. Ud fra

den generelle og nu mere udbredte erkendelse af at en ukritisk import af sprogbrug fra tekstkorpusser også vil importere en række stereotype frem- og forestillinger som vil være kontroversielle og potentielt krænkende (jf. Sigurðsson og Steingrímssons artikel), argumenteres der for at menneskelig vurdering er nødvendig. Men vel at mærke i en mere systematisk metodologisk form. Med inspiration fra sydafrikanske og hollandske undersøgelser gives et bud på en proces hvor en bredt sammensat gruppe af redaktører uafhængigt af hinanden gennemgår og vurderer både eksisterende lemmaer og lemmakandidater ud fra kriterier der inddrager grader af mulig kontroversialitet og de forskellige synsvinkler og pragmatiske kommunikative funktioner som kan være af betydning når et potentielt krænkende ord bruges i en ytringskontekst. En vigtig pointe er at et sådant arbejde må gentages overraskende hyppigt (fx hvert 5. år) – så hurtigt kan ikke kun sproget, men også tilhørende værdier forandre sig.

Endelig gør to artikler i bindet opmærksom på en helt særlig problematik og situation i henseende til leksikografisk grundlagsmateriale, nemlig tilfælde hvor datagrundlaget på en måde slet ikke eksisterer – endnu. Det gælder minoritetssprog, truede sprog, som ikke har haft nogen skrifttradition og dermed heller ikke en normering, i det aktuelle tilfælde de to nært beslægtede sprog kvensk og meänkieli. Der mangler simpelthen tekster til at opbygge et korpus. Men ikke nok med det: Der mangler jævnligt ordforråd inden for forskellige domæner fordi sproget ikke har haft officiel status og dermed ikke er blevet brugt i en række centrale samfundsfunktioner (skole, administration m.m.). Antallet af talere kan desuden være så lavt at sprog og sprogbrug i et vist omfang må skabes på ny.

Anna-Kaisa Räisänen, Aili Eriksen, Thomas Brevik Kjærstad og Trond Trosterud gør rede for situationen for kvensk og beskriver hvordan arbejdet med kvensk-norsk-kvensk ordbog (*Nettidigisanat Kvääni-norja-kvääni-nettisanakirja*) og indsamlingen af

de få tekster der findes, spiller en central rolle i arbejdet med revitalisering af det kvenske sprog og udviklingen af ordforrådet. Arbejdet omfatter ligeledes sprogteknologisk samarbejde med UiT Norges arktiske universitet. En række udfordringer berøres også, fx behovet for at tage hensyn til dialektale forskelle og identitet i normeringsarbejdet og den nødvendige udvikling af et skriftsprog. Næste skridt vil være udnyttelse af en stor mængde lydoptagelser af de forskellige kvenske varieteter i det finske dialektarkiv. Magnus Ahltorp, Lina Lejdebro Enwald, Elina Kangas, Jacob Larsson, Rickard Domeij og Gunnar Eriksson redegør for arbejdet med at etablere sprogteknologiske værktøjer som vil gøre det muligt at skabe et korpus af tekster på meänkieli som lever op til gængse standarder og krav, og som dermed vil være brugbart i en moderne kontekst. En hovedpointe er at ordbog, sprogmodel og korpusværktøj er gensidigt afhængige af hinanden, og at (videre)udviklingen af det ene element dermed også vil fremme (videre)udviklingen af de andre. Også dette arbejde er imidlertid udfordret af at antallet af tekster på meänkieli er begrænset, både i antal og med hensyn til spredning på genrer, teksttyper osv. Forfatterne peger desuden på at interessen for at investere i sprogteknologiske værktøjer og resurser fra kommencielle aktører er begrænset fordi markedet er så lille. Både hvad angår kvensk og meänkieli er arbejdet synligt præget af det muliges kunst, ligesom det i disse situationer i særlig grad er nødvendigt med tæt, målrettet og ofte også opsøgende kontakt med sprogbrugerne.

Det leksikografiske arbejde med minoritetssprogene minder om et forhold som ofte overses i en verden hvor man har en tendens til at sige at den traditionelle ordbog er en saga blot. Ordbøger, eller rettere: de ord og den viden som opsamles systematisk i ordbøger, har stadig en vigtig rolle at spille i en kultur og et samfund, ikke kun kommunikationsmæssigt, men også symbolsk. Det er ikke tomme ord når man siger at ordbøger er centrale for identitet og kulturarv, for bevarelse og stadig levendegørelse af ikke

kun hukommelse, tanker og idéer, men også af konkret praksis og handling.

Efter den tematiske del følger først en anmeldelse og et mindeord. Tor Erik Jenstad anmelder Ove Arild Orvik: *Nordnorsk ordbok. Arven etter Hallfrid Christiansen*, en ordbog over det større, nordnorske område, med historiske perspektiver. Og Lars Svensson skriver et oplysende mindeord om Lars Holm, som optrådte adskillige gange ved NFL's leksikografikonferencer, og som ikke bare var en fremragende historisk leksikograf, men også ydede en stor indsats som tekstudgiver.

Årets nummer afsluttes med en rapport fra Nordisk Forening for Leksikografi ved formand for bestyrelsen, Thomas Widmann, Dansk Sprognævn.

Redaktionen af dette nummer består af de to hovedredaktører Henrik Hovmark og Terje Svardal (nytiltrådt), samt landsredaktørerne Liisa Deth Theilgaard (Danmark, nytiltrådt), Caroline Sandström (Finland, nytiltrådt), Helga Hilmisdóttir (Island, nytiltrådt), Kjetil Gundersen (Norge) og Louise Holmer (Sverige, nytiltrådt).

Temaerne for de to kommende *LexicoNordica*-symposier bliver som følger:

2025: Nordiske ordbøger – opdatering, udvikling og tilgængeliggørelse
2026: Brugerundersøgelser og -involvering i nordisk leksikografi

Alle er velkomne til at komme med forslag til foredrag ved de to symposier, samt med idéer til kommende temaer. Nærmere informationer vil blive annonceret på Nordisk Forening for Leksikografis hjemmeside og i foreningens nyhedsbrev.

Til slut vil vi gerne rette en stor tak til landsredaktørerne for deres meget store indsats i løbet af hele året, og ligeledes en stor tak til Anna Helga Hannesdóttir der fratrådte som hovedredak-

tør sidste år efter et stort, omhyggeligt og yderst kompetent arbejde for tidsskrift og forening. Også tak til bestyrelsen for Nordisk Forening for Leksikografi for godt samarbejde, især til formanden Thomas Widmann, og til kassereren Pär Nilsson som har ydet stor og uvurderlig hjælp i forbindelse med ansøgninger, kommunikationsopgaver og den praktiske gennemførelse af symposiet på Voksenåsen. En varm tak skal også rettes til Laurids Kristian Fahl som endnu engang har påtaget sig opgaven med opsætning og distribution af årets nummer – og har gjort det både omhyggeligt og professionelt. Dette arbejde er af uvurderlig betydning for redaktionen og for Nordisk Forening for Leksikografi. Endelig skal vi takke Voksenåsen kulturcentrum for at huse symposiet og Nordplus Nordens Sprog for velvillig og vigtig støtte til hele projektet: symposium og efterfølgende udgivelse af resultaterne i det nummer af *LexicoNordica* som nu foreligger.

Henrik Hovmark
lektor, ph.d.
Institut for Nordiske Studier og Sprogvidenskab
Københavns Universitet
Emil Holms Kanal 2
DK-2300 København S
hovmark@hum.ku.dk

Terje Svardal
leksikograf
Språksamlingane
Universitetsbiblioteket
Universitetet i Bergen
Haakon Shetligs plass 7
NO-5007 Bergen
terje.svardal@uib.no