LexicoNordica

LexicoNordica

31 · 2024

HVILKET DATAMATERIALE BYGGER NORDISKE ORDBØGER PÅ? SKÆVHEDER, UDFORDRINGER OG LØSNINGER

NORDISK FORENING FOR LEKSIKOGRAFI

LexicoNordica 31 · 2024
Hvilket datamateriale bygger nordiske ordbøger på?
Skævheder, udfordringer og løsninger

Hovedredaktører
Henrik Hovmark
Terje Svardal

Redaktionskomité
Kjetil Gundersen
Helga Hilmisdóttir
Louise Holmer
Caroline Sandström
Liisa Deth Theilgaard

Omslag og sats: Laurids Kristian Fahl
Trykt hos: Tarm Bogtryk a-s, Danmark

Voksenåsen kulturcentrum takkes for at huse
LexicoNordica-symposium 31

LexicoNordica trykkes med økonomisk støtte fra
Nordplus Nordens Sprog

ISSN 0805-2735
ISSN 1891-2206 (online)

Indhold

Henrik Hovmark & Terje Svardal
Hvilket datamateriale bygger nordiske ordbøger på? Skævheder, udfordringer og løsninger 7

Tematiske bidrag

Magnus Ahltorp, Lina Lejdebro Enwald, Elina Kangas, Jacob Larsson, Rickard Domeij & Gunnar Eriksson
Språkteknologi för att samla in texter och analysera språket i korpusverktyg – hur gör man på meänkieli? 21

Kirsten Appel, Nathalie Hau Sørensen & Jonas Jensen
Jagten på hverdagssproget – brugen af tekster fra internetfora i arbejdet med *Den Danske Ordbog* 39

Markus Forsberg & Louise Holmer
Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text 61

Tarja Riitta Heinonen & Caroline Sandström
Skevheter och utmaningar i ordboksbasen för fyra enspråkiga finländska ordböcker 81

Helga Hilmisdóttir
Talspråkskorpusar som resurs för isländska ordböcker 103

Ellert Þór Jóhannsson & Þórdís Úlfarsdóttir
Ordbog over moderne islandsk – udvikling og tilføjelser 127

Sanni Nimb
Fra 'sandheden om sproget' til et opgør med stereotyper: ikke-korpusbaserede metoder til leksikografiske beskrivelser af kontroversielle ord 151

Anna-Kaisa Räisänen, Aili Eriksen, Thomas Brevik Kjærstad & Trond Trosterud
Kvensk revitalisering, normering og leksikografi 175

Einar Freyr Sigurðsson & Steinþór Steingrímsson
Representativeness and biases in Icelandic corpora 201

Anmeldelse

Tor Erik Jenstad
Nordnorsk ordbok, den største i sitt slag 229

Mindeord

Lars Svensson
Minnesord om Lars Holm 243

Meddelelser

Thomas Widmann
Nyt fra bestyrelsen for Nordisk Forening for Leksikografi 253

Redaktionsanvisninger 257

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

TEMATISKE BIDRAG

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

Magnus Ahltorp, Lina Lejdebro Enwald, Elina Kangas, Jacob Larsson, Rickard Domeij & Gunnar Eriksson

The Language Council of Sweden at the Institute for Language and Folklore has a responsibility to support the development of language technology for the languages of Sweden. Together with the Department of National Minorities and Swedish Sign Language at the same institute, we are developing tools to collect and analyse texts in the national minority languages of Sweden. These texts will be available for use in research, for example in lexicographic work. We here present the specific challenges associated with building these tools for the Swedish national minority language Meänkieli.

1. Inledning

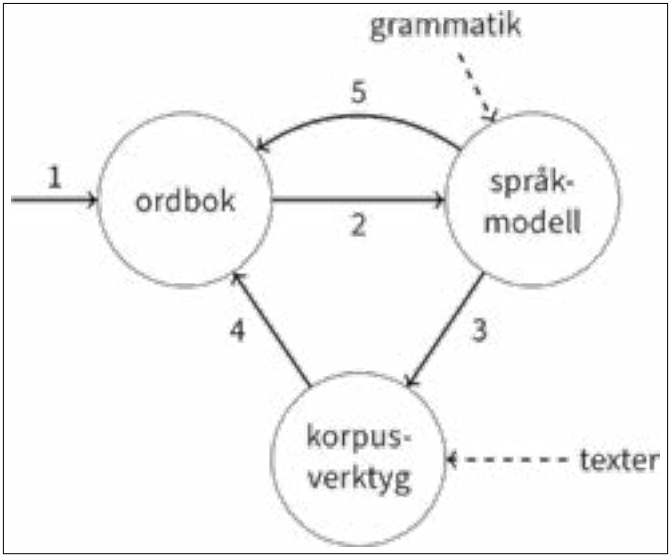
Institutet för språk och folkminnen (Isof) är en statlig myndighet vars uppgift är att samla in, bygga upp och sprida kunskap om språk och kultur i Sverige. Myndigheten ska bedriva språkvård och på vetenskaplig grund öka, levandegöra och sprida kunskaper om språk, dialekter, folkminnen, namn och immateriella kulturarv i Sverige. Språkvård för det svenska språket har bedrivits i semistatlig regi sedan Nämnden för svensk språkvård blev Svenska språknämnden i mitten av 1900-talet. För ett nationellt minoritetsspråk som meänkieli, som har sämre förutsättningar och betydligt mindre resurser än svenskan, har ett sådant uppdrag endast funnits på Isof sedan 2006.

År 2000 ratificerade Sverige två av Europarådets konventioner, *Ramkonventionen om skydd för de nationella minoriteterna* och *Den europeiska stadgan om landsdels- eller minoritetsspråk*. Där-

med fick finska, jiddisch, meänkieli, romska och samiska status som nationella minoritetsspråk i Sverige. I och med undertecknandet har Sverige lovat att trygga rättigheter kopplade till nationella minoritetsspråk.

År 2009 fick Sverige en språklag som slår fast att svenska är huvudspråk i Sverige, men som också säger att det allmänna har ett särskilt ansvar för att skydda och främja de nationella minoritetsspråken. Isof har därför i uppgift att stötta meänkieli, liksom övriga nationella minoritetsspråk, och som ett led i detta att bidra till att skapa en infrastruktur för språkteknologi för språket.

Med ett litet antal språkbrukare, se avsnitt 3 nedan, följer problemet att det saknas kommersiellt intresse för att konstruera språkteknologiska verktyg som stavningskontroll, korpusverktyg eller talsyntes, vilket gör Isofs uppdrag än viktigare (Mattson & Ahltorp 2023; Trosterud 2022).



Figur 1: Översikt över språkteknologisk infrastruktur för meänkieli. Siffrorna visar de olika relationerna komponenterna har till varandra.

För att kunna använda korpusar för att förbättra ordböcker måste det först finnas korpusverktyg. Ett korpusverktyg som är lämpligt för lexikografiska undersökningar förutsätter språkmodeller, som i sin tur förutsätter ordböcker (se figur 1). Med språkmodell menas här ett sätt att automatiskt analysera morfologi, syntax, etc. för språket. Denna artikel beskriver Isofs arbete med korpusverktyg, språkmodeller och språkvård för meänkieli, och hur dessa är beroende av varandra.

2. Meänkieli som språk

Meänkieli tillhör den östersjöfinska grenen av de uraliska språken. De närmaste besläktade språken är finska, där de största likheterna finns i de nordfinska dialekterna, och kvänska, som har minoritetsspråksstatus i Norge. Meänkieli är också besläktat med exempelvis samiska, estniska och på längre håll ungerska.

Meänkieli är sällan inkluderat i exempelvis språkträd och kartor som illustrerar de uraliska språkens släktskap och utbredning, något som delvis beror på att meänkieli inte fick språkstatus förrän år 2000. Enligt Blokland (2024) saknas dock ofta meänkieli även i nyare akademiska källor, alternativt klassificeras som en dialekt av finskan, se t.ex. Unescos *Atlas of the world's languages in danger* (Moseley & Nicolas 2010).

Meänkieli delar många språkdrag med övriga uraliska språk, exempelvis ett rikt kasussystem, nekandeverb, avsaknad av genus, stadieväxling (kvantitativ eller kvalitativ förändring hos stamkonsonanter) och vokalharmoni (främre och bakre vokaler förekommer inte i samma ord). Vad gäller det sistnämnda skiljer det sig emellertid åt mellan olika geografiska varieteter – i den meänkieli som talas kring Malmfälten frångås ibland principen med vokalharmoni.

Allra flest drag delar meänkieli med kvänskan och nordfinska dialekter, och meänkieli, kvänska och finska är sinsemellan hu-

vudsakligen ömsesidigt förståeliga. Meänkieli har emellertid stått under stark påverkan av svenskan, medan kvänskan istället påverkats av norskan. Varken meänkieli eller kvänskan har varit en del av standardiseringsprocessen för finska i Finland, och båda språken har bevarat en del språkdrag som tidigare funnits eller finns i finska dialekter men som inte tagits upp i det finska skriftspråket.

Skillnaderna mellan meänkieli och standardfinska är tydliga exempelvis vad gäller personliga pronomen, verbböjningar, inskott av h-ljud i efterstavelser samt i meänkielis många moderna låneord från svenska.

Exempel på skillnader mot kvänska är vissa kasusändelser, verbböjning och kvänskans dentala tonande frikativa /ð/, ett ljud som tidigare förekom även i meänkieli men som saknas i dagens språk (Paunonen 1987). Både kvänska och meänkieli har generell stadieväxling, men kvänskan har dessutom så kallad specialstadieväxling (t ex *poika* 'pojke' i nominativ singular blir *pojan* 'pojkens' i genitiv singular men *poikkaa* i partitiv singular, en form som saknas i meänkieli).

Precis som många andra uraliska språk har också meänkieli omfattande morfofonologisk växling, vilket nedanstående exempel visar. Denna egenskap komplicerar konstruktionen av språkmodellen.

(1) a. kanala
hönsgård.SG.NOM
'hönsgård'

b. kanalo-i-ta
hönsgård-PL-PAR
'hönsgårdar'

(2) a. katto
tak.SG.NOM
'tak'

b. kato-le
tak-ALL
'till tak(et)'

(3) a. mettä
skog.SG.NOM
'skog'
b. mettä-le
skog-ALL
'till skog(en)'

Exempel (1) visar hur vokalförändring äger rum som en följd av pluralsuffix; det avslutande a:et i *kanala* blir *-o-* på grund av pluralmarkören *-i-*. Exempel (2) visar klassisk stadieväxling där dubbelklusilen *-tt-* blir enkel på grund av allativ-suffixet. Exempel (3) utgör ett undantag från denna regel, vilket beskrivs vidare i sektion 6.

3. Språkbrukare och revitalisering

Traditionellt språkområde för meänkieli är Tornedalen och Malmfälten, områden som idag tillhör de svenska kommunerna Gällivare, Haparanda, Kiruna, Pajala och Övertorneå, men på senare tid har många talare flyttat till andra delar av landet, i synnerhet till storstäderna, som en del av en generell flyttvåg från Norrland. Antalet meänkielitalare är svårbedömt eftersom Sverige inte samlar in sådan statistik – uppskattningarna varierar mellan 20 000 och 75 000, beroende på källa och vilka man valt att inkludera i definitionen (Parkvall 2015; Valijärvi et al. 2022).

Meänkielitalare har under lång tid tvingats gå igenom en språkbytesprocess vilket lett till att meänkieli ofta inte har förts vidare till nästa generation, och språket klassas idag som allvarligt

hotat (Laakso et al. 2016). De flesta talare är äldre och flerspråkiga; alla meänkielitalande är också svensktalande.

Mobiliseringen för ett erkännande av meänkieli som ett eget språk inleddes under tidigt 1980-tal, när meänkielitalande började använda språket i kultursammanhang som teater, musik och litteratur. Intresseföreningen Svenska Tornedalingars Riksförbund – Tornionlaaksolaiset bildades 1981. Den första meänkielispråkiga romanen kom 1985 (*Lyykeri* av Bengt Pohjanen), de första teaterpjäserna på meänkieli kom snart efter det.

Idag görs många åtgärdsförsök för att återta (revitalisera) språket, allt från mindre projekt framtagna av privatpersoner till strukturerade myndighetsuppdrag. Det kan handla om exempelvis språkläger, universitetskurser eller andra språkkurser, särskilda förskolor med inriktning mot meänkieli, språkmyndigheten Isofs revitaliseringsuppdrag, etc.

Isof hade under åren 2006–2013 ett arkivinriktat arbete för meänkieli, och en språkvårdstjänst tillsattes år 2018. Isofs arbete bygger vidare på det arbete som redan gjorts av talarna, exempelvis i form av den digitala ordbok (*Meänkieli–ruotti sanakirja*) som tagits fram av Meän Akateemi-Academia Tornedaliensis (*Meän Akateemis Meänkieli-svensk/svensk-meänkieli digital ordbok* 2024) med delfinansiering från Isof. Detta arbete utgör steg 1 i figur 1.

År 2022 fick Isof därutöver ett treårigt regeringsuppdrag att inrätta ett språkcentrum för meänkieli, finska, jiddisch och romska, i syfte att ge stöd och kunskap som underlättar för språkbärare i hela landet att behålla, ta tillbaka och utveckla sitt språk.

Allt detta innebär att det idag också tillkommit nya talare som inte har förkunskaper i språket hemifrån. Fortfarande råder dock stor brist på läromedel, utbildningar etc. – det hänger till stor del på den enskilda individen att hitta inlärningsmöjligheter.

Språkverktyg som stavnings- eller grammatikkontroll är under utveckling på Isof men finns ännu inte tillgängliga för allmänheten.

4. Särskilda utmaningar för meänkieli

En särskilt komplicerande faktor för språkvården för meänkieli är att det saknas omfattande deskriptiv grammatik. Språkvårdsarbetet på Isof får därför utgå från publikationer inom finsk dialektologiforskning (Mantila 1992; *Suomen murteiden sanakirja* 2020), finsk grammatik (Karlsson 2012), kvänsk grammatik (Söderholm 2017) samt modernare meänkielimaterial i form av exempelvis normativ grammatik (Pohjanen 2022) och meänkielikorpus (Korp för Meänkieli 2024). Vid sidan av detta är utgångspunkten lingvistiska observationer av vardagligt språkbruk i media, i Facebookgrupper, i kontakten med talare (exempelvis vid språkrådgivningssituationer) osv.

Meänkieli är till stor del ett talat språk; många har inte fått lära sig läsa och skriva på meänkieli (Valijärvi et al. 2022). Det finns en förväntan från talare på att få hjälpmedel för att skriva på meänkieli, i synnerhet bland dem som inte fått språket hemifrån, samtidigt som de normativa val som gjorts eller föreslås ofta kan ifrågasättas – som talare kanske man inte alltid identifierar sig eller sitt språk med en viss stavning eller ett visst ordval.

De normeringsförslag som finns sedan tidigare kommer i första hand via aktiva språkaktörer, Facebook-grupper, författare som Bengt Pohjanen som publicerat såväl romaner som språkmaterial i form av ordböcker och skrivregler (Pohjanen 2022), osv. Många av meänkielitalarna utgår inte från dessa normeringsförslag i skrivprocessen, även om man känner till att de finns.

Det mesta som är skrivet på meänkieli är baserat på varieteten som talas i Tornedalen (Kuoppa 2015), trots att exempelvis vokalharmoni, verbböjning och många andra drag uppvisar regional variation. Skälet till detta är att antalet talare från andra områden är få, relativt sett, och att skribenter från dessa geografiska områden är ännu färre – de aktiva skribenterna kommer kort sagt från Tornedalen.

Meänkieli uppvisar dessutom andra typer av variation som språkvården måste hantera än den geografiska; till exempel den mellan äldre/yngre talare, förstaspråks-/andraspråkstalare, uttalsskillnader och skillnaden mellan svenska och finska låneord.

Dessa komplicerande faktorer, i kombination med det faktum att meänkielikorpusen är mycket begränsad (se vidare under avsnitt 7), gör att arbetet med att stödja språket måste bedrivas på två fronter samtidigt; språknormer tas fram samtidigt som språkteknologin utvecklas, och i någon mån tvingar arbetet med språkverktyg också normeringen vidare – för att kunna göra exempelvis en stavningskontroll måste ortografin klargöras, och i och med det upptäcks fenomen som sedan behöver få en lingvistisk förklaring.

5. Språkmodellers betydelse för korpussökningar

Till skillnad från meänkieli har de stora nordiska språken haft både lexikografiska och språkteknologiska resurser under lång tid. Svenska har exempelvis haft ordböcker sedan åtminstone 1712 (Ralph 2009). Språkteknologiska verktyg för svenska var själva grunden till arbetet med *Nusvensk frekvensordbok* (Allén 1970), ett arbete som påbörjades på 1960-talet. Arbetet med ordböcker, språkmodeller och korpusverktyg har utvecklats parallellt under lång tid. För de stora nordiska språken har verktyg för stavningskontroll funnits brett tillgängliga sedan åtminstone tidigt 1990-tal. Dessa språk har en väl utvecklad språkvetenskaplig och språkteknologisk infrastruktur där insamlade texter och språkmodeller kan kombineras i verktyg som analysverktyget Sparv (Hammarstedt et al. 2022) och korpusverktyget Korp (Borin et al. 2012).

Enkla sökningar kan göras i korpusverktyg utan språkmodeller, men är begränsade till exakta strängar eller trunkering

(sökning efter delsträngar). Exempelvis hittar en exakt sökning efter strängen ”bära” i en svensk korpus bara den exakta formen. Vanligtvis kan en sådan sökning även hitta formerna ”Bära” och ”BÄRA”. Även om sökning oberoende av versaler och gemener kan tyckas trivialt är redan det en form av språkmodell, något som är långt ifrån språkoberoende. Enkla exempel är tyskans <ß> som behöver transformeras till <ss>, och turkiskans ortografi där den gemena formen av <I> inte är <i>, utan <ı>. Ännu mer komplicerat blir det i språk som har stor stavningsvariation. Ett extremt exempel är japanska, där den som skriver i de flesta fall kan välja vilka delar av ett ord som ska skrivas med logografiska tecken och uttalstecken, något som förekommer i t.ex. svenskan, men i mycket begränsad omfattning, som i <8:e> och <åttonde>.

Med trunkering kan ytterligare former hittas. Vill någon söka efter olika former av verbet ”bära” kan en sökning efter ”bär*” korrekt hitta formerna ”bära” och ”bär”. Denna sökning kommer dock inte hitta formerna ”bar”, ”burit”, ”buren” (undergenerering), men de orelaterade strängarna ”bärplockning” och ”bärsärk” hittas också (övergenerering). Beroende på vilka former av verbet som är intressanta kan de hittade strängarna ”bäras” och ”bärande” vara önskade eller oönskade.

En mycket enkel sorts språkmodell är s.k. stemming, som till viss del hanterar böjningsformer utan att användaren själv behöver söka med trunkering eller själv mata in alla böjningsformer. Stemming använder mycket enkla regler för att ”skära av” ordet efter dess stam. Även här är undergenerering och övergenerering vanliga problem, och stemming är mer lämpat för språk med okomplicerad morfologi.

Även för språk där en stemmer fungerar någorlunda bra är den oftast bara lämplig för enklare sökningar, som i sökmotorer för webbplatser eller hela webben. Sökningar i korpusverktyg behöver däremot normalt utgå från en fullständig morfologisk analys. En

anledning, förutom att få en ordentlig lemmatisering, är att det även är intressant hur ordet är böjt. Sökningar kan då göras av typen ”alla ord som börjar på ’a’ och står i bestämd form”.

Lemmatisering som tar hänsyn till kontext ger även möjlighet att disambiguera strängar som ”bär” (antingen ’liten frukt’ eller imperativ av *bära*).

För att korpusar ska kunna användas för lexikografiskt arbete enligt steg 4 i figur 1 krävs alltså att en språkmodell för språket är tillgänglig för korpusverktyget (steg 3).

6. En språkmodell för meänkieli

Flera av de stora nordiska språken har omfattande resurser, i form av grammatiska beskrivningar, ordböcker som täcker en stor del av språket och lättillgänglig text författad direkt på språket. Stora mängder text, särskilt annoterad sådan, kan då användas för att träna maskininlärningsmodeller.

När det gäller lågresursspråk, där det inte finns tillräckliga mängder språkdata att träna maskininlärningsmodeller på, måste arbetet med att ta fram en språkmodell i stället göras manuellt och explicit (Lejdebro Enwald 2024; Trosterud 2022): grammatik och morfofonologiska växlingar måste formuleras som maskinläsbara regler. För detta brukar man använda sig av så kallade ändliga tillståndsautomater, Finite State Transducers (FST) med tvånivåmorfologi. Forskningsgrupperna GiellaTekno och Divvun vid Universitetet i Tromsø Norges arktiske universitet (UiT) har under de senaste decennierna tagit fram just sådana modeller för ett stort antal lågresursspråk med rik morfologi, många av dem minoritetsspråk, och samarbetar nu med Isof för att utveckla språkteknologi för meänkieli (Pirinen et al. 2023).

Ett exempel på en sådan regelbaserad språkmodell är den kvänska, som GiellaTekno/Divvun tagit fram i samarbete med

Kvensk Institutt (nationellt center för kvänskt språk och kvänsk kultur) och som också används som en stavningskontroll.

År 2020 tog man vid UiT även fram en grund för en meänkielimodell, baserad på den finska och kvänska språkmodellen men anpassad till meänkieli utifrån en grammatikbok (Kenttä & Pohjanen 1996) och den digitala ordboken (*Meän Akateemis Meänkieli-svensk/svensk-meänkieli digital ordbok* 2024; Trosterud 2020). Ordboken är helt nödvändig indata till språkmodellen (steg 2 i figur 1).

Syftet med språkmodellen är att kunna analysera och generera språkets ordformer. Får språkmodellen input i form av lemma och önskad grammatisk form, exempelvis *pilvi* 'moln', plus den grammatiska taggen "singular illativ", ska den generera korrekt böjd ordform: *pilvheen*, och omvänt; utifrån en böjd ordform ska språkmodellen kunna generera lemma och grammatisk information. Förutom att språkmodellen behövs för korpusverktyget är det viktigt även direkt i ordboken där den kan användas både för sökning och presentation (steg 5 i figur 1).

Isofs språkvårdare och språkteknologer arbetar nu med att vidareutveckla denna språkmodell för meänkieli. Exempel på specifika anpassningar som behöver göras för meänkieli, vid sidan av att korrigera böjningsparadigm och utöka modellens lexikala delar, är de h-metateser som frekvent uppstår i suffix med tonande konsonanter – det som ibland kallats "varumärket för meänkieli" (Kuoppa 2015). Ovan nämnda *pilvheen* kan på grund av detta metatesfenomen också bli *pilhveen*. De kontexter där detta sker har tidigare undersökts för finska dialekter och för ett antal meänkielitexter (Mantila 1992; Kuoppa 2015).

Andra exempel på meänkielianpassningar är de ord som utgör undantag från reglerna om stadieväxling; jämför exemplen (2) och (3) ovan. Normalt uppstår stadieväxling i samband med exempelvis allativ-suffix, men för ord vars motsvarighet i finskt skriftspråk har *-ts-*, exempelvis 'skog' som på finska heter *metsä*

men på meänkieli *mettä*, äger ingen stadieväxling rum. Här styrs ljudförändringen således inte av kontexten utan av det enskilda ordet, vilket kräver en annan typ av regler i språkmodellen än de som är kontextberoende.

Eftersom Isofs uppdrag är att bedriva språkarbetet på vetenskaplig grund, samtidigt som det saknas en deskriptiv grammatik för meänkieli, måste reglerna i språkmodellen tills vidare vara så generösa att en framtida stavningskontroll som bygger på modellen tillåter flera stavningsvarianter.

7. Insamling av texter

Som tidigare nämnts har Isof ett ansvar att bedriva verksamheten på *vetenskaplig grund*. För att kunna göra detta behöver det finnas ett språkmaterial med hög kvalitet som verksamheten kan grunda sig på. Det vore önskvärt med en korpus som är så representativ som möjligt, som täcker flera genrer som skönlitteratur, tidningsprosa, sociala medier m.m. Ett problem Isof dock har mött rätt tidigt i framtagningen av en sådan korpus är att antalet skribenter av publicerad text på meänkieli är lågt.

Arbetet har därför varit tvunget att bli mer pragmatiskt, och fokusera på så stora textmängder som möjligt istället för att fokusera på en bred representation. Det innebär också att större flexibilitet behövs när det gäller källorna till korpusarna och vilka metoder som används i insamlingen av detta material.

Den låga mängden aktiva skribenter innebär att flera vanliga källor för textinsamling, såsom sociala medier och forum, inte är tillgängliga för meänkieli. Dessutom finns det inte en lika lång tradition av att skriva skönlitteratur på meänkieli som det gör för exempelvis svenska, vilket ytterligare begränsar tillgången på textmaterial från denna domän.

En av de största producenterna av texter på meänkieli idag är den offentliga förvaltningen, inklusive kommuner, regioner och myndigheter. En betydande del av denna textmängd är publicerad digitalt och finns spridd över flera olika hemsidor med varierande strukturer och format. Ofta finns det endast enstaka sidor skrivna på meänkieli (Meänkielelä – kiruna.se 2024), medan mer omfattande översättningar förekommer hos andra organisationer (Meänkieli (Tornedalsfinska) – arbetsformedlingen.se 2024).

Man behöver även ha i åtanke att i många fall förekommer dessa texter parallellt med texter på övriga minoritetsspråk. Detta är särskilt besvärande när det gäller finska, som oftast är svårt att automatiskt avskilja från meänkieli (Larsson 2024) och där det är viktigt att undvika att ta med finska texter i korpusen. Att få med finska texter i korpusen riskerar att skapa ett normativt tryck på meänkieli när denna korpus senare används i språkvårdsyften.

Denna fragmentering och variation av texterna på meänkieli utgör ytterligare en utmaning för insamling av material. I andra språk där det hade gått att samla in en stor och representativ korpus med hjälp av texter från några enstaka källor (t.ex. Familjeliv eller Flashback som hos Språkbanken Text) med samma struktur behöver Isof istället ta fram verktyg som kan både samla in brett (över flera olika hemsidor), djupt (texterna ligger långt inbäddade i hemsidorna) och med hög precision så att inte material på andra språk inkluderas.

Allt detta kräver att Isof utvecklar nya anpassningsbara metoder för textinsamling som tar med dessa förutsättningar i ett tidigt stadium av dess utveckling. Ett sådant arbete pågår idag, men innebär även att det inte alltid går att nyttja befintliga verktyg och metoder för textinsamling som är mer anpassade för större språk med en mer etablerad digital närvaro.

8. Avslutande ord

Utan språkteknologisk infrastruktur kan Isof inte utföra sitt uppdrag att bedriva språkvård på vetenskaplig grund för meänkieli, eller för den delen de andra nationella minoritetsspråken. Att kunna använda korpusar för att förbättra ordböcker kräver ändamålsenliga korpusverktyg. Detta förutsätter språkmodeller, som i sin tur förutsätter lexikografiska resurser. Detta kan ses som en grundläggande del av arbetet med standardisering och revitalisering av de nationella minoritetsspråken.

Litteratur

Ordböcker och digitala resurser

Korp för meänkieli – GiellaTekno/Språkbanken Text. <gtweb.uit.no/f_korp> (juni 2024).

Meän akateemis meänkieli-svensk/svensk-meänkieli digital ordbok (2024). <språk.isof.se/meänkieli/> (april 2024).

Meänkielelä – kiruna.se (2024). <kiruna.se/kommun--demokrati/vara -minoritetssprak/meankielela.html> (april 2024).

Meänkieli (Tornedalsfinska) – arbetsformedlingen.se (2024). <arbetsformedlingen.se/other-languages/meankieli-tornedalsfinska> (april 2024).

Suomen murteiden sanakirja (2020): *Kotimaisten kielten keskuksen verkkojulkaisuja*. <kaino.kotus.fi/sms> (april 2024).

Annan litteratur

Allén, Sture (1970): *Nusvensk frekvensordbok baserad på tidningstext 1 Graford Homografkomponenter*. Stockholm: Almqvist & Wiksell.

Blokland, Rogier (2024): Meänkieli som ett uraliskt språk. I: *Föreläsning med anledning av meänkielidagen 27 februari*. Isof.

Borin, Lars, Markus Forsberg, Johan Roxendal (2012): Korp – the corpus infrastructure of Språkbanken. I: *Proceedings of LREC 2012*. Istanbul: ELRA. 474–478.

Hammarstedt, Martin, Anne Schumacher, Lars Borin & Markus Forsberg (2022): *Sparv 5 user manual*. Göteborg: Göteborgs universitet.

Karlsson, Fred (2012): *Finsk grammatik. Nionde, utökade och reviderade upplagan*. Helsingfors: Suomalaisen Kirjallisuuden Seura.

Kenttä, Matti & Bengt Pohjanen (1996): *Meänkielen kramatiikki*. Övertorneå: Kaamos.

Kuoppa, Harriet (2015): *Varumärket för meänkieli – användningen av h i efterstavelsen i skrift*. Umeå: Umeå universitet.

Laakso, Johanna, Anneli Sarhimaa, Athanasia Spiliopoulou Åkermark & Reetta Toivanen (2016): *Towards openly multilingual policies and practices: Assessing minority language maintenance across Europe*. Bristol: Multilingual Matters.

Larsson, Jacob (2024): *Language identification for typologically similar low-resource languages – case study of Meänkieli, Kven and Finnish*. Stockholm: Stockholms universitet.

Lejdebro Enwald, Lina (2024): *En morfologisk modell för meänkieli*. Stockholm: Stockholms universitet.

Mantila, Harri (1992): *Ei tääläkhän senthän jokhaishen sanhan hootakhan panna*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Mattson, Marie & Magnus Ahltorp (2023): Lexikografiska resursers betydelse i utvecklingen av språkteknologiska verktyg för minoritetsspråk. I: *LexicoNordica* 30, 75–94.

Moseley, Christopher & Alexandre Nicolas (2010): *Atlas of the world's languages in danger*. Paris: Unesco.

Parkvall, Mikael (2016): *Sveriges språk i siffror: vilka språk talas och av hur många?* Stockholm: Morfem.

Paunonen, Heikki (1987): De finska dialekterna på Nordkalotten och deras förhållande till de andra finska dialekterna. I: *Studia Historica Septentrionalia nr 14: Nordkalotten i en skiftande värld*. 211–237.

Pirinen, Flammie, Sjur Moshagen & Katri Hiovain-Asikainen (2023): GiellaLT — a stable infrastructure for Nordic minority languages and beyond. I: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn: University of Tartu Library. 643–649.

Pohjanen, Bengt (2022): *Meänkieli: grammatik, lärobok, historik, texter*. Överkalix: Barents.

Ralph, Bo (2009): När ordboken blev en ordlista. I: Martin Gellerstam (red.): *SAOL och tidens flykt: några nedslag i ordlistans historia*. Stockholm: Norstedts.

Söderholm, Eira (2017): *Kvensk grammatikk*. Oslo: Cappelen Damm Akademisk.

Trosterud, Trond (2020): *Språkteknologi för meänkieli*. <giellalt.github.io/lang-fit/ rapport.pdf> (april 2024).

Trosterud, Trond (2022): Normative Language Work in the Age of Machine Learning. I: Željko Jozić & Sabine Kirchmeier (red.): *The role of national language institutions in the digital age*. Budapest: Hungarian Research centre for Linguistics. 61–70.

Valijärvi, Riitta-Liisa, Rogier Blokland, Elina Kangas, Constanze Ackermann-Boström & Harriet Kuoppa (2022): Meänkieli. *Linguistic Minorities in Europe Online*. doi:10.1515/lme.18469972.

Magnus Ahltorp
Språkteknolog
Språkrådet, Institutet för språk och folkminnen
Alsnögatan 7
SE-116 41 Stockholm
magnus.ahltorp@isof.se

Lina Lejdebro Enwald
Språkteknolog
Språkrådet, Institutet för språk och folkminnen
Alsnögatan 7
SE-116 41 Stockholm
lina.lejdebro.enwald@isof.se

Elina Kangas
Språkvårdare i meänkieli
Avdelningen för nationella minoriteter och svenskt teckenspråk, Institutet för språk och folkminnen
Alsnögatan 7
SE-116 41 Stockholm
elina.kangas@isof.se

Jacob Larsson
Språkteknolog
Avdelningen för nationella minoriteter och svenskt teckenspråk, Institutet för språk och folkminnen
Alsnögatan 7
SE-116 41 Stockholm
jacob.larsson@isof.se

Rickard Domeij
Språkteknolog
Språkrådet, Institutet för språk och folkminnen
Alsnögatan 7
SE-116 41 Stockholm
rickard.domeij@isof.se

Gunnar Eriksson
Språkteknolog
Språkrådet, Institutet för språk och folkminnen
Alsnögatan 7
SE-116 41 Stockholm
gunnar.eriksson@isof.se

Jagten på hverdagssproget – brugen af tekster fra internetfora i arbejdet med *Den Danske Ordbog*

Kirsten Appel, Nathalie Hau Sørensen & Jonas Jensen

This paper introduces methods and tools designed to alleviate the issues imposed by an increasingly uniform news wire corpus. The purpose is to help the editors of *The Danish Dictionary* (DDO) find common lemmas from everyday language which are underrepresented in the current corpus due to genre conventions and news values. The primary new tool, *Findor den yngre*, relies on a number of methods to filter the contents of a newly compiled corpus of chat room texts. The paper concludes that this complex filtering is superior to pure frequency-based methods when it comes to lemma selection, and that *Findor* and the new corpus provide editors with a more nuanced representation of contemporary Danish in general and a better selection of lemma candidates in particular.

1. Indledning

Den computerbaserede korpuslingvistiks fremkomst udgjorde et kvantespring i sprogforskningen og udarbejdelsen af ordbøger. Farvel til møjsommelig excerpering af belæg til alfabetisering i velordnede kartotekskasser, og goddag til søgbare databaser fulde af autentisk sprog, konkordanser der kunne sorteres efter ønske, og et overblik over sprogbrugen som man hidtil kun havde kunnet drømme om. Således var tidligere generationer af leksikografer lovligt undskyldt hvis de, grebet af tidsånden, antog at blot man havde et korpus, kunne man uden videre finde den objektive sandhed om sproget.

I tilfældet *Den Danske Ordbog* (DDO) havde man på sin vis mere at have det i da det redaktionelle arbejde indledtes i 1990'erne,

end det er tilfældet i dag: Ganske vist indeholdt DDO's oprindelige korpus blot 40 millioner ord, men det bød til gengæld på både tale- og skriftsprogsmateriale fra et bredt udvalg af kilder og genrer (dagbøger, nyhedsstof, skønlitteratur, transskriberet talesprog m.m.). Nutidens korpus er med sine 1,2 milliarder løbende ord en kæmpe i sammenligning, men hvad det har i størrelse, savner det i mangfoldighed.

Disse begrænsninger har givet anledning til at spørge om det overhovedet er muligt på grundlag af det nuværende korpus at "beskrive sproget sådan som det tales og skrives af et bredt udsnit af den danske befolkning" (*Fakta om DDO*), som var DDO's oprindelige målsætning. Det er desuden en overvejelse værd hvordan skævheden i korpus hidtil har påvirket DDO's lemmaselektion. Mest væsentligt i den aktuelle kontekst er imidlertid spørgsmålet: Hvordan kan en kombination af sprogteknologi og alternativer til det eksisterende korpus give redaktionen et mere nuanceret billede af "sproget sådan som det tales og skrives af et bredt udsnit af den danske befolkning"?

Denne artikels sigte er at besvare ovenstående spørgsmål gennem en præsentation og analyse af værktøjet *Findor den yngre*, der er udviklet netop med henblik på fremsøgning af lemmakandidater fra hverdagssproget i et til formålet kompileret korpus med tekster fra chatfora på internettet. For at forstå formålet med *Findor* er det nødvendigt at kende mere til dels det eksisterende korpus og dels baggrunden for første iteration af algoritmen, *Findor den ældre*. Begge dele beskrives i afsnit 2 nedenfor.

2. Baggrund

2.1. Korpus

Siden 2001 er det oprindelige, 40 millioner ord store DDO-korpus

løbende blevet udbygget med nye tekster, i altovervejende grad fra nyhedsmedier.

Resultatet er, som nævnt i afsnit 1, et korpus der er langt større end i den trykte ordbogs tid, men uden samme mangfoldighed. Det betyder at nogle sprogbrugere, fx børn og unge, er meget dårligt repræsenteret i vores korpus, mens professionelle sprogbrugere, især journalister, er kraftigt overrepræsenteret. Også korpussets ordvalg og emnemæssige sammensætning bærer præg af at teksterne er underlagt særlige krav til aktualitet og register. Trods adskillige metoder til fremfinding af lemmakandidater (bl.a. *Årets Ord*, *Månedens Ord*, brugerforslag, komposita og afledninger fra den trykte ordbog samt ord der optræder i redaktionel tekst) er det blevet tydeligt at der er ord og emner der er gået under radaren i den hidtidige lemmaselektion. Samtidig kan vi konstatere at ikke alt det vi faktisk finder, er lige velegnet til inklusion i DDO. Figur 1, som viser *Månedens Ord* fra 2023, illustrerer problemet, idet emner som krig og terror (fx *infanterikampkøretøj* og *terrorsag*) fylder uforholdsmæssigt meget, mens hverdagsemner (fx *læseglæde*) fylder tilsvarende lidt.

2.2. Automatiske metoder – hvad har andre prøvet?

I jagten på de tidligere oversete lemmakandidater i vores avistunge korpus blev den automatiske metode *Findor den ældre* (Sørensen et al. 2023) udviklet i 2023. Metoden viste at man med de rette sprogteknologiske værktøjer kan finde gode lemmakandidater blandt de lav- og mellemfrekvente ord i et aviskorpus.

Inspirationen til *Findor den ældre* kommer fra automatisk detektion af neologismer (Kerremans, Stegmayr & Schmid 2012, Falk, Bernhard & Gérard 2014, Langemets et al. 2020, Halskov & Jarvad 2010). Målet – at fremfinde henholdsvis lemmakandidater og neologismer – er omtrent det samme, dog med den væsentlige forskel at detektion af neologismer opererer med strengere kriterier for

Figur 1: Ord der optrådte på den frekvensbaserede liste over *Månedens Ord* i løbet af 2023. *Månedens Ord* udtrækkes ved at måle overhyppighed sammenlignet med andre måneder. Ordet *koranafbrænding* optræder tre gange, *koranlov*, *læseglæde* og *mobilitetsplan* optræder hver to gange. Resten af ordene optræder en enkelt gang.

hvad en god kandidat er. I stedet for kun at lede efter neologismer søger vi bredere efter ord som ikke nødvendigvis er nye i sproget, men som endnu ikke er opdaget af de frekvensbaserede værktøjer.

En bredt anvendt metode til automatisk detektion af neologismer er brugen af eksklusionslister. De bruges til at frasortere ord som man på forhånd ved ikke har interesse – fx ord som allerede findes i ordbogen, navnelister og stavefejl. Efter frasorteringen ved hjælp af eksklusionslisterne kan man så filtrere og postprocessere de resterende data til man får en mere overskuelig kandidatliste. Den største udfordring ved denne metode er effektiv frasortering

af støj i data som især kommer fra proprier, stavefejl, tokeniseringsfejl og gennemskuelige komposita.

I *Findor den ældre* omgås støjen gennem en lemmascore. Lemmascoren kombinerer information fra en række nøje konstruerede karaktertræk, også kaldet features, som er relevante for lemmaselektionen. Metoden kommer fra maskinlæringens feature engineering, hvor man bruger ekspertviden til at udvælge og udtrække en række egenskaber eller karakteristika om et fænomen som man vil modellere. Normalt vil man bruge features til at lave et datasæt man kan træne en model på. Denne metode er fx benyttet i Falk, Bernhard & Gérard (2014). Det kræver dog de rigtige data at træne en pålidelig model. Vi har positive eksempler fra online DDO, men mangler et tilstrækkeligt antal gode negative eksempler, det vil sige eksempler på ord der ikke er relevante for ordbogen (fx propriet *Juventus* eller den gennemskuelige sammensætning *hestenavn*). At en type ikke er med i DDO nu, er ikke et godt nok kriterium, da typen kan være en overset lemmakandidat. *Findor den ældre* brugte en mere simpel metode, nemlig et vægtet gennemsnit med nøje manuelt justerede vægte, som viste sig at være tilstrækkelig effektiv. Vi har nu videreudviklet *Findor* og præsenterer i denne artikel *Findor den yngre* som i stedet for avistekster er målrettet chatforumtekster.

3. Metode

3.1. Indsamling af internettekster

Som nævnt i afsnit 2.1 udgør den store overvægt af avistekster en udfordring ved vores nuværende korpus, og vi har derfor sat os for at finde tekster der både er frit tilgængelige, og som kan udgøre en modvægt til det eksisterende korpus.

Vi mangler først og fremmest hverdagssprog – ting vi taler om

sammen, men som ikke nødvendigvis lever op til nyhedskriterierne, og som derfor sjældent kan genfindes i avisartikler. En oplagt kilde til dette sprog er åbne internetfora hvor der typisk skrives i et uformelt sprog om hverdagsproblemstillinger, fx hvordan man undgår at blive syg i vinterperioden, eller hvilken tørretumbler andre kan anbefale.

Sproget på internetfora adskiller sig fra avisteksternes på flere områder. For det første er brugerne ikke professionelle skribenter, og teksterne er i mindre omfang redigeret og korrekturlæst, og vi må derfor forvente at se flere stavefejl. Teksterne er desuden kortere, mere uformelle, og de lægger op til interaktion. Vi ser fx en del emojis og slang i data. Teksterne bærer præg af at handle om hvad skribenten har på hjerte nu og her, altså øjeblikstanker, i modsætning til avisteksternes krav om relevans for en bred læserskare.

Da internetfora ligger frit tilgængelige på internettet, kan data derfra høstes via web scraping (det vil sige automatisk indsamling af internettekster). Vi har identificeret syv forskellige internetfora med hverdagssprog (Hestenettet, baby.dk, bold.dk, HardwareOnline, Pokernet, debatten.net og reddit.com/r/denmark). Hvert forum indeholder tekster fra perioden 2005-2023. Derudover har vi udvalgt specifikke subfora med en forventning om at de indeholder meget hverdagssprog. Disse subfora har typisk overskrifter som "fri-snak-fredag", "hyggesnak", "off-topic", "generelt" osv. Vi har bestræbt os på at indhente tekst fra fora med forskellige emneområder, fx heste, graviditet, sport og teknik. Til hvert forum har vi skræddersyet et pythonscript der gemmer indlæg og kommentarer som separate tekstfiler med følgende metadata: id, tekstens titel, selve teksten, sektion, url, type, dato, parent id og parent url. De to sidstnævnte sikrer at vi kan genskabe indlæggets kontekst. I alt har det resulteret i et korpus på 174 millioner løbende ord.

3.2. Værktøjet *Findor den yngre*

Findor den yngre bygger på samme arkitektur som sin forgænger, *Findor den ældre*, nemlig de tre faser: 1) præprocessering af korpus, 2) beregning af en score for "lemmahed" (altså egnethed som lemma i DDO) og 3) sortering efter denne score. Forskellen på de to iterationer af *Findor* er datagrundlaget (avistekster versus chatforumtekster) og indmaden i fase 1 og 2. Vi vil i det følgende kun gå i dybden med opbygningen af *Findor den yngre*, og betegnelsen *Findor* vil derfor referere til den nyeste version. For en grundig beskrivelse af *Findor den ældre* henviser vi til Sørensen et al. (2023).

For at kunne undersøge kvaliteten af *Findor* deler vi DDO's materiale op i det der optrådte i den trykte udgave (2002-2005), og det som siden er blevet føjet til onlineudgaven (2006-). I opbygningen af *Findor* bruger vi kun information tilgængelig i den trykte udgave og gemmer derfor alle opdateringer som en "guldstandard" der kan bruges som led i evalueringen.

3.2.1. Fase 1: Præprocessering

Formålet med præprocesseringen er at gå fra korpus i rå tekst til en bruttoliste med lemmakandidater. Det vil sige at ikke alle ord på listen nødvendigvis er gode lemmakandidater, men at vi sorterer alt fra som vi allerede ved ikke har interesse for DDO. Da det kan diskuteres om alt indhold på listen kan kaldes ord eller lemmaer, vil vi bruge "tokens" om enkelte eksempler på vilkårlige tekststrenge, "typer" om en samlet gruppe af tokens med samme form, og "kandidater" om typer som er behandlet af *Findor*.

Præprocesseringen tager udgangspunkt i et årsopdelt korpus (årgangene 2005-2023). Herefter udfører vi en simpel tokenisering ved at adskille teksten med mellemrum. Vi renser også data yderligere ved at fjerne URI'er, transformere store bogstaver til små og ved at fjerne tokens som indeholder ugyldige tegn (fx ه¤þü)[1]. Vi

1 Gyldige tegn = abcdeéfghijklmnopqrstuvwxyzæøå0123456789-&/.´"¹²³⁴⁵⁶⁷⁸⁹⁰₀₁₂₃₄₅₆₇₈₉ₓ

fjerner også al tegnsætning undtagen bindestreg, som bruges til at rense tokeniseringsfejl. Herefter grupperer vi tokens som typer og optæller deres frekvens for hvert år og i hele korpusset. Frekvensoptællingerne tillader os at fjerne alle typer som har en frekvens på mindre end fem. Denne liste med typer og frekvenser udgør den første kandidatliste.

På dette stadie indeholder kandidatlisten stadig uinteressante typer, det vil sige enten støj, navne eller ord som allerede er i den trykte ordbog, og som derfor ikke skal tilføjes til DDO. Formålet med det næste trin er derfor at fjerne så mange uinteressante typer som muligt. Vi fjerner derfor i første omgang alle typer som indeholder tal, eller som starter med en bindestreg. Dernæst fjerner vi alle typer som forekommer på én af fem forskellige eksklusionslister: en fuldformsliste fra den trykte ordbog og fire lister med fornavne (mandlige og kvindelige), efternavne og stednavne som er registreret af Danmarks Statistik. Da flere typer kan være eksempler på det samme ord i forskellige bøjninger, bruger vi også værktøjet *CSTLEMMA* til automatisk at lemmatisere typerne og gruppere dem igen. Til sidst fjerner vi alle typer som optræder i færre end tre årgange. En oversigt over antallet af typer efter hvert rensningstrin kan ses i tabel 1:

Trin	Fjernet	Antal på kandidatliste
Første kandidatliste		335.401
Fjern tal og bindestreg	7.494	327.907
I den trykte DDO	122.482	205.425
Findes på navnelister	23.340	182.085
Lemmatisering	21.263	160.822
Mindre end tre årgange	25.491	135.331

Tabel 1: Antal typer der resterer efter de respektive trin i præprocesseringen.

3.2.2. Fase 2: Udregning af "lemmahed"

Kernen i *Findor* er en måling af "lemmahed", som vi kalder lemmascoren. Lemmascoren er et vægtet gennemsnit af flere delscorer som hver især afspejler et karaktertræk ('feature') der kan være relevant for lemmaselektionen. Det er blandt andet stabilitet over tid, overensstemmelse med dansk ortografi og morfologi samt semantisk lighed med lemmaer som allerede er i DDO. I denne fase bliver ingen typer fjernet på baggrund af et enkelt kriterium som i præprocesseringen. I stedet skaber det vægtede gennemsnit et samlet billede på tværs af delscorerne, så en dårlig score ét sted kan opvejes af flere gode scorer andre steder. I det følgende forklarer vi baggrunden for hver delscore, og hvordan den er udregnet.

Stabilitet over tid

Et kriterium i DDO's lemmaselektion er at et ord skal optræde stabilt i korpus hen over en årrække. Vi antager derfor at en mere udbredt tidsmæssig repræsentation i korpus korrelerer med egnethed som lemma, hvorfor vi tæller antallet af årgange et ord optræder i – uden dog at tage højde for frekvensen pr. år. Scoren for stabilitet over tid er antallet af år en type forekommer i, delt med det totale antal år i korpus (18). Dermed er scoren højere, desto flere år en type optræder i.

Bøjningsformer

Vi antager at et ord med større sandsynlighed er etableret i dansk hvis det følger genkendelige mønstre for dansk morfologi. Som led i den automatiske lemmatisering i præprocesseringen (se afsnit 3.2.1) har vi optalt og gemt antallet af unikke former et ord optræder i. Vi har ikke taget højde for ordklasse i optællingen. Derfor kan ordklasser med få mulige bøjningsformer (fx adverbier) risikere at blive nedprioriteret i denne score. Vi kan dog også se at det er få typer som har mere end tre former i vores data. Vi vurderer derfor at uligheden har begrænset indflydelse på den en-

delige lemmascore. Scoren for bøjningsformer er antallet af unikke former for en type, men justeret så scoren ligger mellem 0 og 1.

Frasortering af proprier
I præprocesseringen har vi allerede fjernet en lang række person- og stednavne fra kandidatlisten. Vi mangler dog stadig at frasortere andre proprier som ikke forekommer på vores eksklusionslister. Det gælder bl.a. firma- og produktnavne og personnavne fra andre kulturer. Til dette formål anvender vi en sprogteknologisk metode ved navn Named Entity Recognition igennem modellen *ScandiNER*. Named Entity Recognition går ud på at få en model til at genkende proprier i en kort tekst, typisk én sætning ad gangen. For hver type har vi derfor tilfældigt udvalgt op til ti sætninger fra korpus og tagget dem ved hjælp af modellen. Scoren er den procentvise andel af sætninger hvor en type blev tagget som et proprium. Denne score vægtes negativt, og dermed er scoren lavere, jo oftere en type bliver anset for at være et proprium.

Dansk ortografi og nedgradering af fremmedord
For at give ord der følger dansk ortografi, større vægt end fx tokeniseringsfejl og fejlstavninger, har vi udviklet en score som måler "danskhed", altså overensstemmelse med typisk dansk ortografi. Til formålet har vi trænet en tetragrammodel på en fuldformsliste fra DDO, som beregner sandsynligheden for at en bogstavsekvens ligner kendte danske ord. Scoren er enten 0 eller 1, alt efter om sandsynligheden overstiger en given grænseværdi.

Der kan også være fremmedord på listen som er relevante lemmakandidater, især anglicismer og indlån fra engelsk og tysk. Vi har derfor tilføjet endnu en score som ser på om en type følger dansk, engelsk eller tysk ortografi. Vi har derfor også trænet en tetragrammodel på ordlister for engelsk (*Moby Crosswords word list*) og tysk (*Aspell-de*). Sprogscoren er højest hvis en type med stor sandsynlighed følger dansk ortografi, næsthøjest hvis typen

følger engelsk ortografi, og tredjehøjest hvis den følger tysk ortografi. Sprogscoren er 0 hvis typen er under grænseværdien for alle sprogene.

Som noget nyt i forhold til *Findor den ældre* har vi indført endnu en score der ser på typernes kontekst. I stedet for kun at se på ordene isoleret undersøger vi nu også om en type optræder i en dansk kontekst eller en kontekst fra andre sprog. På denne måde kan vi adskille faktiske indlån, som ofte vil optræde blandt danske ord, fra kodeskift eller citater på et andet sprog, som netop ikke vil have mange danske ord omkring sig. Denne score er særligt relevant for internetdata, da vi ikke kan være sikre på at der altid skrives på dansk i de webscrapede tekster. Til at udregne denne score bruger vi de samme tilfældigt indhentede sætninger fra propriumscoren, og vi bruger vores tetragrammodeller til at finde det mest sandsynlige sprog for hvert token inden for fem pladser fra den undersøgte type. Indlånsscoren er den gennemsnitlige sprogscore for alle tokens i alle sætningerne. Det resulterer i en højere score hvis ordet optræder i danske kontekster.

Semantisk lighed

Vi antager at et ord er en bedre lemmakandidat hvis det har synonymer eller nærsynonymer der allerede optræder i DDO. Vi har derfor benyttet en word2vec-model til at finde de 20 nærmeste naboer for hvert ord på vores lemmaliste. Det giver en højere score hvis ordet har flere synonymer i ordbogen blandt sine nærmeste naboer. I Sørensen et al. (2023) anvendtes en word2vec-model for dansk trænet på avistekster, men til denne undersøgelse har vi trænet en ny model specifikt på internettekster. Denne foranstaltning sikrer at alle ord i materialet faktisk er repræsenteret i vores semantiske model.

4. Resultater

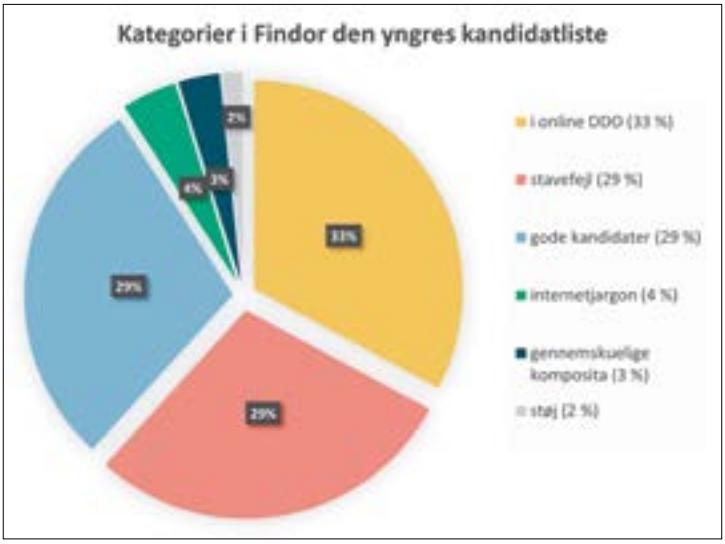
Vi har kørt *Findor* på de 174 millioner løbende ord i chatforumkorpusset, og det har resulteret i 135.331 unikke lemmakandidater rangeret efter lemmascore. Det er ikke tanken at alle kandidater skal med i DDO, men snarere at redaktørerne kan vælge fra et filtreret udsnit af kandidatlisten.

Vi evaluerer *Findors* kandidatliste på tre måder. Først analyserer vi indholdet i kandidatlistens øverste 2.000 kandidater. Herefter bruger vi DDO's opdateringer siden 2005 til at sammenligne *Findors* fundne lemmakandidater med lemmakandidater fundet ved ren frekvens. Til sidst foretager vi en kvalitativ evaluering, hvor en leksikograf manuelt vurderer tre forskellige udsnit af *Findors* kandidatliste.

4.1. Hvad finder *Findor*?

I figur 2 ses de øverste 2.000 typer på den kandidatliste som *Findor* har genereret, fordelt på seks forskellige kategorier. Kategorien "i online DDO" er automatisk annoteret og indeholder de lemmaer som er blevet tilføjet til DDO siden 2005. De resterende kategorier er annoteret manuelt. Med "gennemskuelige komposita" menes fx *laksetærte* og *broccolitærte* der ikke er decideret uegnede som lemmakandidater, men som det på grund af deres gennemskuelighed aktuelt er mindre oplagt at føje til ordbogen end mere svært afkodelige alternativer som *pletbløde*, *gummirøjser* og *pigefarve*. Sidstnævnte tilhører de "gode kandidater", hvilket vil sige at de er gode nok til at komme i betragtning til DDO.

Resultaterne er lovende da 62 % af top-2.000 enten allerede er i DDO eller egner sig til at komme det. Der er dog stadig en del stavefejl – mere herom i afsnit 5.3.



Figur 2: Kandidatlistens øverste 2.000 ord fordelt på kategorierne "i online DDO", "stavefejl", "gode kandidater", "internetjargon", "gennemskuelige komposita" og "støj".

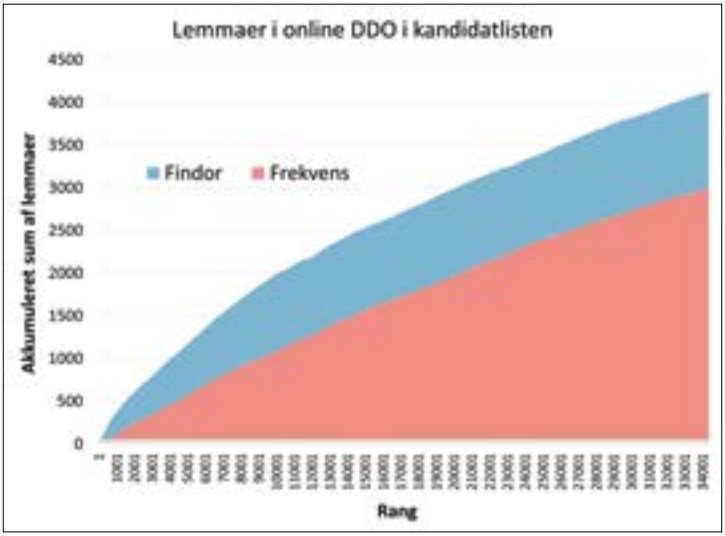
4.2. *Findor* versus frekvens

Findor den yngre er udviklet til at finde kandidater i nyindsamlet materiale som vi ikke tidligere har bearbejdet med vores gængse, frekvensbaserede værktøjer. Et oplagt spørgsmål er derfor om *Findor den yngre* klarer sig bedre end ren frekvens, når det gælder om at finde lemmakandidater i et helt nyt materiale, eller om vi kunne have opnået samme resultat med ren frekvens.

I denne undersøgelse bruger vi de lemmaer som er tilføjet til DDO siden 2005 som guldstandard, jf. afsnit 3.2. Vi antager at jo flere DDO-lemmaer en metode finder og placerer højt på listen, desto bedre er metoden til at finde nye lemmakandidater.

Vi sammenligner *Findors* rangering af lemmakandidaterne med en rangering fra højeste til laveste frekvens. I figur 3 ses den akkumulerede sum af lemmaer fra onlineudgaven af DDO for de

øverste 35.000 kandidater fra *Findor* (blå) og frekvens (rød). Her kan vi se at *Findor* konsekvent finder flere DDO-lemmaer, men er bedst i toppen af listen, hvor kurven er stejlest. Fx er 51 af *Findors* top-100 med i online DDO, svarende til 51 % af lemmaerne, mens det samme kun gør sig gældende for 12 lemmaer eller 12 % for frekvenslisten. Det samme mønster ser vi for top-1.000 hvor 37 % af *Findors* lemmaer er med i online DDO, mens kun 11 % er det for frekvenslisten.



Figur 3: Akkumuleret sum af antal lemmaer der allerede findes i onlineudgaven af DDO, fundet af henholdsvis *Findor* og via ren frekvens.

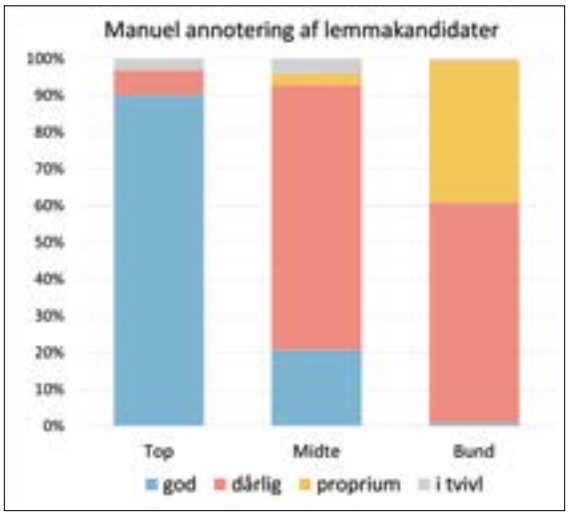
4.3. Manuel evaluering

En ulempe ved at bruge DDO's egne tilføjelser som guldstandard er at kandidaterne i listen sagtens kan være gode lemmakandidater uden at være føjet til ordbogen endnu. Dertil kommer at formålet med at udvikle *Findor* netop var at finde ord som vi ikke

tidligere har fundet – og som vi måske end ikke tidligere har haft adgang til at finde fordi vores sædvanlige korpus ikke indeholder de pågældende ord. For at kunne evaluere om metoden også fanger interessante kandidater uden for DDO's lemmaliste, iværksatte vi en evaluering med manuel annotation.

Vi udvalgte et tilfældigt udsnit på 222 kandidater fra henholdsvis de øverste, midterste og nederste 2.000 ord fra kandidatlisten, i alt 666 kandidater. Først frasorterede vi stavefejl, dernæst blandede vi de tre lister sammen, således at top-, midter- og bundkandidaterne optrådte sammen på en randomiseret liste.

Vi gav efterfølgende den samlede liste til en erfaren leksikograf med den opgave at placere hver kandidat i én af de fire kategorier "god", "dårlig", "proprium" og "i tvivl". Resultatet kan ses i figur 4. Her fremgår det at 90 % af ordene taget fra de øverste 2.000 er gode lemmakandidater, mens det samme kun gælder for 21 % fra midtergruppen og 1 % fra bunden.



Figur 4: Antal lemmaer vurderet af en leksikograf til at passe i kategorien "god", "dårlig", "proprium" eller "i tvivl" fra toppen, midten og bunden af listen over lemmakandidater.

Ud af de i alt 247 lemmaer som leksikografen har givet kategorien "god", findes 207 af dem ikke i forvejen på vores interne lister over lemmakandidater. *Findor* har altså overvejende fundet nye lemmakandidater.

Det er desuden værd at bemærke at proprierne helt overvejende befinder sig i den nederste del af listen, og at *Findor* altså har haft succes med at nedprioritere navne.

5. Diskussion

5.1. Repræsentativitet

Afsættet for denne undersøgelse er netop vores eksisterende korpus' repræsentativitet – eller mangel på samme. Vi er som nævnt i afsnit 2.1 blevet tiltagende bevidste om de mange domæner og sprogbrugere som ikke optræder i vores korpus, og forsøg på at udbygge det eksisterende korpus har endnu ikke båret frugt. Det er især juridiske udfordringer der spænder ben for denne proces, idet de mest attraktive teksttyper (fx moderne skønlitteratur) er belagt med copyright.

At sammensætte et repræsentativt korpus er af praktiske, økonomiske og ikke mindst juridiske årsager så godt som umuligt, men det er værd at huske på at en repræsentativ og fordomsfri beskrivelse af sproget er endnu mere umulig uden adgang til et korpus. Vi mener derfor ikke at manglen på et perfekt afbalanceret korpus skal spænde ben for at arbejde med et tekstkorpus, men vi må som leksikografer erkende korpussets fejl og mangler og ikke drage konklusioner ud over hvad vores korpus rent faktisk kan bære. Samtidig har vi et ansvar for gradvis at gøre vores korpus mere repræsentativt. Nærværende undersøgelse har netop til hensigt at råde bod på den aktuelle skævhed i forhold til genrer og sprogbrugere, idet vi har tilføjet tekster fra chatfora. Et internetkor-

pus er selvsagt ikke i sig selv mere repræsentativt for skriftsprog i Danmark end et aviskorpus, men ved at stykke forskellige korpusser sammen får vi et mere alsidigt og balanceret indtryk af moderne dansk, idet flere sprogbrugere og domæner er repræsenteret.

5.2. Substantiver og komposita

Et kritikpunkt ved *Findor den ældres* lemmakandidater var den relativt høje andel af såkaldt gennemskuelige komposita. Vi fjernede derfor værktøjets kompositumsplitter, som vi formodede kunne øge mængden af komposita, netop fordi den belønnede ord der var sammensat af eksisterende ord i DDO, med en højere score.

Findor den yngre finder fortsat først og fremmest komposita. DDO har imidlertid eksisteret i omkring 30 år, og vi må derfor også forvente at beskrivelsen af kerneordforrådet er et afsluttet kapitel. Vi er længere ude i periferien, og derfor er der nødvendigvis færre simpleksord – og i det omfang vi stadig tilføjer simpleksord til ordbogen, er de som oftest låneord. Vi vil på et senere tidspunkt undersøge om vi kan fremfinde de mest relevante sammensætninger for DDO, fx ved at kigge på brugsstatistikker som det er gjort på norsk i Paulsen (2023).

De lemmakandidater *Findor* har fundet, er i altovervejende grad substantiver. Det skyldes dels det faktum at netop denne ordklasse er meget produktiv (bl.a. i kraft af de mange komposita), dels at *Findors* parametre for søgning og filtrering især tilgodeser substantiver. Det kunne være interessant i fremtiden at forsøge målrettet at opspore eksempelvis adjektiver, men det har ikke været prioriteten i denne omgang.

5.3. Stavefejl

Et iøjnefaldende resultat er mængden af stavefejl på den endelige lemmakandidatliste. Vi forventede, jf. afsnit 3.1, flere stavefejl i et

chatforumkorpus sammenlignet med vores aviskorpus, fordi brugere af internetfora sjældent er professionelle sprogbrugere med adgang til korrekturlæsere og redaktører, men 29 % stavefejl oversteg alligevel vores forventning.

En mulig forklaring er vores brug af en word2vec-model. Modellen finder nemlig semantisk lignende ord, og stavefejl har selvsagt en betydning der er identisk med den korrekt stavede udgave. Denne forklaring underbygges af at stavefejlene i altoverskyggende grad befinder sig i toppen af listen og altså er blevet vægtet højt af *Findor*.

Imidlertid er mængden af stavefejl ikke udelukkende et problem for os: Dels føjer vi dem til vores liste over fejlstavninger, sådan at de udelades fra fremtidige lister med lemmakandidater, dels er stavefejl en ressource i sig selv, idet vi indlemmer dem i DDO's søgehjælp, sådan at brugere i fremtiden får nemmere ved at finde det ønskede opslagsord – uanset stavekundskaber.

5.4. Sorterer vi for meget fra?

At sortere store mængder støj fra er hele formålet med vores algoritme, men det er klart at gode lemmakandidater også kan ryge i svinget. Som nævnt i afsnit 3.2.1 frasorterer vi alle ord der indeholder tal, for fx at slippe for produktive dannelser som *1-1-sejr* – til gengæld ofrer vi muligheden for i denne omgang at finde gode lemmakandidater der indeholder tal. Det samme gør sig gældende for låneord, idet vi ved hjælp af en tetragrammodel nedprioriterer ord der ikke følger dansk, engelsk og tysk ortografi. Det betyder at vi måske går glip af gode kandidater fra andre sprog, eksempelvis *romanesco*. På samme måde frasorteres forkortelser som *uvb-stråle* 'ultraviolet stråle fra fx solen'. Det er dog ikke tanken at *Findors* lemmakandidatliste skal stå alene, og vi anser det derfor ikke som et problem at vi måske frasorterer for meget – for frasortering er en forudsætning for hele metoden.

6. Konklusion

Ved hjælp af *Findor den yngre* har vi fundet lemmakandidater fra hverdagssproget som vi ellers ikke kunne finde automatisk. *Findor* er rent frekvensbaserede værktøjer overlegen i kraft af sin komplekse filtrering der tager højde for bl.a. lemmakandidaternes semantiske lighed med eksisterende lemmaer og deres overensstemmelse med dansk ortografi og morfologi. Konkret udmærker 207 af de 247 ord som leksikografen godkendte fra *Findors* liste, sig ved at være nye i forhold til de kandidater vi har fundet via vores hidtidige, rent frekvensbaserede metoder.

Findor retter i et vist omfang op på aviskorpussets iboende begrænsninger – men metoden bør ad åre afprøves på og tilpasses et endnu bredere udvalg af teksttyper og -genrer.

Litteratur

Ordbøger, korpusser og digitale resurser

Aspell-de. <ftp.gnu.org/gnu/aspell/dict/oindex.html> (januar 2024).

Baby.dk. <baby.dk/debat/grupper.aspx> (december 2023).

Bold.dk. <bold.dk/snak> (november 2023).

CSTLEMMA. <github.com/kuhumcst/cstlemma> (oktober 2023).

Danmarks Statistik. <sprogteknologi.dk/dataset/fornavne-og-efternavne-i-befolkningen-i-danmark-i-januar-2020> (marts 2023).

DDO = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ddo> (marts 2024).

Debatten.net. <debatten.net/forum/> (december 2023).

Fakta om DDO. <ordnet.dk/ddo/fakta-om-ddo/ordbogens-tilblivelse> (marts 2024).

HardwareOnline. <hardwareonline.dk/forum_list.aspx?fid=23> (november 2023).
Hestenettet. <heste-nettet.dk/forum/1/> (august 2023).
Moby Crosswords word list. <gutenberg.org/files/3201/files/> (januar 2024).
Pokernet. <pokernet.dk/forum/kategorier/frontpage/off-topic.html> (december 2023).
Reddit.com/r/denmark. <reddit.com/r/denmark> (november 2023).
ScandiNER. <huggingface.co/saattrupdan/nbailab-base-ner-scandi> (januar 2024).
Tetragrammodel for dansk. <github.com/dsldk/lexiscore> (januar 2024).
Word2vec-model for dansk. <korpus.dsl.dk/resources/details/word2vec.html> (oktober 2023).
Wordclouds. <wordclouds.com> (marts 2024).

Anden litteratur

Falk, Ingrid, Delphine Bernhard & Christophe Gérard (2014): From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. I: *LREC-The 9th edition of the Language Resources and Evaluation Conference*. Reykjavik, Iceland. 4338-4344. <lrec-conf.org/proceedings/lrec2014/pdf/288_Paper.pdf>.
Halskov, Jakob & Pia Jarvad (2010): Manuel og maskinel excerpering af neologismer. I: *NyS – Nydanske Sprogstudier* 38, 39-68.
Kerremans, Daphné, Susanne Stegmayr & Hans-Jörg Schmid (2012): The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. I: Kathryn Allan & Justyna A. Robinson (eds.): *Current Methods in Historical Semantics* 73. Berlin, Boston: De Gruyter Mouton. 59-96.

Langemets, Margit, Jelena Kallas, Kaisa Norak & Indrek Hein (2020): New Estonian Words and Senses: Detection and Description. I: *Dictionaries: Journal of the Dictionary Society of North America* 41(1), 69-82.

Norling-Christensen, Ole & Jørg Asmussen (1998): The Corpus of the Danish Dictionary. I: *Lexikos* 8, 223-242. doi.org/10.5788/8-1-955.

Paulsen, Mikkel Ekeland (2023): Wheat or Chaff? A Compound Selection Model Based on Look-Up Data. I: *International Journal of Lexicography* 36(3), 306-324.

Sørensen, Nathalie Hau, Nicolai Hartvig Sørensen, Kirsten Lundholm Appel & Sanni Nimb (2023): Trawling the corpus for the overlooked lemmas. I: Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček & Simon Krek (eds.): *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. 392-409.

Kirsten Appel
Seniorredaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
ka@dsl.dk

Jonas Jensen
Seniorredaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
jj@dsl.dk

Nathalie Hau Sørensen
Assisterende redaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
nats@dsl.dk

Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text

Markus Forsberg & Louise Holmer

This article discusses data access, methodology development, and lexicographical work at Språkbanken Text at the University of Gothenburg.

Although large, the different corpora accessible through Språkbanken Text's research infrastructure relevant for the work on contemporary dictionaries have mainly built upon newspaper texts from limited geographical areas, as well as texts from other genres, but often limited to specific time periods. However, since 2021, several joint efforts and cooperations have been initiated to develop and refine this aspect of Språkbanken Text's material. In this article, we describe the composition of the new corpus material, the development of new tools, and point out some possible research areas that have now appeared. One focus point is how the contemporary dictionaries SAOL (*Svenska Akademiens ordlista*) and SO (*Svensk ordbok utgiven av Svenska Akademien*) may benefit from new corpus material and new methods.

1. Inledning

I den här artikeln fokuserar vi på materialtillgång och metodutveckling vid Språkbanken Text vid Göteborgs universitet, framför allt i förhållande till arbetet med ordböckerna *Svenska Akademiens ordlista* (SAOL) och *Svensk ordbok utgiven av Svenska Akademien* (SO). Syftet är att ge en översikt över de nytillkomna material som finns tillgängliga via Språkbanken Text, särskilt genom ordforskningsplattformen Korp (Borin et al. 2012) i kombination med nyutvecklade verktyg för ordforskning. I artikeln diskuteras också möjliga framtida studier och lexikografisk utveckling.

För att närmare kunna beskriva de nutida materialen ges först

en introduktion till Språkbanken Text liksom det lexikografiska arbetet med SAOL och SO. Artikeln är i övrigt disponerad så att materialtillgången beskrivs och exemplifieras i huvudsakligen kronologisk ordning, och därefter fokuseras på metodutvecklingen. Avslutningsvis sammanfattas innehållet och vi ger förslag på framtida utvecklingsmöjligheter.

2. Språkbanken Texts lexikografiska verksamhet

Språkbanken Text är en del av en nationell forskningsinfrastruktur för språkliga data med syftet att stödja främst språkteknologisk och språkvetenskaplig forskning, men även annan forskning där språkliga data har en framträdande roll (Språkbanken Text 2024a). Språkbanken Text är placerad vid Institutionen för svenska, flerspråkighet och språkteknologi vid Göteborgs universitet. Sedan 2021 ingår dessutom den verksamhet och de lexikografer som arbetar med samtidsordböckerna SAOL och SO i Språkbanken Text (Borin & Holmer 2024). Den lexikografiska verksamheten vid Göteborgs universitet bedrivs alltså inom den mer renodlat språkteknologiska ramen såväl som inom det ordboks- och produktorienterade området, med mänskliga användare i åtanke.

2.1. Språkbanken Texts forskning

Språkbanken Text bedriver språkteknologisk forskning som ofta bidrar till att utveckla dess forskningsinfrastruktur. De språkteknologiska forskningsprojekten med lexikografiskt fokus har varit särskilt fruktsamma, speciellt ansatsen som gick under benämningen Svenskt Frasnät++ (Dannélls et al. 2021, jfr Borin & Holmer 2024:46) eftersom den knöt samman många tidigare lexikografiska projekt vid Göteborgs universitet. Dessa hade i några fall löpt över flera decennier, och i och med Svenskt Frasnät++ väcktes

därmed även tidigare arbeten, som var nära att falla i glömska, till liv.

De språkteknologiska arbetena vid Språkbanken Text med lexikografiskt fokus har länge haft ett särskilt uttalat syfte: att med hjälp av datorns hjälp göra storskalig automatisk ordanalys av stora mängder text, för att därmed göra texterna mer tillgängliga för forskning. Det gynnar också tillämpningen och det vetenskapligt grundade arbetet inom lexikografin.

Blickar vi bakåt ännu mer ser vi att Språkbanken Text är sprungen ur just ordboksverksamhet. Att den lexikografiska verksamheten vid institutionen sedan 2021 ingår i Språkbanken Text är på så vis en välkommen återgång till en tidigare ordning.

Att utveckla ordböcker med hjälp av språkteknologiska verktyg och språkteknologiskt förädlade textmaterial vid Göteborgs universitet går tillbaka ända till 1960-talet (Malmgren & Sköldberg 2013:125). Det arbetet fick sedermera en organisatorisk form 1975, när regeringen inrättade Logoteket (senare Språkbanken, nu Språkbanken Text) under Sture Alléns ledning. Mycket har hänt sedan dess, men den starka kopplingen mellan lexikografi och språkteknologi finns fortsatt kvar.

2.2. SAOL- och SO-redaktionen

SAOL och SO har länge haft sin redaktionella hemvist vid Göteborgs universitet. SAOL:s bakomliggande material flyttades till Göteborg från Lund på 1980-talet medan SO har sin grund i arbetet med Lexikalisk databas, till vilken grunden lades under 1960-talet (jfr Malmgren & Sköldberg 2013, se också Borin & Holmer 2024).

Traditionellt har det lexikografiska arbetet rent praktiskt gått till så att det redaktionella arbetet med en upplaga av ett visst verk har pågått under några år, sedan har den tryckta ordboken givits ut, och därefter har insatserna inriktats på nästa ordbok eller nästa

upplaga av SAOL eller SO. Arbetet med de två ordböckerna har alltså ofta gått omlott. Det här har också varit ett naturligt arbetssätt när det primära målet för verksamheten har varit att ge ut ordböcker i form av tryckta böcker.

När nu redaktionen för SAOL och SO ingår i Språkbanken Text har verksamheten blivit betydligt mer dynamisk. Detta sammanfaller också med att marknaden för tryckta ordböcker har gått tillbaka kraftigt. Ett exempel på det är att SO (2021) enbart publicerades digitalt (jfr Sköldberg 2022) och att den delvis historiska *Svenska Akademiens ordbok* (SAOB) enbart kommer att uppdateras digitalt. Planen för SAOL är dock att publicera den kommande upplagan som tryckt bok, liksom i olika elektroniska varianter. Det grundar sig bl.a. i SAOL:s långa tradition med tryckta ordböcker, liksom – faktiskt – efterfrågan från allmänheten.

De som utgör den aktiva forskargruppen involverade i SAOL och SO under perioden 2024–2028 består idag av runt tio lexikografer och språkteknologer, alla med olika specialkompetenser inom sina respektive områden.

3. Materialtillgång för svensk lexikografi

När ordboken *Svensk ordbok* 1986 (SOB) utvecklades vid Göteborgs universitet var det den första svenska ordbok som hade utarbetats med korpusbaserade metoder (Malmgren & Sköldberg 2013, Borin & Holmer 2024). Sedan dess har det lexikografiska arbetet vid Göteborgs universitet vilat på just digitala material och metoder, vilket numera är förhållandevis vedertaget i fråga om vetenskapligt grundade ordböcker över allmänspråket (jfr Atkins & Rundell 2008).

Det korpusmaterial som ordböckerna SAOL och SO (liksom deras föregångare) bygger på, har förändrats och utvecklats under åren. I detta avsnitt ges en översikt över dessa materials mest framträdande drag.

3.1. Press 65 – den första tidningskorpusen

Det tidigaste materialet som låg till grund för det stora forskningsprojektet *Nusvensk frekvensordbok* (NFO, Allén et al. 1970–1980) utgjordes av det som kom att kallas Press 65, fortfarande tillgängligt i Korp. Press 65 består av en miljon löpord från fem större svenska dagstidningar: Göteborgs Handels- och Sjöfartstidning, Svenska Dagbladet, Stockholmstidningen, Dagens Nyheter och Sydsvenska Dagbladet – Snällposten. Det insamlade materialet kom från noggrant utvalda delar av tidningarna, kategoriserade som utrikeskorrespondenters rapporter, kulturartiklar jämte recensioner och allmänna reportage. Ett antal texter som t.ex. sportartiklar, anonyma insändare, annonser med mera uteslöts medvetet under materialinsamlingen. Ytterligare en faktor som påverkade var att materialet skulle vara tillgängligt på hålkort för maskinell bearbetning (Språkbanken Text 2024b).

Press 65-innehållet består alltså av tidningstext från fem större morgontidningar där materialet är noggrant utvalt av den dåvarande forskargruppen i enlighet med ett antal kriterier. Denna korpus låg till grund för omfattande frekvensbaserade undersökningar om svenskans skriftliga ordförråd. De uppgifter som presenteras i *Tiotusen i topp* (Allén 1972) med frekvensinformation om svenskt skriftspråk kommer också från Press 65. Alltjämt finns Press 65 liksom dess efterföljare sökbara i Korp, något som bland annat tillåter diakroniska studier över det svenska ordförrådets utveckling.

3.2. SAOL och SO: äldre och nyare material

Från materialet i Press 65 har korpusunderlaget successivt utökats. För det lexikografiska arbetet med ordböckerna SAOL och SO liksom dess föregångare, har det framför allt varit aktuellt att använda textmaterial i form av korpusar med nyhetstexter. Efter Press

65 utökades materialet med ytterligare liknande nyhetskorpusar (t.ex. Press 76 på ca 1,3 miljoner löpord och Press 98 på ca 10,7 miljoner löpord) och därefter flera årgångar av morgontidningen Göteborgs-Posten, den sista årgången från 2013 med 16,9 miljoner löpord. I takt med att dessa material har tillgängliggjorts för forskning har de också legat till grund för *Nationalencyklopedins ordbok* (NEO, 1995–1996), SAOL 12 (1998), SAOL 13 (2006), SO (2009) och SAOL 14 (2015).

Allt eftersom korpusmaterialen har inkorporerats i Språkbanken Texts samlingar har också de material som det praktiska lexikografiska arbetet vilar på kunnat utökas sett till mängden, och det har även kunnat breddas genremässigt. I korpussamlingarna ingår också romankorpusar som består av både originalsvensk text och översättningar från framför allt engelska men även andra språk. Romanmaterialet är dock inte lika omfattande som tidningstextmaterialet, och även om det är välstrukturerat och t.ex. tillåter studier av enskilda författares språk, har materialet varit begränsat till att omfatta romaner från 1970–1990-talet. De ovannämnda romankorpusarnas storlek uppgår sammanlagt till ca 18,5 miljoner löpord. Till dessa kommer också hela textmängden från Litteraturbanken som också finns tillgänglig via Språkbanken Text, men den har inte tidigare använts aktivt i arbetet med ordböckerna. Förutom de moderna romankorpusarna och material från Litteraturbanken har det dessutom tillkommit ytterligare en romankorpus genom SAOB-redaktionens försorg (se avsnitt 3.3).

Under sent 1990-tal började texter publicerade på internet bli vanliga som skriftspråksform i det svenska samhället och det avspeglas även i Språkbanken Texts samlingar. De innehåller t.ex. ett antal omfattande korpusar med bl.a. bloggtexter. De 69 korpusarna med text från sociala medier innehåller sammanlagt ca 11,8 miljarder löpord, där ca 9,06 miljarder kommer från diskussionsforum och resten från t.ex. bloggar och twittertexter.

SAOL är en mer normativ ordlista medan SO är en mer

deskriptiv ordbok. Detta faktum har också inverkat på vilka texter som har legat till grund för materialvalen för de två ordböckerna. De något ledigare utformade bloggtexterna har främst använts i samband med SO-arbetet medan SAOL företrädesvis har grundats på det ofta mer vårdade skriftspråket från nyhetstexter. Båda ordböckerna har dock framför allt haft skriftspråk från nyhetstexter från de större svenska rikstäckande dagstidningarna, och i viss mån romantexter, som sin främsta materialkälla. För en översikt över arbetet med SO 2021, se Sköldberg (2022), och se Holmer (2022) för en något fylligare bild av framför allt utvalda tidnings-, roman- och bloggkorpusar, liksom hur dessa korpusar har använts i lexikala studier.

3.3 Språkbanken Texts nutida material

Språkbanken Text gör språkteknologiska bearbetningar av de material som läggs till i dess samlingar. Dessa bearbetningar syftar till att maximera tillgängligheten och nyttan för materialen som forskningsdata, utan att kränka lagar såsom upphovsrätten (jfr Bouma et al. 2024). Det kan till exempel handla om att skapa s.k. meningsmängder, där meningarna i ett material kastas om slumpvis, för att texterna inte ska vara intakta, samtidigt som referenserna till originalet bibehålls. Denna praktik liknar det som redan görs på internet av sökmotorer, där man kan få se ett utdrag av texten i sököversikten, medan man behöver besöka sidan för att få läsa hela texten.

Däri ligger även begränsningen i de material som finns tillgängliga via Språkbanken Text: det finns en tydlig slagsida mot material som är publicerade på internet, eftersom det är dessa material som verksamheten har teknisk tillgång till.

Men till skillnad från Språkbanken Text, har Kungliga biblioteket (KB) full tillgång till nästan allt som ges ut i Sverige, via lagarna om pliktexemplar (Kungliga Biblioteket 2024). Samtidigt har

KB inte laglig rätt att dela sina data vidare till Språkbanken Text – eventuella bearbetningar som måste göras av juridiska skäl för att datamängder ska kunna spridas fritt, behöver göras inom KB:s servermiljö. År 2021 påbörjades ett samarbete mellan Språkbanken Text och KB-labb, en forskningsinfrastruktursenhet vid KB, där just detta görs, något som har resulterat i nya, fritt tillgängliga, orddatamängder via Språkbanken Text, exempelvis i Korp: *Kubord*. Datamängderna fokuserar på lexikal information och syftar till att stödja ordforskning.

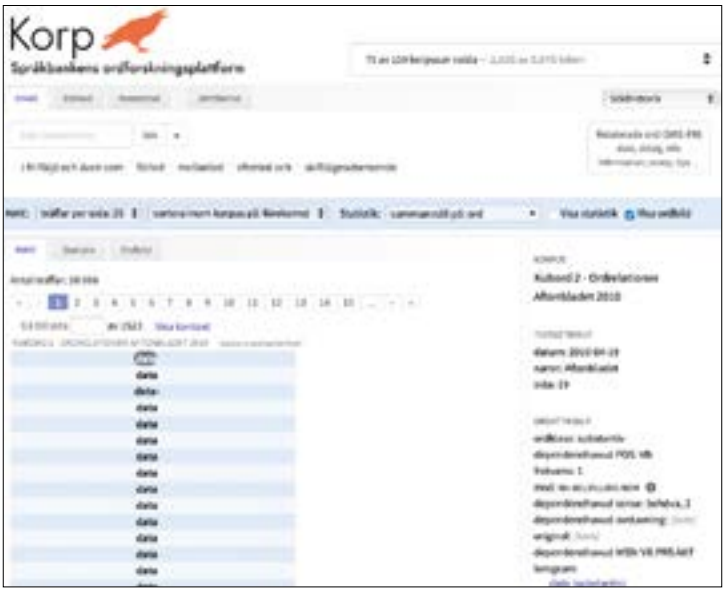
Kubord består av tidningstext från i huvudsak år 2010 till år 2021 och har hämtats från de större morgontidningarna Dagens Nyheter, Svenska Dagbladet och Göteborgs-Posten. Dessutom ingår Sydsvenska Dagbladet, som framför allt täcker Skåne, och Östgöta-Correspondenten, som ges ut i Östergötland (östra Mellansverige). Förutom dessa mer klassiska morgontidningar ingår också de två kvällstidningarna Aftonbladet (oberoende socialdemokratisk) och Expressen (obunden liberal). Till skillnad från de äldre materialen (Press 65 och liknande utvalda textmängder) innehåller Kubord-mängderna i princip de flesta delarna från de digitaliserade papperstidningarna, alltså både inrikes- och utrikesnyheter, sport, nöje m.m. De speglar därmed en större bredd av presstexterna än de tidigaste korpusarna.

I och med Kubord är det första gången som modernt kvällstidningsmaterial över huvud taget finns tillgängligt via Språkbanken Text. På sikt är det möjligt att det kommer att kunna gå att utöka antalet tidningar med olika typer av regional hemvist, vilket skulle balansera materialet ytterligare geografiskt. Det vanliga är annars att de största tidningarna anses tillräckliga, men de behandlar främst rikstäckande nyheter och har ofta storstäderna Stockholm, Göteborg och Malmö i fokus. Nyheter från t.ex. landsdelen Norrland kommer inte lika ofta med i tidningarna, och som en följd av det finns sådana nyhetstexter inte heller representerade i referenskorpusarna.

Delmängden Kubord 1 i Korp består av en fritt tillgänglig ordlistning med källreferenser som blivit språkteknologiskt förädlad för att innehålla sådant som ordklass, lemmatisering, ordbetydelser med mera. Kubord 1 är både nedladdningsbar via Språkbanken Texts datasida och tillgänglig via Korp.

Kubord 2 är en vidareutveckling av Kubord 1. Kubord 2 består av en listning med par av ord som är relaterade via en syntaktisk relation, exempelvis ett verbs direkta objekt eller substantivs framförställda attribut. Detta möjliggör att man exempelvis kan skapa en s.k. ordbild i Korp, vilken ger uppgifter om ordets syntaktiska relationer och liknande (jfr Borin et al. 2012:476). Sammanlagt bidrar Kubord med ca 6 miljarder ord.

För att illustrera visas i Figur 1 och 2 exempel från Kubord 2 i Korp via en sökning efter lemmat *data*.



Figur 1. Resultat vid sökning efter *data* i Kubord 2 när fliken KWIC är aktiverad (keyword in context)

I figur 1 visas sökresultatet för *data* med 38 066 träffar. Resultatet består alltså av sökträffar utan kontext, men med rik extra information (en delmängd visas till höger). Även om det kan se något fattigt ut fungerar det mesta av Korps funktionalitet för Kubord 2, förutom att användaren inte får någon kontext presenterad. I de fall användaren behöver gå till den ursprungliga texten finns källhänvisningar. För första träffen noteras Aftonbladet, 2010-04-19, sidan 19 (visas till höger i figur 1). Vi kan även observera t.ex. vilket syntaktiskt huvud som första träffen har, nämligen *behöva*.

Figur 2. Resultat vid sökning på *data* i Kubord 2 när fliken Ordbild är aktiverad

I figur 2 visas ordbilden för *data* i Kubord 2. Typiska attribut för *data* i Kubord-materialet är *extra*, *biometrisk* och *personlig*. Ordet *data* är också något som typiskt *köps*, *ingår*, *hämtas* och *samlas*.

Utöver Kubord-arbetet, kan tilläggas att redaktionen för SAOB i Lund nyligen utfört ett omfattande arbete med att skanna in romaner från perioden 1950 till 2007. Korpusen har fått namnet SAOB 1950 och finns i Språkbanken Texts samlingar, genom Korp och som öppet tillgänglig nedladdningsbar meningsmängd (Nationella språkbanken 2023). I och med den insatsen har mängden romanmaterial, ett tidigare ofta förbisett material, ökat avsevärt.

Dessutom täcker korpusen in även åren 1950 till 1980, decennier som i övrigt inte har varit särskilt välrepresenterade tidigare. Korpusen SAOB 1950 omfattar ca 50 miljoner ord, vilket innebär ett rejält tillskott till det befintliga romanmaterialet (se avsnitt 3.2).

3.4. Sammanfattning av materialutvecklingen

Både SAOL:s och SO:s nuvarande datamängder utgör resultatet av en mångårig tradition i kombination med nya samarbeten, nya material och nya metoder. Ett uppenbart problem ur materialsynvinkel är att tillgången till webbmaterial för forskningsändamål, som respekterar upphovsrätten och andra ekonomiska värden, minskar i takt med att betalväggar och andra tekniska lösningar begränsar tillgängligheten och därmed tillgången till data. Därför är det avgörande för svensk ordforskning att en verksamhet som Språkbanken Text ger sig in i strategiska partnerskap med de organisationer som har datatillgång via lagen eller på andra sätt har tillgång till relevanta data.

Från de tidigaste korpusarna med text från 1960- och 1970-talet omfattande någon miljon ord vardera, har tillgängliga material i Språkbanken alltså vuxit och omfattar nu många miljarder löpord fördelade på runt 300 korpusar med moderna material. Utöver det finns det numera också korpusar över fornsvenska, historiskt material från KB m.m. och diverse specialkorpusar som kan vara insamlade av enskilda forskare.

Förhoppningen är att på sikt kunna lägga till fler årgångar av moderna dags- och kvällstidningar i Korp, liksom tidningsmaterial från fler regioner så att materialet breddas geografiskt. Samtidigt får vi också konstatera att med moderna tidningskorpusar på 6 miljarder ord, som uppdateras löpande, i kombination med det tidigare och mycket omfattande textmaterialet, måste tillgången till svensk text i forskningssyfte sägas vara mycket god.

4. Nyutvecklade metoder

I samband med att de nya materialen har kunnat tillgängliggöras via Språkbanken Texts forskningsinfrastruktur (se avsnitt 3), har forskargruppen som arbetar med ordböckerna också kunnat utveckla särskilda verktyg för nyordsexcerpering och göra storskaliga jämförelser mellan befintliga ordboksmaterial och de nya korpusmaterialen. I följande avsnitt beskrivs först det nyordsverktyg som håller på att utvecklas internt inom gruppen och därefter arbetet med så kallade ordvektorer. Naturligtvis används också andra metoder och verktyg i kombination – förutom det här beskrivna nyordsverktyget används till exempel manuell excerpering av dagsaktuella texter, förslag på nya ord från allmänheten, loggfiler med icke-träffar från de elektroniska versionerna, jämförelser med motsvarande nordiska ordböckers listor över nytillkomna ord m.m. Här fokuseras dock på de nya metoder som har utvecklats sedan 2021.

4.1. Nyordsverktyg

Ett självklart led i att utveckla samtidsordböckerna är att uppdatera lemmalistan med nya ord, liksom att utmönstra föråldrade ord (jfr Diamond 2016). I grund och botten är maskinellt stöd för en sådan process enkel och ytterst beroende av relevant materialtillgång. Ordboksredaktionen vill undersöka vilka ord som ökar signifikant i frekvens i materialet över åren och vilka som minskar. Ord med ökad frekvens kan utgöra nyordskandidater medan ord med minskad frekvens kan utgöra strykningskandidater (jfr Berg, Holmer & Sköldberg 2010, Holmer et al. 2024).

Samtidigt blir det snabbt viktigt med en mer kritisk hållning till ordens frekvensinformation och deras ursprung. Om en signifikant frekvensökning kommer explosionsartat över ett par dagar

i bara en tidning, är det förmodligen en mycket sämre nyordskandidat än den vars frekvensökning är jämnare fördelad över tid och över flera material. Den senare fångar bättre vad forskarna och redaktionen är ute efter, nämligen att ordet har ökat i bruk i hela samhället.

Man kan också fråga sig vilken tidsenhet som är mest lämplig för frekvensjämförelser. Som regel väljs ofta år, inte minst för att korpusmaterial brukar delas upp årsvis, men det finns andra varianter, t.ex. ett glidande medelvärde. Därtill kommer hur det statistiskt kan avgöras att det alls skett en ökning och minskning. Alla varianter kommer med sina fördelar och nackdelar, så ett användbart nyordsverktyg behöver kunna hantera den variationen.

Sedan har vi frågan om analysenhet: vad menar vi egentligen med ett ord? Menar vi bara en ordform eller bör vi gruppera ordformer under ett visst lemma? Hur är det med de sammansättningar ordet ingår i, ska de också tas med i beräkningarna? Ska exempelvis *selfiemuseum* och *gymselfie* bidra till bedömningen av *selfie*?

Till sist har vi frågan hur en nyordkandidat ska presenteras. Ett idealiskt nyordsverktyg skulle gå hela vägen och också skapa ett ordboksartikelutkast med färdig information som böjningsinformation, typiska sammansättningar, språkexempel m.m.

Inom ordboksverksamheten jobbar vi med att utveckla ett nytt integrerat nyordsverktyg som tar sig an dessa frågeställningar och målsättningar. Hittills har detta verktyg framför allt kunnat användas för att vaska fram ett antal nyordskandidater. Följande handfull exempel, excerperade med hjälp av detta verktyg, utgör t.ex. potentiella nyord i SAOL och självständiga uppslagsord eller sammansättningsexempel i SO: *hållbarhetstänk*, *höjdrädsla*, *kärleksfront* (i frasen *på kärleksfronten*) *näringsjäst*, *samhällsutmaning*, *spontanköp*, *webbenkät* och *viltkamera*.

4.2. Ordvektorer

Under 2024 håller Språkbanken Text och KB-labb tillsammans på att utveckla en ny sorts Kubord-datamängd benämnd *Kubord-fasttext*, med ordvektorer. Enkelt uttryckt så kan man med ordvektorers hjälp få svar på frågan om vilka ord som har liknande språkliga kontexter i ett visst material.

Ordvektorer definieras algoritmiskt utifrån sin språkliga kontext på så vis att de ord som har liknande språkliga kontexter hamnar nära varandra matematiskt i en så kallad vektorrymd. För ett ord som *sjunga* kan det betyda att verb som *gnola, skråla* och *pratsjunga* ligger nära i vektorrymden, tillsammans med ord från andra ordklasser, som exempelvis *sång* och *skolkör*. Det senare är en indikation på en egenskap hos dessa ordvektorer, nämligen att syntaktisk position inte är en del av en ordvektors uppbyggnad. Så länge två ord förekommer i samma mening med liknande andra ord omkring sig hamnar de nära varandra.

I samarbetet använder vi oss av verktyget fastText (Bojanowski et al. 2017; verktygets namn skrivs med versalt T) för att skapa våra ordvektorer. Det som skiljer detta verktyg från andra verktyg som bygger ordvektorer, exempelvis Word2Vec, är att fastText konstruerar en ordvektor utifrån ordets delar. Det betyder i praktiken att ord med sammanfallande delar hamnar närmare varandra i vektorrymden, dvs. att ord som *sjunga* och *provsjunga* kommer närmare varandra på grund av att de delar *sjunga*.

Vad kan då dessa ordvektorer användas till i det lexikografiska arbetet? De kan användas för att identifiera lemmaluckor, hitta nya ordrelationer mellan lemman eller för att lokalisera typiska sammansättningar för ett ord, för att nämna några exempel. En ordvektor för *sommar* har exempelvis grannar som höst, vår och *vinter,* typiska sammansättningar som *sommarsäsong* och *försommar*, och även mer indirekta ordrelationer, som *båtsäsong*.

För mer detaljerade redogörelser om arbetet med ordvektorer

i en lexikografisk kontext, se Forsberg & Sköldberg (under utgivning).

5. Avslutande ord

I artikeln presenteras det arbete med nya material och nya metoder som har gjorts vid Språkbanken Text framför allt från och med år 2021. Fokus ligger på det lexikografiska arbete som ligger till grund för ordböckerna SAOL och SO. Vid Språkbanken Text pågår (eller har pågått) ytterligare projekt som är relaterade till arbetet med ordböcker och vars projektmedlemmar ibland är involverade i flera projekt samtidigt (se t.ex. Språkbanken Text 2024c).

Svenskan är trots allt ett välbeskrivet språk med flera ordböcker, grammatikor och korpusar. Som kontrast till artikelns fokus på svenskt skriftspråk kan nämnas arbetet med isländska talspråkskorpusar (Hilmisdóttir 2024) och arbetet med ett språk som meänkieli (Ahltorp et al. 2024). Att ständigt vilja utöka de svenska korpusmaterialen med ännu mer skriftspråk kan kanske tyckas onödigt. Vi ser dock möjligheterna i att, som nämnts, kunna öka den geografiska spridningen, något som möjliggör studier av t.ex. regional variation. Det finns också möjlighet till mer avancerade studier av ordförrådet över tid. Till de utökade materialen kommer en satsning på att metoder för identifiering och beskrivning av nyord.

Att ha tillgång till så dagsaktuella texter som möjligt är också helt nödvändigt för utvecklingen och uppdateringen av de två samtidsordböckerna SAOL och SO. Samarbetena med KB-labb har gjort att det lexikografiska arbetet kan baseras på ett mer kontinuerligt datainflöde av samtidstext än vad som har varit fallet tidigare, och i och med romankorpusen SAOB 1950 tillgängliggörs ett urval av 1900-talets skönlitterära texter i Korp.

Med tanke på att AI-genererade texter har blivit vanligare se-

dan åtminstone 2023, har det också blivit en viktig fråga att kunna skilja autentisk text från maskingenererad sådan, för att kunna göra en korrekt beskrivning av det svenska ordförrådet. Till viss del kan det avhjälpas genom tillgång till texter som man vet är (i stort sett) mänskligt producerade, exempelvis tidningstexterna tillgängliga via Språkbanken Text (t.ex Kubord 1 och Kubord 2). Men även i tidningstexter har användning av maskingenerad text börjat öka. En framtida utmaning blir att skilja dessa två texttyper åt, antingen via kunskap om materialet, till exempel genom att märka upp de delar som man vet att en tidning maskingenererar, eller genom att utveckla ny automatisk analys som försöker göra detsamma.

Avslutningsvis vill vi återigen lyfta hur central redaktionens inkorporering i Språkbanken Text har varit för båda verksamheterna, där det finns en växelvis draghjälp.

Litteratur

Ordböcker, korpusar och digitala resurser

Allén, Sture (1972): *Tiotusen i topp. Ordfrekvenser i tidningstext*. Stockholm: Almqvist & Wiksell.

KB-labb. <kb.se/samverkan-och-utveckling/kb-labb.html> (april 2024).

Korp = Språkbankens ordforskningsplattform. Version 9.0.6. <spraakbanken.gu.se/korp/> (augusti 2024).

Kungliga Biblioteket (2024). <kb.se/om-oss/det-har-gor-vi.html> (juli 2024).

Nationella språkbanken (2023). <sprakbanken.se/aktuellt/nyheter/2023-12-07-sprakteknologi-forenklar-arbetet-med-nya-saob> (juli 2024).

NEO = *Nationalencyklopedins ordbok*, band 1–3 (1995–1996). Höganäs: Bokförlaget Bra Böcker.

NFO = *Nusvensk frekvensordbok baserad på tidningstext* (1970–1980). Utarbetad av Sture Allén m.fl. Fyra band. Stockholm: Almqvist och Wiksell international (distr.).

SAOB = *Svenska Akademiens ordbok* (1898–2023). <www.saob.se/> (augusti 2024).

SAOL = *Svenska Akademiens ordlista över svenska språket*, 14 uppl. (2015). Stockholm: Norstedts.

SO = *Svensk ordbok utgiven av Svenska Akademien* (2021). <svenska.se/so/> (april 2024).

SOB 1986 = *Svensk ordbok* (1986). Göteborg: Språkdata & Esselte Studium AB.

Språkbanken Text 2024a. <spraakbanken.gu.se/> (mars 2024).

Språkbanken Text 2024b. <spraakbanken.gu.se/resurser/press65> (april 2024).

Språkbanken Text 2024c. <spraakbanken.gu.se/forskning> (augusti 2024).

Svenska.se = Svenska Akademiens ordboksportal. <svenska.se/> (mars 2024).

Annan litteratur

Ahltorp, Magnus, Lina Lejdebro Enwald, Elina Kangas, Jacob Larsson, Rickard Domeij & Gunnar Eriksson (2024): Språkteknologi för att samla in texter och analysera språket i korpusverktyg – hur gör man på meänkieli? I: *LexicoNordica* 31 (denna volym).

Atkins B.T. Sue & Michael Rundell (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Berg, Sture, Louise Holmer & Emma Sköldberg (2010): Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL). I: *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010)*. Leeuwarden: Fryske Akademy. 567–576.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov (2017): Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* (2017) 5, 135–146.

Borin, Lars, Markus Forsberg & Johan Roxendal (2012): Korp – the corpus infrastructure of Språkbanken. I: *Proceedings of LREC 2012. Istanbul: ELRA*. 474–478.

Borin, Lars & Louise Holmer (2024): Tradita innovare, innovata tradere. The Gothenburg way of computational lexicography. I: *Proceedings of the Huminfra Conference* (HiC 2024). 41–50.

Bouma, Gerlof, Markus Forsberg, Justyna Sikora & Emma Sköldberg (2024): Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten. I: *Proceedings of the Huminfra Conference* (HiC 2024). 161–167.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (2021): *The Swedish FrameNet++ Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins.

Diamond, Graeme (2016): Making Decisions about Inclusion and Exclusion. I: Philip Durkin (ed.): *The Oxford Handbook of Lexicography*. Oxford. 532–545.

Forsberg, Markus & Emma Sköldberg (Under utgivning). Ord med liknande kontext sökes! Om ordvektorers roll i svensk lexikografi.

Hilmisdóttir, Helga (2024). Talspråkskorpusar som resurs för isländska ordböcker. I: *LexicoNordica* 31 (denna volym).

Holmer, Louise (2022): Neutrala substantiv på *-ande* i text och ordbok. Meijerbergs arkiv 47. Göteborg.

Holmer, Louise, Ann Lillieström, Emma Sköldberg & Jonatan Uppström (2024): SAOL och svensk språkvetenskaplig infrastruktur – nu och i framtiden. I: *Proceedings of the Huminfra Conference* (HiC 2024). 68–75.

Malmgren, Sven-Göran & Emma Sköldberg (2013): The Lexicography of Swedish and other Scandinavian Languages. I: *International Journal of Lexicography*, 26(2), 117–134.

Sköldberg, Emma (2022): Andra upplagan av Svensk ordbok: förutsättningar och redaktionella val. I: *LexicoNordica* 29, 139–152.

Markus Forsberg
Föreståndare för Språkbanken Text
Inst. för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
Box 200
SE-40530 Göteborg

Louise Holmer
Forskare, lektor
Inst. för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
Box 200
SE-40530 Göteborg

Skevheter och utmaningar i ordboksbasen för fyra enspråkiga finländska ordböcker

Tarja Riitta Heinonen & Caroline Sandström

This article gives an overview of what kind of data the four monolingual dictionaries compiled at the Institute for the Languages of Finland are based on and what kind of data-related challenges and biases we have encountered. Major issues that generally affect all dictionaries include the reliability and representativity of the source material. A further concern for us is how to react to the passage of time, since all these dictionary projects trace back over a hundred years. We present a few examples on how we update the dictionary of contemporary language, taking into account that it is partially based on older dictionaries. We also show that current language usage impacts choices made in historical and dialect dictionaries.

1. Inledning

På Institutet för de inhemska språken i Finland redigeras fyra stora enspråkiga ordböcker: en finsk dialektordbok *Suomen murteiden sanakirja* (SMS), en svensk dialektordbok *Ordbok över Finlands svenska folkmål* (FO), en ordbok över äldre finskt skriftspråk *Vanhan kirjasuomen sanakirja* (VKS) och en nufinsk deskriptiv och preskriptiv ordbok *Kielitoimiston sanakirja* (KS). De tre dokumenterande, historiska ordböckerna SMS, FO och VKS bygger på omfattande arkivmaterial. De är alla typiska flergenerationsordböcker (Vikør 1999). Redigeringen inleddes på 1950–1960-talen och pågår fortfarande. Med tiden har redaktörer avlöst varandra och redaktionella principer och praxis har justerats. Teknisk utveckling och digitalisering har gett nya redigeringsredskap, men också medfört vissa utmaningar, i synnerhet vid övergången till digital publicering under 2010-talet. Den nufinska ordboken KS grun-

dar sig i sin tur delvis på tidigare materialbaserade samtidsordböcker.

Samtliga har karaktären av nationella storprojekt och deras tillkomsthistoria sträcker sig långt tillbaka i tiden. De finska ordböckerna är ett resultat av det ordboksprogram för tre stora vetenskapliga finska ordböcker som språkforskaren och politikern E.N. Setälä introducerade 1896. Programmet skulle omfatta en allmänfinsk ordbok, en ordbok över äldre finskt skriftspråk och en ordbok över folkmålen (Setälä 1896; Vilppula 1999:407f.; Onikki-Rantajääskö 2011:544–547; Ruppel & Sandström 2014:148). När det gäller de svenska dialekterna har lexikografin likaså av hävd varit en del av det nationella projektet för svenskan i Finland (Ahlbäck 1946:1; Sandström 2013:92f. och 2016:80f.).

En allmängiltig målsättning för vetenskapliga ordböcker är att ordboksbasen för beskrivningen i en ordbok bör bestå av ett så representativt, heltäckande och pålitligt materialurval som möjligt. Inledningsvis kan vi också konstatera att redigeringen av nuspråkliga ordböcker i högre grad påverkas av språkliga och samhälleliga förändringar än ordböcker som dokumenterar äldre språkskeden.

I denna artikel gör vi jämförelser och fokuserar på utmaningar i ordboksbaserna för de fyra ordböckerna. I avsnitt 2 ges först en kort bakgrund till ordboksbas som begrepp. I avsnitt 3 presenteras ordböckerna, deras omfång och typen av material som de bygger på. I avsnitt 4 diskuterar vi utmaningar som framträder vid redigeringen av dessa ordböcker. Vi tar fram skillnader och likheter i de problem som redaktörerna möter, men fokuserar också på hur redaktionerna har löst och hanterat skevheter och utmaningar i materialen, t.ex. när det gäller representativitet och pålitlighet i förhållande till språkliga och samhälleliga förändringar. Huvudpunkterna i vår artikel sammanfattas i avsnitt 5.

2. Ordboksbas och källor för ordböcker

I *Nordisk leksikografisk ordbok* (NLO 1997:199) definieras *ordboksbas* som de källor som valts ut för en ordbok och enligt Svensén (2004:50) som ”[d]e källor som används för utarbetandet av en ordbok”. Både Svensén och NLO gör en indelning i primärkällor och sekundärkällor. Primärkällorna består av autentiskt språkligt material, antingen muntligt eller skriftligt. De kan vara av olika slag: beläggsamlingar, korpusar eller material som man kommer åt genom introspektion (lexikografens egen kompetens eller informantdata). Sekundärkällorna utgörs av existerande beskrivningar av språket, såsom andra ordböcker, grammatikor eller specialundersökningar.

Svensén (2004:51) framhåller att då data tas fram genom introspektion bör denna metod användas med försiktighet. Vid introspektion kan det egna språkbruket och bedömningen av språkliga data påverkas av mer eller mindre medvetna föreställningar om språkliga normer. Om en lexikograf producerar egna språkprov är det därför viktigt att kontrollera språkprovet mot autentiska språkprov från andra källor. I diskussionen om datainsamling och dataurval skärskådar Svensén (2004:52–57) olika metoder att samla in material till en ordbok från primärkällor och konstaterar att beläggsamlingar tidigare var det enda säkra sättet, men numera är korpusmetoden den vanligaste. Svensén betonar att det finns för- och nackdelar med bägge metoderna, och att korpusmetoden och beläggmetoden kan komplettera varandra. En korpus ger information om det typiska språkbruket och den ger även möjlighet till kvantifiering. I en beläggsamling, som tillkommit genom urval, betonas ofta lågfrekventa och ovanliga språkliga företeelser.

När det gäller de fyra finländska ordböckerna finns det skillnader mellan ordboksbaserna för en dialektordbok eller en historisk ordbok som VKS, och en modern allmänspråklig ordbok. De två dialektordböckerna bygger på traditionella beläggsamlingar som

sammanställts för ordböckerna. Ordboksbasen för den historiska ordboken VKS utgör en kombination av en äldre beläggsamling och en korpus. I flera omgångar har redaktionen för VKS breddat sin ordboksbas och utökat materialet genom en textkorpus där texterna gjorts sökbara. Den allmänfinska ordboken KS bygger dels på två äldre samtidsordböcker, dels utnyttjas såväl introspektion som kontroller i tillgängliga korpusar och varierande primär- och sekundärkällor.

3. Översikt över de fyra ordböckerna SMS, FO, VKS och KS: bakgrund, material, publicering

Det finns många likheter i bakgrund och villkor för de enspråkiga finländska ordböckerna på Institutet för de inhemska språken. Gemensamt för dem alla är att materialet huvudsakligen var färdigt insamlat och att redigeringen hade påbörjats, innan föregångaren Forskningscentralen för de inhemska språken (1976–), senare Institutet för de inhemska språken (2012–), existerade. Likheter framträder speciellt i fråga om de två finländska dialektordböckerna, som är monumentala nationella projekt med uppgift att dokumentera, bevara och publicera det språkliga kulturarvet för nationalspråken finska och svenska.

3.1. Den finska dialektordboken SMS

Den finska dialektordboken SMS (*Suomen murteiden sanakirja*) grundar sig på en enorm beläggsamling som sammanställts i ordbokens arkiv (*Suomen murteiden sana-arkisto*). Arkivet består av originalsamlingar och omfattar ca 8 miljoner ordsedlar med uppgifter om ca 400 000 ord. Samlingarna utgörs dels av svar på frågelistor, dels av uppteckningar gjorda av både vetenskapligt utbildade stipendiater och dialektkunniga lekmän. Materialet

samlades in 1890–1973 i Finland, men också i finsktalande områden utanför Finlands gränser, bl.a. i Värmland, Tornedalen och Finnmarken i Norge (Vilppula 1999:407). Det sammanställdes för redigering 1924–1974. Redigeringen av SMS inleddes på 1960-talet och ordboken publicerades i tryck 1985–2008 i åtta band (av planerade 20 band). Nätpublicering av SMS påbörjades 2012, och 2024 har avsnittet *a–noukkia* utgivits. Omkring 187 500 av planerade 350 000 artiklar och dessutom över 74 000 utbredningskartor har blivit publicerade.

3.2. Den svenska dialektordboken FO

Ordbok över Finlands svenska folkmål (FO) är en dialektordbok som behandlar dialektmaterial insamlat under perioden 1860–1970 i regionerna Österbotten, Satakunta, Åland, Åboland och Nyland inom nuvarande Finlands gränser. Ordboken redigeras utgående från en beläggsamling som omfattar omkring 1 miljon ordsedlar, som excerperats ur dialektologiska och etnologiska uppteckningar, bandinspelningar, dialektologiska avhandlingar och också ur texter av dialektkunniga lekmän. I materialet ingår två äldre dialektordböcker utgivna av Herman Vendell (1904–1907) och V.E.V. Wessman (1925–1932). Beläggsamlingen FOreg. (Register för FO) sammanställdes 1945–1980. Redigeringen inleddes på 1960-talet och ordboken utgavs 1976–2007 i fyra band (av planerade sju band). Nätpublicering av FO påbörjades 2013, och 2024 har avsnittet *a–rättvisa*, ca 79 500 av planerade 120 000 artiklar publicerats.

3.3. Ordbok över äldre finskt skriftspråk VKS

Ordboken över äldre finskt skriftspråk VKS (*Vanhan kirjasuomen sanakirja*) är en historisk ordbok som behandlar över 1 500 skriftliga finska källor utgivna 1543–1810. Ordboksbasen består av en

beläggsamling på ca 0,5 miljoner ordsedlar och en korpus på ca 3,5 miljoner ord. Beläggsamlingen sammanställdes under perioden 1930–1980 och korpusen under 1990-talet och början av 2000-talet. Redigeringen inleddes under 1950-talet. VKS utkom i tryck 1985–1994 i två band (*a–k*). Från 2014 publiceras den som nätordbok. Våren 2024 hade avsnittet *a–puhdistaminen*, 46 400 av planerade 80 000 artiklar, blivit publicerade på nätet.

3.4. Den nufinska ordboken KS

Ordboken KS (*Kielitoimiston sanakirja* 'Språkbyråns ordbok') är en deskriptiv och preskriptiv allmänspråklig ordbok. Den är utgiven i sin helhet *a–ö*, och omfattar drygt 100 000 lemman. KS uppdateras kontinuerligt och den har tidigare utkommit i olika utgåvor. KS grundar sig via mellansteget *Suomen kielen perussanakirja* 'Finsk basordbok' (PS) på den äldre, betydligt mer omfattande, *Nykysuomen sanakirja* 'Nufinsk ordbok' (NS) med drygt 200 000 lemman.

KS är alltså en direkt fortsättning av PS, som publicerades i tryck under åren 1990–1994, och i slutet av 1990-talet som CD-rom. Ordboken utgavs första gången under det nuvarande namnet *Kielitoimiston sanakirja* som betald nätordbok och CD-rom 2004. Den utkom också i tryck i två upplagor 2006 och 2012. Från 2014 publiceras KS som en öppet tillgänglig nätordbok.

Arbetet med NS hade påbörjats 1929 och ordboken utkom i sex delar under perioden 1951–1961. NS grundar sig på ett omfattande arkivmaterial (ca 4,5 miljoner ordsedlar med uppgifter om ca 850 000 ord). Materialet, som sammanställdes på 1930–1950-talen, excerperades ur skönlitteratur och facklitteratur utgiven under perioden 1880–1950, men också ur äldre språkligt och kulturellt epokgörande verk som Kalevala och Aleksis Kivis produktion (Haarala 1999:151–155; Ruppel & Sandström 2014:151–153).

Då arbetet med PS inleddes utnyttjades den äldre ordboken NS

som bas. Syftet med PS var att den skulle vara en aktuell, kompaktare version av NS (Haarala 1994). Vid revideringen rensades över hälften av de gamla uppslagsorden ut och i stället togs nya lemman in i PS. Också själva ordartiklarna förkortades och omarbetades.

Ordboksbasen för KS består alltså till stor del av samma ordboksbas som föregångarna, vilket fortfarande skiner igenom i dagens KS. Över 70 procent av uppslagsorden är desamma som i NS, även om deras relativa andel har minskat då nya lemman tagits in.

Det nya materialet för KS kan beskrivas som öppet; underlaget består av en elektronisk beläggsamling på 260 000 ord, som redaktörerna vid sidan av det övriga redaktionella arbetet fortlöpande kompletterar med nya ord och uttryck ur varierande källor. Att ordboksbasen är öppen innebär bl.a. att redaktionen vid valet av nya lemman måste ta ställning till vilka ord som är allmänspråkliga respektive termer inom ett visst specialområde. Eftersom KS också har rollen av en preskriptiv ordbok ingår drygt 1 000 språkvårdande rekommendationer. Vid uppdateringar ses dessa språkliga rekommendationer över i samarbete med de finska språkvårdarna på Språkinstitutet.

4. Skevheter och problem i ordboksbaserna

Som framgått ovan har de fyra ordböckerna en lång historia. Dialektordböckerna beskriver huvudsakligen de traditionella finska respektive svenska dialekterna som talades på landsbygden under 1900-talets första decennier. De är inte inriktade på dagens regionala, utjämnade dialekter, vilket skulle kräva insamling av ett nytt uppdaterat material. För VKS har insamlingen av källmaterial genomgått många olika skeden, men eftersom ordboken beskriver det äldre finska skriftspråket har det bevarade källmaterialet inte förändrats över tid. Däremot har sättet att tillgodogöra sig materialet inom redaktionen utvecklats under årens lopp.

Till skillnad från de historiska, dokumenterande ordböckerna förändras föremålet för beskrivningen från dag till dag i en modern allmänspråklig ordbok med syfte att beskriva och normera det aktuella språkbruket. Det är därför en stor utmaning för KS att hålla sig i takt med tiden; se vidare diskussion om kravet på ständig uppdatering av KS i avsnitt 4.2.

4.1. Representativitet och materialets pålitlighet i äldre ordböcker

Syftet med en dialektordbok eller en historisk ordbok är att beskriva och dokumentera dialekterna eller ett äldre språkskede och därför ges ofta rikligt med autentiska exempel i ordboksartiklarna. Målsättningen för en allmänspråklig ordbok är däremot att beskriva och exemplifiera det samtida språket för språkbrukarna. KS bygger på sina föregångare, och den äldre av dem, NS, grundar sig på en stor beläggsamling. I NS förekommer ofta dels autentiska exempel, dels förkortade exempel som belyser rektion och typiska kollokationer. I KS har exemplen för det mesta kraftigt nedkortats och autentiska exempel är sällsynta. I de få fall autentiska exempel ges är de i regel hämtade ur vissa unika källor såsom Bibeln.

Ett ansenligt antal lemman och exempel i NS är excerperade ur skönlitterära texter. I källorna för ordboken finns många avledningar och parallellformer och för finskan typiska s.k. deskriptiva eller onomatopoetiska ord. I (1) återges ur NS några prov på lemman som vid redigeringen av PS rensades ut, därför att de blivit gammaldags, ovanliga eller verkar konstiga i dagens språkbruk. Exempelorden i (1) har alla ett skönlitterärt sammanhang.

(1) **häkättää** (onomatopoetiskt) skratta på ett visst sätt (Ilmari Kianto)
hälläkkä (deskriptivt) om flicka: ’toka, fjolla’ (Aleksis Kivi)

> **hämäröittää** (sällsynt avledning av finska *hämärä* 'dunkel, skum; suddig') 'göra dunkel' (Otto Manninen)
> **himerä**, **himeä** (sällsynt) jfr finska *himmeä* och *hämärä* 'dunkel, skum; dämpad; oklar' (F.E. Sillanpää, Aleksis Kivi) (våra egna övers.)

Inom äldre finsk lexikografi har litterära källor haft en stark ställning och skönlitterära exempel har fått stort utrymme. I NS återges de som citat med en namngiven författare som källa. En orsak till att skönlitteraturen utnyttjades var att de finska ordböckerna var knutna till det nationella projektet att höja det finska språkets status, därför sågs skönlitterära författare som förebilder. Från 1960-talet framåt har sakprosa och framför allt korpusar bestående av presstexter fått större roll vid uppbyggnaden av ordboksbaser för allmänspråkliga ordböcker (jfr Svensén 2004:54–59; Forsberg & Holmer 2024).

I det följande vill vi lyfta fram representativiteten när det gäller olika områden inom språkbruket. Vilka skevheter och lemmaluckor påträffas i ordböckerna? För dialektordböckernas del har det sedan gammalt varit mycket vanligt att förutom att beskriva själva ordförrådet också ge etnologiska, s.k. sakliga, materiella, uppgifter om orden. Detta kan ses som ett naturligt element, typiskt för dialektordböckerna, men det har också lett till en slagsida, där ingående beskrivningen av bondekulturen kraftigt dominerat på bekostnad av andra levnadssätt och kulturer. I Lilja (1996) lyfts denna snedvridning till förmån för bondekulturen fram med utgångspunkt i en undersökning av traditionssamlande arkiv i Uppsala 1914–1945. Lilja (1996:26) konstaterar att "[d]en folkkultur som undersöktes [---] nästan uteslutande [har] varit liktydig med landsbygdens och böndernas". Detta gäller även för dialektordböckerna SMS och FO.

Det kan ibland vara utmanande att avgöra vad representativitet innebär när det gäller en dialektordbok. Utgående från arbetet

med FO har redaktörerna lagt märke till att uppgifter om vissa ord eller betydelser kan saknas i ordboksbasen från ett område eller en socken. Orsaken kan vara att insamlaren förbisett eller missat ordet, men det kan också vara så att ordet eller betydelsen inte används i dialekten i detta område. Det är därför viktigt att redaktören tolkar uppgifterna om ordets utbredning noggrant och korrekt. När det gäller materialet för FO har redaktionen också sett en geografisk snedvridning i att det insamlade materialet från stora delar av Österbotten samlats in av vetenskapligt utbildade insamlare i mitten och slutet av 1900-talet. Uppgifterna från Österbotten tenderar att vara mer detaljerade, pålitliga och mångsidiga i fråga om uttal, betydelser och exempel än i materialet från Nyland, som ofta är äldre och insamlat med mindre systematik i slutet av 1800-talet.

En slagsida för dialektordböckerna kan också vara manér hos vissa kreativa upptecknare. Det kan i FO:s material handla om rena skrivbordskonstruktioner av produktiva partikelverb eller flerledade sammansättningar, om ett stort antal pejorativa benämningar på kvinnor eller ovanligt många ord för sex och könsumgänge från samma orter. I de fall när det finns återkommande mönster där ord inom vissa områden är överrepresenterade i samlingar från en viss ort är det viktigt att redaktörerna känner sitt material väl, är medvetna och idkar källkritik. Lösningen i de här fallen är helt enkelt att utelämna sådana belägg vid redigeringen av ordboksartiklar.

SMS:s arkiv omfattar som sagt över 8 miljoner ordsedlar, det är därför ingen större överraskning att beläggkorten inte genomgående håller en jämn kvalitet. En del av insamlarna hade språkvetenskaplig utbildning, andra var lekmän. En del av lekmännen hade också flyttat och bott på flera olika orter, vilket gjorde dem till mindre pålitliga informanter för en viss dialekt. Med tiden har erfarenheten att arbeta med materialet inom SMS-redaktion gett upphov till en gemensam insikt om hur pålitliga de olika käl-

lorna är, något som även har beaktats i ordbokens redigeringsmanual.

Utmaningar när det gäller materialets pålitlighet i SMS kan bero på att en del orduppgifter har samlats in med hjälp av frågelistor som behandlar ett begränsat ämnesområde, t.ex. skörd, fiske eller olika jordbruksredskap. I vissa fall har det funnits bakgrundsuppgifter och frågeställaren kan ha gett ett dialektord som bjudord. Informanten kan då ha angett att ordet också används i hens dialekt, även om detta inte får stöd i andra källor från samma ort.

Misstankar när det gäller materialets pålitlighet kan också uppstå när en redaktör möter ett enstaka enda belägg. Har den som besvarat frågan inte haft en korrekt minnesbild eller handlar det om ett ord som bara använts i en mindre krets, inom en familj? Den här sortens ströbelägg förekommer även då informanter besvarat frågor om olika delar på ett bruksföremål. Det kan handla om att informanten i stunden använt ett tillfälligt beskrivande ord för just den delen av bruksföremålet.

I ordboksartiklarna syns detta ofta genom en spretighet i betydelsebeskrivningen. Typiska exempel på detta ur SMS är en del av de vaga betydelser som ges i artikeln *aluspuu* 'nedre el. undre trä'. Artikeln är indelad i fem betydelsemoment som anger olika typer av ribbor eller stänger och i betydelsemoment 5 ges ytterligare "muuta" 'övrigt' enbart i form av exempel. Artikeln *etupuu* 'främre trä' är på liknande sätt indelad i tre betydelsemoment, varav de två första avser främre trä på en släde eller en vävstol, och i det tredje ges enbart exempel. Enligt dagens uppdaterade redigeringsprinciper strävar redaktörerna efter att undvika lösningar där exempel ges utan betydelseförklaring i egna betydelsemoment.

I fråga om dialektordböcker påverkar även samhälleliga förändringar, t.ex. när det gäller vad som är politiskt korrekt att publicera i vår tid jämfört med då materialet samlades in för över hundra år sedan. På FO har redaktionen noterat en skevhet när det gäller pejorativa benämningar på folkgrupper. En jämförelse av antalet

sammansättningar som anger pejorativa personbeteckningar under huvudorden *finne* 1 och *ryss, rysse* 1 visar att redigeringspraxis justerats på denna punkt; antalet pejorativa sammansättningar med *finn-* som förled (som utgavs på 1980-talet) är märkbart fler än sammansättningarna med *ryss-* som förled (publicerade 2023). Här uppstår lätt ett dilemma då ordboksredaktören ska navigera mellan vad som är politiskt korrekt idag, och i vilken utsträckning vi ger synlighet åt ett äldre samhälle som återspeglas i ordboksbasen.

VKS, som beskriver det tidiga finska skriftspråket utifrån skriftliga källor utgivna 1543–1810, domineras av religiöst språk, såsom Agricolas översättning av Nya testamentet och andra bibelöversättningar samt lagspråk i lagar och förordningar. Förklaringen till att dessa genrer är så starkt företrädda är att lejonparten av de texter som skrevs och utgavs på finska under perioden bestod av religiösa och administrativa texter. Vid redigeringen strävar redaktörerna efter att ge exempel som företräder varierande texttyper med jämn fördelning från olika decennier.

4.2. Krav på ständig uppdatering i en samtidsordbok

Då en äldre ordbok utgör ordboksbas för en efterträdare kan detta ses som ett slags halvfabrikat till en ordbok. Det försnabbar i viss mån redigeringen, men kan göra arbetet mer invecklat och begränsa möjligheterna att skapa nya lösningar. I synnerhet när det gäller exempel finns det en uppenbar risk att en del uttryck och fenomen föråldras oförmärkt. Övergången till KS från NS, som utkom 1951–1961, via PS på 1990-talet, har gått via processer av att sovra, utmönstra och komplettera. Till följd av resursbrist har det inte heller varit möjligt att påbörja arbetet helt från början. KS innehåller därför genom sin bakgrund flera olika tidsskikt där nya och gamla artiklar varvas.

4.2.1. Korpusmetoden

I dagens digitala värld förväntas en allmänspråklig ordbok vara baserad på stora korpusmaterial (Svensén 2004:58). Atkins & Rundell (2008:95–96) framhåller att lexikografin gärna ska vara evidensbaserad med en trovärdig förankring i det allmänna varierande språkbruket, vilket enligt dem kräver en tillräckligt stor och mångsidig korpus:

> for lexicography, the best, most useful kind of corpus is one that combines very large volumes of data with diversity in a number of broad categories (like mode, medium, and domain), and a level of linguistic and textual annotation which aspires to high quality but does not seek perfection. [---] the single biggest benefit is the access it gives us to the 'regularities' of the language – the typical and recurrent features and patterns which make up the norms that lexicographers seek to identify and describe. There is no longer any serious argument about whether or not to use corpora in creating dictionaries.

För det finska språket saknas en heltäckande korpus som skulle ge tillräckligt underlag för att mångsidigt belysa olika användningsområden. Allmänt tillgängliga korpusar utnyttjas däremot regelbundet i arbetet med KS. I övrigt följer redaktörerna med vad som händer omkring dem, gör Google-sökningar och frågar experter. Vid uppdateringen av en ordbok behöver korpusmetoden inte heller alltid vara den enda eller den bästa lösningen (jfr Svensén 2004:56–58).

De mest centrala korpusarna finns i *Språkbanken i Finland* med konkordansverktyget Korp, som anpassats till finska språkbanken. Språkbanken innehåller språkresurser inom digital humaniora och samhällsvetenskap och har ett brett utbud av text- och talkorpusar. Den omfattar ett flertal olika texttyper, men dagspress och ore-

digerade texter från diskussionsforum på internet dominerar för finskans del. En viktig poäng är att de oredigerade texterna delvis innehåller osakliga och alltför talspråkliga formuleringar. Men de kan vara ett hjälpmedel för att belysa bruket av vissa ord och uttryck.

4.2.2. Äldre ordböcker som ordboksbas

I KS påträffas gamla artiklar sida vid sida med nytillkomna artiklar. Nya ord föds genom nya fenomen och samhällsförändringar. Ett typiskt exempel är den administrativa termen *hyvinvointialue* 'välfärdsområde' som togs in i KS 2022, eftersom redaktionen i förväg kände till att termen skulle införas officiellt 1.1.2023. Också tillströmningen av lånord från engelska såsom verbet *chillata* 'chilla' har ökat. Ordet kom in i KS vid uppdateringen 2016, se (2).

(2) **chillata** [tš-] slg. oleilla, hengailla, chillailla. *Nuorisoporukka chillasi nurtsilla*. (KS 2016)
'chilla' [tš-] slang. 'hålla till, hänga, chilla' *Ungdomsgänget chillade på gräsmattan*. (förf. övers.)

I redigeringsprocessen har gamla artiklar i regel uppdaterats i större utsträckning än nya uppslagsord har kommit in i ordboken. Vid uppdateringarna av KS 2022 och 2024 omarbetades ca 1 000 artiklar men bara omkring 400 nya artiklar infördes. De redaktionella resurserna har varit alltför små för heltäckande systematiska genomgångar. Därför kan man på många ställen i ordboken hitta sådana artiklar som skulle kräva mer omfattande uppdatering. Exempelvis artikeln *kaulin* 'kavel' i (3) står fortfarande i samma form som den hade i PS 1990. Sett med dagens ögon framstår artikeln *kaulin* gammaldags och stereotypisk, eftersom det enda exemplet berättar om ett huskors. Å andra sidan var motsvarande artikel i föregångaren NS från 1950-talet i (4) ännu mer rakt på sak, för där belyser exemplet hur en hustru slår sin man i huvudet med en kavel.

(3) **kaulin** (puu)tela jolla kaulitaan taikinaa. Leik. ”naisvallan” symbolina. *Kaulinta heilutteleva pirttihirmu*. (PS 1990, KS 2004–)
’rulle av trä med vilken man kavlar deg’. Lekfullt symbol för ”kvinnomakt”. *Ett huskors som svingar en kavel*. (förf. övers.)

(4) **kaulin** puutela, jota edestakaisin pyörittämällä taikina kaulitaan. *Löi miestään kaulimella päähän*. (NS 1950-tal)
’rulle av trä, med vilken man kavlar deg’. *(Hon) slog sin man med en kavel i huvudet*. (förf. övers.)

Redaktionen för KS har med vissa intervall försökt systematisera genomgången av äldre material utifrån användningsområde och ämne. Under de senaste åren har redaktionen tagit itu med könsstereotypier genom att systematiskt gå igenom ordboken och granska hur orden har använts. Bland annat har användningen av orden *tyttö* ’flicka’ och *poika* ’pojke’ i exemplen granskats och åtgärdats. Ordet *poika* har varit överrepresenterat i olika negativa sammanhang, såsom i exempelfrasen *Pojan ryökäle* ’pojkslyngel’ i artikeln *ryökäle* i (5) A. Vid uppdateringen 2022 lades det i stället in ett mer neutralt exempel, se (5) B.

(5) A **ryökäle** – – *Pojan ryökäle*. (KS –2021)
’lymmel, slyngel, skurk’ *Pojkslyngel*.
(5) B **ryökäle** – – *Kuka ryökäle on syönyt keksipurkin tyhjäksi?* (KS 2022–)
– –’*Vem sjutton har ätit upp alla kakor i kakburken*’ (förf. övers.)

Ordet *tyttö* ’flicka’ har tidigare används mer generellt i betydelsen ’ung kvinna’. Exempel på denna användning ges i (6), betydelsemoment 4 b, nedan. Där vi kan följa hur det i KS 2020 lagts in en kommentar om att detta i synnerhet gäller i äldre språkbruk.

(6) **tyttö** – – 4. b. tyttöystävä, rakastettu, heila. *Puistossa oli meripoikia tyttöineen*. (KS –2018)
flicka – – 4.b. 'flickvän, käraste, flamma'. *I parken var (några) unga sjömän med sina flammor*.
tyttö – – 4. b. **vars. aik.** tyttöystävä, rakastettu, heila. *Puistossa oli meripoikia tyttöineen*. (KS 2020–).
flicka – – 4. b. 'i synnerhet i äldre språkbruk' [---] (förf. övers.)

Lösningen har alltså dels varit att göra ändringar i exempel, såsom i (5) ovan, dels att markera att en viss användning av ett ord är föråldrad, vilket exemplifieras i (6).

4.2.3. Gränsdragning och lemmaselektion

I KS-nätordbok finns det i varje artikel en länk genom vilken användarna per e-post kan sända respons till redaktionen. Många ändringar och uppdateringar i ordboken har gjorts tack vare respons från användarna. Till exempel termer inom fysiken uppdaterades för några år sedan genom tips från en aktiv ordboksanvändare.

Eftersom ordboksbasen för KS är öppen och inte har fastställts i förväg i form av en korpus måste alla beslut om gränsdragning tas inom redaktionen. Exempelvis vilka ord inom ett visst specialområde som ingår i allmänspråket. Trots att KS är en allmänspråklig ordbok tas också termer från en mängd olika specialområden med i ordboken. Det kan handla om ord och termer som lekmän stöter på då de läser en dagstidning eller tar del av sin egen eller en familjemedlems patientjournal. Frågan är om vissa specialområden är överrepresenterade och andra specialområden fått för liten uppmärksamhet, vilket också påverkar vid uppdatering av ordboken.

Särskilt medicinska termer förekommer ofta i KS, men också ämnesområden som sport och matlagning är väl företrädda. Då sökningar med hjälp av markörer för användningsområden gjor-

des visade resultatet att det i KS finns sammanlagt: 2 180 sporttermer och 2 126 medicinska, 1 115 inom matlagning, 455 inom datateknik, men bara 173 inom handarbete. I viss mån har redaktörernas egna särintressen och specialkunskaper styrt de här processerna. Redaktörerna kan med hjälp av markörerna söka ord och termer inom ett visst ämnesområde för granskning och genomgång. Hittills har bl.a. språkvetenskapliga och zoologiska termer granskats gruppvis.

En annan fråga inom ramen för användningsområden berör allmänspråkets förhållande till det informella språket, alltså till vardagsspråket, dialekter och slang. I vilken utsträckning kan man säga att allmänspråket innehåller stilistiskt informella drag? Eller handlar det snarare om att vissa texttyper eller genrer i så hög grad förändrats i en mer informell riktning att man inte längre kan tala om ett helt neutralt allmänspråk? Tidningsspråk och talspråket i tv-program tillåter för finskans del numera en tämligen stor variation och innehåller fler och fler anglicismer. Linjedragningen gentemot lånord har alltmer blivit en kompromiss mellan den lavinartade tillströmningen och ett försiktigt godkännande och kodifiering av lånord. Vid valet av nya uppslagsord i KS påverkar även språkvårdens rekommendationer, eftersom ordboken också är normerande för språkbruket. Språkvården har i allmänhet favoriserat förfinskade former, och försökt undvika skrivformer som är främmande för finskan: bland annat har w-bokstaven fram till nyligen i allmänhet ersatts med v. Först i den version av ordboken som publiceras 2024 har det för verbet *vokata* 'tillreda wokmat' i samråd med språkvården förts in en parallellform *wokata* stavat med w. I årets uppdatering av KS ändras också stavningen av ordet *unisex*, till stavning med bokstaven x – fram till 2022 hade KS formen *uniseksi*, en stavning med ett -ks- enligt finska stavningsregler. Formen slog inte igenom eftersom den uppfattades som komisk, bl.a. eftersom det blir mångtydigt och kan tolkas som *uniseksi* 'sömnsex; drömsex'.

5. Sammandrag

Gemensamt för de fyra finländska enspråkiga ordböckerna SMS, FO, VKS och KS som vi har diskuterat är att deras tillkomsthistoria sträcker sig långt tillbaka i tiden. När det gäller ordboksbasen för dessa ordböcker och vissa skevheter och utmaningar som de innehåller är det just tidsaspekten som kan vara en viktig förklaring till utmaningarna.

Materialet för en ordbok är förmodligen aldrig fullkomligt; inom ordboksredaktionerna har det därför utvecklats olika lösningar för att vid redigeringen av ordboken rätta till och komplettera de skevheter som finns i materialet. I vissa fall fylls ordboksbasen på, t.ex. genom att redaktionen för KS excerperar nya ord och nya betydelser som vid uppdatering kan tas in i ordboken. Redaktionen för VKS har i flera omgångar breddat sin ordboksbas. Dessutom har man t.ex. vid redaktionerna för dialektordböckerna SMS och FO lärt sig att handskas med problem genom källkritik; lösningen är då vanligen att utelämna opålitliga belägg. I vissa fall kan det också vara ett medvetet val att låta skevhet och överrepresentation av vissa ord, betydelser och användningsområden framträda i ordboken, därför att syftet med ordboken är att beskriva en dialekt eller ett äldre språkbruk. När det är frågan om långvariga, omfattande ordboksprojekt är det inte heller möjligt att lättvindigt förändra redigeringsprinciper och praxis, eftersom ordboken helst ska vara så logiskt sammanhållen och enhetlig som möjligt från a till ö.

För såväl den allmänspråkliga ordboken KS som dialektordböckerna SMS och FO gäller att gränsen mellan vad som idag uppfattar som politiskt korrekt eller partiskt förändrats sedan materialen för ordböckerna samlades in. När det gäller en historisk dialektordbok vars material återspeglar ett äldre språkbruk och ett samhälle för hundra år sedan, innebär det en balansgång. Hur låter man ett äldre samhälle med andra värderingar återspeglas men

undgår att publicera artiklar med betydelser och exempel som i dagens värld kan uppfattas som stötande?

Litteratur

Ordböcker, korpusar och digitala resurser.

FO (2013–) = *Ordbok över Finlands svenska folkmål*. Helsingfors: Institutet för de inhemska språkens webbpublikationer 33. <kaino.kotus.fi/fo/> (juni 2024).

Korp = Språkbankens konkordansverktyg, anpassad till Finska språkbanken. <kielipankki.fi/korp/>.

KS (2004–) = *Kielitoimiston sanakirja*. Helsingfors: Institutet för de inhemska språkens webbpublikationer 35. <kielitoimistonsanakirja.fi> (juni 2024).

NLO (1997). = *Nordisk leksikografisk ordbok*. Henning Bergenholtz, Ilse Cantell, Ruth Vatvedt Fjeld, Dag Gundersen, Jón Hilmar Jónsson, Bo Svensén (red.). Skrifter utgivna av Nordiska föreningen för lexikografi. Skrift nr 4. Universitetsforlaget.

NS (1951–1961) = *Nykysuomen sanakirja*. Matti Sadeniemi, Jouko Vesikansa m.fl. (red.). Suomalaisen Kirjallisuuden Seura, WSOY.

PS (1990–1994) = *Suomen kielen perussanakirja*. Kotimaisten kielten tutkimuskeskus, Painatuskeskus.

PS (1997) = *CD-Perussanakirja*. Kotimaisten kielten tutkimuskeskus, Oy Edita Ab.

SMS (2012–) = *Suomen murteiden sanakirja*. Helsingfors: Institutet för de inhemska språkens webbpublikationer 30. <kaino.kotus.fi/sms/> (juni 2024).

Språkbanken. <kielipankki.fi/sprakbanken/> (juni 2024).

Vendell, Herman (1904–1907): *Ordbok över de östsvenska dialekterna*. Helsingfors: Svenska litteratursällskapet i Finland.

Wessman, V. E. V. (1925–1932): *Samling av ord ur östsvenska folkmål.* Tillägg till H. Vendells ordbok över de östsvenska dialekterna. Helsingfors: Svenska litteratursällskapet i Finland.

VKS (2014–) = *Vanhan kirjasuomen sanakirja*. Helsingfors: Institutet för de inhemska språkens webbpublikationer 38. <kaino.kotus.fi/vks/> (juni 2024).

Annan litteratur

Ahlbäck, Olav (1946): *Studier över substantivböjningen i Finlands svenska folkmål*. Studier i Nordisk Filologi 33–34. Helsingfors: Svenska litteratursällskapet i Finland.

Atkins, Sue & Michael Rundell (2008): *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Forsberg, Markus & Louise Holmer (2024): Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text. I: *LexicoNordica* 31 (denna volym).

Haarala, Risto (1994): De viktigaste ordböckerna över finskt allmänspråk. I: *LexicoNordica* 1, 53–61.

Haarala, Risto (1999): Vem är de finska ordböckerna gjorda för? I: Peter Slotte, Pia Westerberg & Eva Orava (red.): *Nordiska studier i lexikografi* 4. Rapport från Konferensen om lexikografi i Norden Esbo 21–24 maj 1997. (Skrifter utgivna av Nordiska föreningen för lexikografi. Skrift nr 5.). 149–158.

Lilja, Agneta (1996): *Föreställningen om den ideala uppteckningen. En studie i idé och praktik vid traditionssamlande arkiv – ett exempel från Uppsala 1914–1945*. Skrifter utgivna genom Dialekt- och folkminnesarkivet i Uppsala, Ser. B: 22. Uppsala.

Onikki-Rantajääskö, Tiina (2011): Mestari–kisälli-mallin soveltuvuus fennistiikan historiaan. I: *Virittäjä* 2011, 542–574.

Ruppel, Klaas & Caroline Sandström (2014): Stora finska ordböcker i ett historiskt perspektiv. I: *LexicoNordica* 21, 141–160.

Sandström, Caroline (2013): Från folkmålsstudier till interaktionell dialektologi. I: *Folkmålsstudier* 51, 87–113.

Sandström, Caroline (2016): Perspektiv på svensk lexikografi i Finland med Ordbok över Finlands svenska folkmål och Finlandssvensk ordbok som exempel. I: Anna Helga Hannesdóttir (red.): *Framtidens lexikografi. Rapport från ett symposium*. Mejerbergs arkiv för svensk ordforskning 42. Göteborg. 75–109.

Setälä, E. N. (1896): Suomalaisen sanakirjatyön ohjelmasta. I: *Suomi* 3:13, 89–95.

Svensén, Bo (2004): *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. Andra omarbetade och utökade upplagan. Stockholm: Norstedts Akademiska Förlag.

Vikør, Lars S. (1999): Fleirgenerasjonsordbøker og tida. I: Peter Slotte, Pia Westerberg & Eva Orava (red.): *Nordiska studier i lexikografi* 4. Rapport från Konferensen om lexikografi i Norden Esbo 21–24 maj 1997. (Skrifter utgivna av Nordiska föreningen för lexikografi. Skrift nr 5.). 395–405.

Vilppula, Matti (1999): Ordbok över finska dialekter och evigheten. I: Peter Slotte, Pia Westerberg & Eva Orava (red.): *Nordiska studier i lexikografi* 4. Rapport från Konferensen om lexikografi i Norden Esbo 21–24 maj 1997. (Skrifter utgivna av Nordiska föreningen för lexikografi. Skrift nr 5.). 407–411.

Tarja Riitta Heinonen
ordboksredaktör, fil. dr
Institutet för de inhemska språken
Hagnäskajen 6
FI-00530 Helsingfors
tarja.heinonen@kotus.fi

Caroline Sandström
ordboksredaktör, fil. dr
Institutet för de inhemska språken
Hagnäskajen 6
FI-00530 Helsingfors
caroline.sandstrom@sprakinstitutet.fi

Talspråkskorpusar som resurs för isländska ordböcker

Helga Hilmisdóttir

This article focuses on spoken corpora and how they can be used in lexicographic work involving Icelandic. A particular focus is on the *Icelandic youth language* corpus, which gives users access to both texts and recordings. I discuss three domains in which a corpus of spoken language can supplement large text corpora: 1) information on anglicisms, 2) the pragmatic aspects of adverbs, and 3) discourse functions of words or collocations. In addition to this, the paper discusses an alternative method for presenting the peculiarities of spoken language: *Samtalsorðabók*, a dictionary that builds on transcribed excerpts of conversation with sound clips.

1. Inledning

Denna studie fokuserar på talspråkskorpusar och vad de kan bidra med i lexikografiskt arbete på Island. Enligt den europeiska ordbokstraditionen bygger ordboksarbete först och främst på skrivna källor, nuförtiden oftast stora textkorpusar (se diskussion i Eyþórsson 2005; Fjeld 2008; Selback & Svardal 2021). Detta gäller även moderna, isländska ordböcker som *Íslensk nútímamálsorðabók* 'Ordbok över nutida isländska' (ÍNO), som är den enda isländska definitionsordboken som ständigt uppdateras (se Jóhannsson & Úlfarsdóttir, denna volym).

I ÍNO förekommer en del ord som först och främst förknippas med talspråk, t.ex. svordomar, olika dialektala ord, slang och partiklar. Precis som när det gäller andra ordböcker kommer talspråksfenomen dock oftast in ordboken via stora textkorpusar. Uppslagsorden beskrivs utifrån belägg t.ex. från litterära dialoger eller direkta anföringar i dagstidningar och texter som publicer-

as på webben. Med andra ord beskrivs det som vi uppfattar som talspråk på skriftspråkets villkor (jfr Linell 1982 om ”the written language bias”). På så sätt kan man säga att begreppet *talspråk* tolkas som en stilmarkör (formellt-informellt) snarare än beteckning för språkets medium (talspråk-skriftspråk) (jfr kategorisering i Svensén 2004:379). Samtalsforskningen har dock visat de senaste decennierna att ord och fraser kan ha funktioner i samtal som får en ganska styvmoderlig behandling i ordböcker (Hilmisdóttir 2021). Detta gäller i synnerhet adverb, interjektioner och diskurspartiklar. För att fånga ordens funktioner i det talade språket har Hilmisdóttir (2024) utvecklat en webbordbok som beskriver talspråksfenomen som förekommer i inspelade samtal. *Samtalsorðabók* ’Samtalsordboken’ är en deskriptiv ordbok där användarna har tillgång till såväl ljud som text.

I denna artikel ligger fokus på korpusar som består av talat språk. Syftet med artikeln är att diskutera hur talspråkskorpusar skulle kunna användas i lexikografiskt arbete på Island och eventuellt i övriga Norden och andra ordboksmiljöer. Frågeställningen är som följer: Hur kan talspråkskorpusar användas för att ge en mer nyanserad bild av modern isländska och för att inkludera mer material som är representativt för vardagliga samtal och inte minst ungdomars språkbruk? Vad hör hemma i en traditionell ordbok och hur kan en ordbok som *Samtalsorðabók* komplettera den? Det material som studien baseras på består av tre korpusar som innehåller talat språk: den isländska gigakorpusen *Risamálheildin*, textkorpusen *Íslenskt textasafn* och samtalskorpusen *Íslenskt unglingamál* ’Isländskt ungdomsspråk’. I diskussionen fokuserar jag på tre olika fenomen i talspråkskorpusarna: 1) anglicismer, 2) pragmatiska funktioner och 3) diskursstrukturerande element.

2. Talspråk i moderna isländska ordböcker

För att ge inblick i hur talspråket har representerats i isländska ordböcker kommer jag att inleda med en jämförelse av tre olika verk: *Islandsk-dansk ordbog* från 1920–1924, *Íslensk orðabók* från 1963 och *Íslensk nútímamálsorðabók* som publicerades på webben 2016.

Islandsk-dansk ordbog av Sigfús Blöndal (1920–1924 och supplement från 1963) var den första ordboken som behandlade modern isländska. Ordboken är deskriptiv till sin karaktär och enligt redaktören var det hans syfte att fånga bl.a. det som han kallar för *almugesprog* 'allmogespråk'. I ordboken som har 154 000 uppslagsord förekommer ett ganska stort antal ord och uttryckssätt som fångats upp i talspråk. Huvudkällorna för talspråksinslagen är först och främst ordinsamlingar på fältet, bl.a. ordlistor som samlades in av redaktörens vänner och kollegor (Blöndal 1920–1924:VIII; Ingólfsson 1997:22; Eyþórsson 2005:16). Talspråksinslagen består av dialektala ord och fraser (Ingólfsson 1997), svordomar, barnspråk, lånord (Óskarsson 1997) och olika diskursmarkörer och interjektioner. En del lånord, dvs. de som ansågs mer problematiska än andra, markerades dock som "Udenlansk Laaneord, alm. i daglig Tale, is. i Byerne; ikke anerkendt i Skriftsprog" (se även Svavarsdóttir & Jónsdóttir 2020:123). Under redigeringsarbetet lades stor vikt vid transkribering av uttal, inklusive lokalt betingade uttalsvarianter (Árnason 1997; Hilmisdóttir & Svavarsdóttir 2023:153–156).

Den första enspråkiga isländska ordboken, *Íslensk orðabók*, kom först ut 1963 men kom ut i ny utgåva 1983, 2002 och 2007. Ordboksbasen bygger delvis på Blöndals ordbok och detta gäller inte minst dialektala och arkaiska ord och uttryck som först och främst förknippas med talspråk. Ordboken bygger även på två stora ordboksarkiv som samlades in av det Lexikografiska institutet i Reykjavik (isl. *Orðabók Háskólans*) men förvaras på Árni

Magnússon-institutet för isländska studier: a) skriftspråksarkivet som består av sedlar med belägg som excerperades från manuskripter och tryckta böcker från 1540 till 2000, och b) talspråksarkivet som består av sedlar som visar kommentarer om ord och fraser som kom från Lexikografiska institutets informanter. *Íslensk orðabók* har dock färre uppslagsord, ca 88 000. I förordet beskrivs den som en deskriptiv ordbok som omfattar ”det språk som talas vid alla omständigheter” (förf. övers.). Samtidigt understryks ordbokens normativa roll och att den ska ge fingervisningar om erkänd språkanvändning (*Íslensk orðabók* 2002:VI). Den har med andra ord en språkpuristisk utgångspunkt (*Íslensk orðabók* 2007:VI–VII). Språklig variation framhävs t.ex. inte på samma sätt som i Blöndal (t.ex. färre dialektala ord) och den är konservativ angående ord som anses främmande. I den andra och senare upplagor blev inställningen mindre konservativ. In i lemmalistan togs en del anglicismer som tillhör det vardagliga ordförrådet som t.ex. *ælæner* ’eyeliner’ och slangord och uttryck som har sitt ursprung i ungdomsspråk som *bömmer* ’bummer’ och *sjitt* ’shit’. De markerades dock som slang eller tveksamma främmandeord. I den sista reviderade utgåvan av *Íslensk orðabók* från 2007 reviderades även beskrivningar av olika markörer i talspråk som t.ex. interjektionen *jæja* ’jaha, nå’ som används bl.a. för att markera topikövergångar och attityder. I den nya utgåvan framhävdes mer markörernas funktioner i samtal.

Som tidigare nämnts är *Íslensk nútímamálsorðabók* den enda ordboken över modern isländska som uppdateras löpande (se Jóhannsson & Úlfarsdóttir, denna volym). Den har inte samma lemmalista som de två tidigare ordböckerna, utan bygger på en lista som utvecklades för ordböcker från isländska till nordiska språk (Jónsdóttir & Úlfarsdóttir 2011). Målet med ordboken, som har 56 000 uppslagsord, är i första hand att beskriva centralt, nutida språkbruk. Ordboken har t.ex. endast ett fåtal lokalt präglade ord. Vid ordboksarbetet har redaktörerna delvis använt textkorpusar

som *Íslenskt textasafn* och *Risamálheildin* (se avsnitt 3). ÍNO är än så länge mindre än de tidigare två ordböckerna vilket kan medföra konsekvenser för beskrivningen av talspråk, i synnerhet dialektala former och nya lånord. Eftersom ordbokens främsta målgrupp är unga språkanvändare och andraspråkstalare är tanken att skriva korta och relativt enkelt formulerade förklaringar, åtminstone i jämförelse med *Íslensk orðabók*. Det betyder att ord som är ganska komplicerade och har många olika funktioner, såsom partiklar, har minimala förklaringar med huvudvikt på semantiskt innehåll. Till exempel beskrivs uppslagsordet *nú* som ett temporalt adverb utan att kommentera ordets funktion som icke-temporal diskurspartikel (jfr Hilmisdóttir 2016). Den icke-temporala funktionen syns dock i artikeln som två fasta eller halvfasta fraser: *er það nú* <*bíll*> 'ska det här föreställa <en bil>?' och *þó það nú væri* 'naturligvis'. I ÍNO har varje uppslagsord en ljudfil där man kan höra ordets uttal (uppläsning av skådespelare) men inga kommentarer finns om uttalsvarianter. De tre ordböckera jämförs i Tabell 1.

	Islandsk-dansk ordbog (1920–1924)	**Íslensk orðabók (2002)**	**Íslensk nútímamáls-orðabók (inget datum)**
uppslagsord	154 000	88 000	56 000
källor för talspråk	ordinsamlingar på fältet	Islandsk-dansk ordbog och ordinsamlingar	lemmalista som gjordes för ISLEX med tillägg
dialektalt ordförråd	stor vikt vid dialektala ord	en del dialektala ord	sporadiskt
lånord	ganska liberal inställning	mest mycket etablerade ord	mest mycket etablerade ord
partiklar	minimal beskrivning, fasta idiomatiska fraser	etablerade partiklar förklaras med synonymer och ibland kommenterar om funktion	enkel beskrivning, semantiskt innehåll och synonymer

uttal	transkription inkl. varianter	inga uttals-anvisningar	uppläsning av isolerade ord

Tabell 1: Jämförelse av talspråksinslag i *Islandsk-dansk ordbog*, *Íslensk orðabók* och *Íslensk nútímamálsorðabók*.

Som tabellen visar har utvecklingen gått från att visa språklig variation, inklusive talspråkliga och dialektala drag, till att lägga huvudtyngd på att visa normaliserat språk med huvudtyngd på skriftspråk. Med tanke på den samhällsutveckling som har ägt rum på Island de senaste decennierna där antalet andraspråkstalare har ökat markant finns det dock ett starkt behov för att beskriva talspråksfenomen. Talspråkskorpusar är i det avseendet en viktig resurs.

3. Isländska talspråkskorpusar

I jämförelse med skriftspråksbaserade textkorpusar är talspråkskorpusar begränsade till sitt omfång. På grund av detta spelar det stor roll vilken talare som spelas in, under vilka omständigheter och vilka samtalsämnen det rör sig om. För att ge en bra bild av vardaglig interaktion skulle det vara optimalt att ha en stor korpus med autentiska samtal som spelas in i vardagliga situationer, dvs. samtal som inte är planerade eller styrda. Det kan dock vara svårt att få in tillräckligt mycket material, i synnerhet om korpusen ska vara öppen och finnas tillgänglig på webben.

Än så länge finns det inte allmänt tillgängligt talspråksmaterial för isländska, och det material som finns består mest av transkriberingar av nyhetssändningar och intervjuer som gjorts inom språkforskning. Tabell 2 visar en översikt över tre korpusar som består helt eller delvis av talspråk: talspråksmaterial från den isländska gigakorpusen *Risamálheildin*, talspråksmaterial från textkorpusen *Íslenskt textasafn* och talspråkskorpusen *Íslenskt unglingamál*.

	Risamál-heildin	**Íslenskt textasafn**	**Íslenskt unglingamál**
år	2004–2021	2000–2006	2019–2020
genre	nyhetsprogram på radio och tv	vardagliga samtal och gruppintervjuer	gruppintervjuer
storlek	ca 77 miljoner löpord (av 2770 miljoner)	ca 0,5 miljoner löpord (av 65 miljoner)	ca 0,2 miljoner löpord
angiven kontext	en mening, dvs. från stor bokstav till punkt	en rad, 140 tecken	hela inspelningen
ljudfiler	nej	nej	ja

Tabell 2: Översikt över talspråk i isländska korpusar.

Som tabellen visar varierar storleken på de tre korpusarna mycket. Den största korpusen består av en samling nyhetsprogram som finns med i den isländska gigakorpusen. Den består av nästan 77 miljoner löpord och består dels av uppläsning av journalisternas manus, dels av korta ljudklipp från intervjuer. Det finns ingen tillgänglig information om fördelningen mellan planerat och oplanerat tal. Sökorden och meningen som ordet förekommer i visas på skärmen, dvs. från stor bokstav till punkt. Detta betyder att t.ex. interjektioner och partiklar som t.ex. *já* 'ja' eller *nei* 'nej' som ibland förekommer som en egen replik kan förekomma i korpusen utan kontext. Man kan heller inte kontrollera meningarnas yttre kontext, dvs. det som kommer före eller efter repliken i fråga. Användarna har inte tillgång till ljudfilerna för att kontrollera transkriberingen eller för att tolka beläggens prosodiska utformning.

Textkorpusen *Íslenskt textasafn*, som har runt en halv miljon löpord, består av transkriberat talspråk, vardagliga samtal och gruppintervjuer om språkattityder. Sökorden visas i sin kontext, en rad per belägg (140 tecken). När replikerna är korta betyder

det att användaren ibland kan se det som sägs före och efter det aktuella yttrandet, vilket inte är möjligt i gigakorpusen. Ljudinspelningar är däremot inte tillgängliga.

Talspråkskorpusen *Íslenskt unglingamál* samlades in i grund- och gymnasieskolor på Island skolåret 2019–2020. Syftet med insamlingen var att dokumentera unga islänningars samtalsspråk och göra en korpus för framtida forskning. Korpusen gjordes i samarbete med Tekstlaboratoriet i Oslo och finns tillgänglig på webben. Den består av 22 inspelade intervjuer med ungdomar på fyra orter runtom landet. De intervjuade delades in i två åldersgrupper: grundskoleelever i årskurs 9 (14–15 år) och gymnasieelever som gick sitt sista år i skolan (18–19 år). I korpusen deltar 120 olika talare, inklusive intervjuare. I varje intervju deltar fyra ungdomar och en moderator. Under inspelningarna ställer moderatorn några förbestämda frågor till ungdomarna, bl.a. om hur de trivs i skolan, vad de gör på fritiden och hur de ser på framtiden. I slutet spelar moderatorn upp musiksnuttar för att få igång en diskussion om musik, trender och musiksmak.

4. Nedslag i isländska talspråkskorpusar

Men vad kan man använda talspråkskorpusar till i ordboksarbete? Som Selback & Svardal (2021:104) påpekar beror det på vilka delar av ordförrådet man fokuserar på och vilken typ av ordbok det rör sig om. I traditionella definitionsordböcker är det innehållsorden som dominerar, dvs. främst substantiv och sedan utgör adjektiv och verb något mindre kategorier (Selback & Svardal 2021:97). För att fånga en stor variation innehållsord behövs dock mycket stora korpusar. Talspråkskorpusar är däremot särskilt användbara när det gäller beskrivningen av funktionsord som t.ex. prepositioner, interjektioner och pragmatiska partiklar (jfr Selback & Svardal 2021:91).

I följande tre underkapitel diskuterar jag hur talspråkskorpusar kan användas som resurs vid lexikografiskt arbete och visar samtalsutdrag som stöd: resurs för att identifiera anglicismer som är frekventa i talspråk (4.1.), resurs för att identifiera och definiera uppslagsordens pragmatiska (4.2) och diskursstrukturerande funktioner i samtal (4.3).

4.1. Anglicismer i talspråk

Isländsk språkpolitik har ofta beskrivits som konservativ och puristisk och detta har avspeglats i lexikografiskt arbete (Svavarsdóttir & Jónsdóttir 2020:120; Kristinsson 2021). Lånord och ord som uppleves som ”främmande” är fortfarande kontroversiella i det isländska språksamhället. Detta skapar problem för lexikografer som måste ta ställning till vilka ord som ska inkluderas i en ordbok och hur de bör presenteras (Svavarsdóttir & Jónsdóttir 2020:122).

En del danska och engelska lånord har funnits länge i isländskt talspråk men förekommer sällan i skrift (Kvaran 2007:17; se även diskussion i Hilmisdóttir & Peterson 2024:86–87). Förr betydde detta att det var svårt att uppskatta ordens spridning och frekvens i vardagligt språk. I och med sociala medier har detta dock ändrats och t.ex. vid uppdateringen av ÍNO har redaktörerna kunnat använda den isländska gigakorpusen *Risamálheildin* för att få fram information om förekomster av anglicismer i informellt webbspråk.

Tabell 3 visar en översikt över frekvensen på ett urval anglicismer som används i isländska samtal men finns inte i ÍNO. Materialet som jämförs är som följer: 1) talspråksmaterial från gigakorpusen (Risamálheildin, delkorpus som består av nyheter på radio och tv), 2) talspråksmaterial från textkorpusen *Íslenskt textasafn*, och 3) talspråkskorpusen *Íslenskt unglingamál*.

	Risa-málheildin	Íslenskt textasafn	Íslenskt unglingamál
random	19	0	6
kreisí/crazy	15	0	2
level	36	0	0
sökka 'suck'	9	1	3
beisiklí/basically	29	0	14
jess/yes	4	1	31
what/vott	0	0	23

Tabell 3: Belägg på anglicismer i isländska talspråkskorpusar.

Som tabell 3 visar ger talspråkskorpusar endast ett fåtal belägg av varje ord. De flesta orden i tabell 3 används frekvent i mycket informella texter som en sökning i gigakorpusens material från sociala medier visar: *random* (N=10 236), *kreisí/crazy* (N=7745), *level* (N=16 610), *sökka* 'suck' (N=13 446), *basically/beisiklí* (N=5449), *jess/yes* (N=1920) och *what* (N=0). Dokumenteringen av frekventa anglicismer är på det sättet inte ett starkt argument för att inkludera talspråkskorpusar som resurs för ordböcker som det är nu. Beläggen som där finns ger dock autentiska exempel från talspråk och de visar om och hur orden integreras prosodiskt och morfologiskt. Man kan lyssna på beläggen och höra hur de användes i sin kontext. Eventuellt kan talspråkskorpusar också ge fingervisningar om vilka anglicismer som är de mest frekventa och hur ordförrådet fördelar sig mellan åldersgrupper och olika minoritetsgrupper. Till exempel verkar substantivet *level* 'nivå' vara ett ord som används mycket av vuxna talare medan *what* förekommer enbart i ungdomsspråkskorpusen (som är den nyaste korpusen). Som tabellen visar är det i synnerhet de funktionella orden som är frekventa i talspråkskorpusarna, ord som *beisiklí*, *jess* och *what*, dvs. ord som inte är lika kontextbundna som innehållsorden som

styrs mer av diskussionsämnet. Detta visar att det finns ett behov för en mycket stor talspråkskorpus som skulle kunna ge en mera nyanserad bild av isländskt talspråk och i synnerhet innehållsord.

4.2. Pragmatiska funktioner

En talspråkskorpus kan ge oss information om ordens pragmatiska funktioner i tal. Ett exempel på det är uppslagsordet *örugglega* som är ett relativt frekvent ord i såväl tal som skrift. Enligt ÍNO har adverbet *örugglega* två betydelser (förf. övers.):

> 1) *án efa*, *vafalaust*, *áreiðanlega* 'utan tvekan, definitivt, som man kan lita på'. Ex. *það verður* ***örugglega*** *skemmtilegt í ferðalaginu* 'det blir **säkert** roligt på resan', *ertu* ***örugglega*** *búinn að læsa útidyrunum?* 'har du **definitivt** låst ytterdörren?'
>
> 2) *á sannfærandi hátt, af öryggi* 'på ett övertygande sätt, med säkerhet'. Ex. *liðið sigraði* ***örugglega*** *á mótinu* 'laget segrade **på ett övertygnde sätt** i cupen'.

Söker man på *örugglega* i korpusar ser man dock snabbt att adverbet ofta förekommer i sammanhang där talaren är tveksam eller osäker. I talspråkskorpusen använder ungdomarna adverbet *örugglega* oftast i svar på frågor. En vanlig kontext är t.ex. när de berättar om sina önskemål och planer för framtiden. I följande utdrag frågar moderatorn eleverna om de tänker ansöka om plats vid en gymnasieskola eller något annat motsvarande (rader 1–2).[1]

1 I transkriptionerna används följande notationer: (0,3) paus angiven i sekunder; (.) mikropaus; [överlappning inleds;] överlappning slutar; ja- avbr; :: förlängd vokal; hh utandning; .hh inandning; ** med skratt i rösten.

1. Ansöka till ett gymnasium: IU
(M=moderator, L=Lárus, S=Siggi, T=Tommi)

01 M ætliði í einhvern- ætliði í menntaskóla
ska ni gå nån, ska ni gå nån gymnasieskola
02 eða eitthvað svoleiðis
eller något sånt
03 L .hnf
04 S uh
05 T mm já
mm ja
→ 06 L **örugglega** bara Emm E: sko
***örugglega** bara ME*
07 M mhm
08 S hhh já hh
hhh ja hh
09 L eða (0,4) Verkmenntaskólann á
eller yrkesskolan i
10 Seyðisfirði
Seyðisfjörður
11 S ókei
okej

Efter tvekan (rad 4) och ett kort positivt svar (rad 5) utvecklar Lárus svaret och förklarar vilken gymnasieskola han vill söka till. Turen inleds med *örugglega* vilket här signalerar att svaret inte är definitivt. Lárus osäkerhet blir ännu synligare i hans följande tur där han tillägger namnet på en annan skola han kanske kan tänka sig söka till (rad 9–10).

Utan *örugglega* i exempel (1) skulle svaret ge ett mera bestämt intryck. Beläggen i korpusen tyder på att i talspråk används adverbet *örugglega* uttryckligen för att indikera att talaren är osäker på sitt svar. Svaren är oftast korta och har formen *örugglega*

+ 'ett möjligt eller sannolikt svar'. För att representera detta i ordboken borde man tillägga ett betydelsemoment som avspeglar detta.

> *ég held, ég geri ráð fyrir* 'jag tror, jag utgår ifrån' Ex. *-hvað verður í hádegismatinn? -**örugglega** fiskur* '-vad får vi till lunch? -fisk **tror jag**', *-ertu að vinna á morgun? -**örugglega*** 'jobbar du imorgon? -jag **utgår ifrån** det'

Osäkerheten som *örugglega* indikerar framgår tydligt när man ser ordets kontext, dvs. man måste ha tillgång till de repliker som kommer före och efter för att kunna tolka funktionen. I en textkorpus som inte visar kontexten kan det vara svårare att sätta fingret på denna användning. Beläggen i gigakorpusen är många, eller drygt 345 000, och man ser inte det som kommer före och efter meningen som innehåller belägget. Genom användningen av en talspråkskorpus får ordboksredaktören tillgång til större kontext och kan lyssna själv på samtalsutdraget. På så sätt är det enklare att tolka ordets egentliga funktion eller innebörd i samtal.

4.3. Diskursstrukturerande funktioner

I talspråkskorpusar kan man hitta funktioner som inte syns i skrivna texter. Detta gäller t.ex. vissa diskursstrukturerande funktioner som bl.a. adverb kan ha. Ordet *allavega* (*alla vega*) och *allavegana* (*alla vegana*) är ett exempel på detta.

Enligt ÍNO har uppslagsordet *allavega* två betydelser (förf. övers.):

> 1) *af mörgum gerðum, fjölbreytilega* 'av olika slag'. Ex. *hann hlustar á **allavega** tónlist* 'han lyssnar på **all slags** musik'.
>
> 2) *að minnsta kosti* 'åtminstone'. Ex. *við getum ekki klifrað*

upp í þetta tré, ***allavega*** *ekki ég* 'vi kan inte klättra upp i detta träd, **åtminstone** inte jag'.

Varianten *allavegana* finns också med i ÍNO men ordboksanvändaren hänvisas direkt till *allavega.* Samma gäller för de särskrivna varianterna *alla vega* och *alla vegana.*

I talspråkskorpusen *Íslenskt unglingamál* förekommer totalt 202 belägg: *allavega* (N=120) och *allavegana* (82). Inget av beläggen används i första betydelsen, dvs. 'av olika slag'. Istället förekommer *allavega* som någon typ av diskursstrukturerande element som används antingen för att precisera eller avgränsa ett tidigare påstående som i exemplet ovan, eller som en återgångsmarkör (jfr Ottesjö 2005 om *iallafall* i svenska).

En återgångsmarkör är en markör som används för att återgå till en berättelse efter ett avbrott, som t.ex. reparationer som har som syfte att utreda oklarheter i det som sagts. Följande utdrag visar ett exempel på detta. I utdraget berättar eleven Embla för de andra hur hennes släkting har byggt robotar som användes i en musikvideo utan att få någon uppmärksamhet för det (rader 1–9). Hon nämner inte namnet på låten men moderatorn kommer med ett förslag (rad 10) som leder till att det blir ett avbrott i Emblas berättelse, dvs. en reparationssekvens som skjuts in. Avbrottet eller reparationen markers med bokstaven A i marginalen (rader 10–12).

2. Robotar för musikvideo: IU
(M=moderator, E=Embla)

```
01  E  eitt tónlistarmyndbandið þar er
       ett av musikvideorna, där finns
02     bjó frændi minn til eitthver
       gjorde min släkting några
03     ótrúlega flotta hvíta: (.)
       otroligt fina vita
```

	04		vélmenni fyrir (.) myndavé- (.)
			robot för kamer- ...
	05		myndir og hérna: .hhh hann fékk
			bilder och ehm han fick
	06		ekkert svona kredit fyrir það
			inget credit för det
	07		það er pínu s-[(.)] pínu
			det är lite s... lite
	08	M	[já]
			ja
	09		leiðinlegt
			tråkigt
A	10	M	love and happiness
A	11	E	já [ábyggilega]
			ja säkert
A	12	E	[held ég] að það heiti
			tror jag att det heter
→	13	E	**allavegana** já hann- hann bjó
			allavegana, ja han- han byggde
	14	E	til þann róbót
			den roboten
	15	M	[já ókei]
			ja okej
	16	E	[og (.)] hann er bara með þá í
			och han har dem i
	17		kjallaranum sínum *heh*
			sin källare heh

Efter att reparationen är slutförd återgår Embla till sin berättelse och upprepar delvis det som hon har tidigare sagt (att det var hennes släkting som byggde roboten). Sedan tillägger hon ny information, dvs. att roboten fortfarande finns bevarad i hans källare (rader 16–17). Återgången till den pågående berättelsen markeras med *allavegana.*

Återgångsmarkören *allavega* förekommer som ett initialt annex, dvs. den är inte syntaktiskt integrerad i den återstående delen av yttrandet. Den används som en turinledning som knyter ihop det som talaren kommer att säga med det som hon har sagt tidigare. Med andra ord har *allavegana* en funktion som inte kan hänföras till ÍNO:s beskrivning av uppslagsordet, dvs. det betyder varken 'av olika slag' eller 'åtminstone'. För att inkludera även denna funktion kunde man tillägga ett tredje moment: 'används för att återknyta till huvudtråden efter avbrott eller utvikning'.

Belägg på *allavega* och *allavegana* som visas i exempel (2) är sällsynta i textkorpusar. Dessutom visas de belägg som eventuellt finns i korpusarna inte i tillräckligt stor kontext för att man ska kunna avgöra vad de har för funktion i sammanhanget.

5. Alternativa metoder för att presentera talspråk

En relevant fråga som kommer upp efter denna genomgång är om en definitionsordbok som ÍNO ska inkludera talat språk. Svaret beror först och främst på vad man har för syfte med ordboken och vilken målgrupp den riktar sig till (se diskussion i Selback & Svardal 2021; Eyþórsson 2005). Som tidigare nämnts är ÍNO en medelstor ordbok som fokuserar mest på standardskriftspråk. Den används i grund- och gymnasieskolor och har en normativ roll, även om där också finns uppslagsord som enligt ordboken inte är helt accepterade och markerade som sådana. Eventuellt kunde man fokusera på alternativa lösningar för att beskriva talspråket på ett grundligare sätt.

Den talspråksbaserade ordboken *Samstalsorðabók* är ett exempel på en alternativ lösning som kan komplettera ÍNO (jfr *Ordbog over Dansk Talesprog*, se t.ex. Hansen 2015). *Samtalsorðabók* fokuserar på interaktiva element som t.ex. pragmatiska markörer, dis-

kurspartiklar, svordomar, hälsnings- och avskedsfraser och tilltal (Hilmisdóttir 2023). Som analysen i avsnitt 4 visar kan markörer i samtal ha en lång räckvidd, dvs. de kan knyta ihop yttranden som inte kommer direkt efter varandra. Enligt traditionen är exempel i vanliga definitionsordböcker korta, oftast en replik eller högst två. I *Samtalsorðabók* används längre utdrag, oftast 6–12 rader. Ordboksanvändaren kan läsa transkriptionen på skärmen, lyssna på hela ljudklippet eller höra själva ordet utan kontext. I ordboken formuleras uppslagsordets funktion och användaren kan slå upp andra ord som har liknande funktioner. Ordboken är deskriptiv och gör ingen skillnad på ord som har funnits länge i språket (t.ex. *ha* 'va') eller nya lån från engelska (t.ex. *whatever*, *basically*). Uppslagsorden kan vara enstaka ord (t.ex. *þúst* 'du vet') eller fraser (t.ex. *allt í lagi*). I ordboken finns även element som inte klassificeras som ord i ordböcker, som t.ex. *ah* och *mhm*.

Figur 1 visar uppslagsordet *ah* som definieras som 'indikerar att talaren har förstått något som han eller hon inte visste tidigare'. Samma exempel används även för att belysa ett av tre betydelsemoment för uppslagsordet *ha* 'va': 'används som reparationsmarkör som visar att talaren inte har hört det som den andra sa och att han måste upprepa sina ord, dvs. 'skulle du kunna upprepa det du sa just nu?'. Inspelningen är från ett samtal mellan två 15-åriga pojkar.

Uppslagsorden i *Samtalsorðabók* begränsas till ord som har pragmatiska och diskursiva funktioner. Ordboken har använts bl.a. i undervisning för andraspråkstalare vid Islands universitet som är en av ordbokens främsta målgrupper.

Vid arbetet med *Samtalsorðabók* blev det dock tydligt att det finns ett behov av större talspråkskorpusar med varierat material. En korpus som *Íslenskt unglingamál* är ganska ensidig eftersom alla inspelningarna består av gruppsamtal som styrs av en moderator. När man samlade in och transkriberade ljudklipp för *Samtalsorðabók* stötte man även på ett problem angående personssekretess. Samtal som spelas in som forskningsmaterial kan

Figur 1: Exempel i *Samtalsorðabók*.

användas som korpusmaterial men ljudfilerna kan inte läggas ut öppet på webben. Vid redigeringen av *Samtalsorðabók* samlade man därför in radiosamtal och poddar som redan var tillgängliga på nätet och tog kontakt med moderatorerna för att få tillstånd att använda ljudklipp i en ordbok.

6. Sammanfattning och avslutande diskussion

I denna artikel har jag diskuterat hur talspråk har inkluderats i moderna isländska ordböcker och hur talspråkskorpusar skulle kunna ge en bättre och mer nyanserad bild av uppslagsordens betydelse eller funktion i talspråk. I artikeln diskuterade jag tre olika fenomen i talspråkskorpusarna: 1) anglicismer, 2) pragmatiska funktioner och 3) diskursstrukturerande element.

Eftersom talspråkskorpusarna tillsvidare är små bygger deras värde först och främst på vad de kan berätta om funktionsord och hur de används i samtal. I ÍNO är funktionsordens semantik ofta underbelyst, eventuellt på grund av att a) den bygger först

och främst på textkorpusar och b) den har som mål att beskriva modern isländska på ett tillgängligt sätt, inte minst med tanke på de yngre generationerna och andraspråkstalare. Ibland fokuseras enbart på uppslagsordens semantiska innebörd (t.ex. *örugglega* 'med säkerhet') även om detta inte nödvändigtvis återspeglar nutida språkbruk. Däremot visar exempelmeningarna ibland belägg som har partikelfunktioner, dvs. de står utanför yttrandets propositionella innehåll (t.ex. *nú* 'nu', *sko* 'ser du'). För att komma åt dessa funktioner måste man ha tillgång till en stor och varierande talspråkskorpus som består av såväl inspelningar som transkriptioner.

För tillfället har ÍNO runt 56 000 uppslagsord vilket betyder att ordboken är betydligt mindre än de äldre verken (156 000 respektive 88 000 uppslagsord). ÍNO är ganska normativ medan *Islandsk-dansk ordbog* från 1920–1924 är utpräglat deskriptiv och *Íslensk orðabók* från 1963 enligt förordet eftersträvar en balansgång mellan att vara normativ och deskriptiv. Syftet med ÍNO är först och främst att fokusera på standardspråk och centralt ordförråd och tanken är att den inte ska innehålla "onödiga" och föråldrade uppslagsord. Eventuellt beror detta på att ÍNO bygger på en lemmalista som har sitt ursprung i ISLEX, som är en isländsk-skandinavisk ordbok avsedd för språkinlärare på Island och i Norden (se Jóhannsson & Úlfarsdóttir, denna volym). I lemmalistan finns varken många lokalt präglade ord och betydningar eller nya lånord. En mycket stor och varierande talspråkskorpus skulle kunna användas för att få fram ny och uppdaterad information om centralt ordförråd i modernt isländskt talspråk, men en sådan korpus återstår att samla in.

Avslutningsvis kan man fundera på om allmänna ordböcker kan göra anspråk på att omfatta såväl tal som skrift. Går det att integrera så olika former i ett och samma uppslagsverk? Talat språk kännetecknas först och främst av att det är kontextbundet och dialogiskt (se t.ex. Linell 1982:5–11). De som ingår i ett samtal befin-

ner sig oftast i samma rum och de kan använda rösten, ansiktsuttryck, kroppsspråk och fysiska föremål för att kommunicera med varandra och försäkra att det inte uppstår några missförstånd. I samtal kan vi därför använda mer deiktiska och allmänna ord än vi gör i skrift, och i samtal kan det uppstå ord som är t.ex. bundna till orter (dialektala ord) eller vissa samhällsgrupper (slang). När vi talar måste vi också förhandla med vår samtalspartner om när och hur länge vi får tala och hur det vi säger anknyter till det pågående samtalet, vilket förklarar det frekventa partikelbruket i talspråk. När vi uttrycker oss i skrift används språket däremot utan yttre kontext. Den som skriver tilltalar läsare som oftast befinner sig i en annan tid och annat rum. Skriftspråket är monologiskt till sin natur och skribenten har ett mindre behov av partiklar och andra diskursstrukturerande element än den som talar. Innehållsorden har däremot en tendens att få ökad tyngd i skrift. Ordförrådet blir större och mera beskrivande och precist och vi behöver inte lika stort sammanhang för att tolka meddelandet. Det ser vi t.ex. i traditionella ordböcker där exempelmeningar ofta består av en kort replik. Som min analys har visat behövs ofta en större kontext i talspråk, ibland flera repliker.

Idealiskt ska en allmän ordbok behandla både tal- och skriftspråk så att en vanlig ordboksanvändare kan hitta svar på ett ställe. Men med tanke på att den lexikografiska traditionen först och främst har grundats på skriftspråk finns det också behov för alternativa ordböcker som fokuserar på talspråk och möjliggör en fördjupad analys på talspråkets egna villkor.

Referenser

Ordböcker, korpusar och digitala resurser

Blöndal, Sigfús (1920–1924): *Íslensk-dönsk orðabók*. Reykjavík: Íslenzk-danskur orðabókarsjóður.

Blöndal, Sigfús (1963): *Islandsk-dansk ordbog. Supplement*. Reykjavík: Íslenzk-danskur orðabókarsjóður.

Íslensk nútímamálsorðabók. Þórdís Úlfarsdóttir & Halldóra Jónsdóttir (red.). Árni Magnússon-institutet för isländska studier. <islenskordabok.arnastofnun.is> (februari 2024).

Íslenzk orðabók handa skólum og almenningi (1963). Árni Böðvarsson (red.). Reykjavík: Bókaútgáfa Menningarsjóðs.

Íslensk orðabók handa skólum og almenningi. Önnur útgáfa, aukin og bætt (1983). Árni Böðvarsson (red.). Reykjavík: Bókaútgáfa Menningarsjóðs.

Íslensk orðabók. Þriðja útgáfa, aukin og bætt (2002). Mörður Árnason (red.). Reykjavík: Edda.

Íslensk orðabók. Fjórða útgáfa byggð á 3. prentun frá 2005 með allnokkrum beytingum (2007). Mörður Árnason (red.). Reykjavík: Edda.

Íslenskt textasafn. Árni Magnússon-institutet för isländska studier. <corpus.arnastofnun.is> (april 2024).

Íslenskt unglingamál. Helga Hilmisdóttir (red.). Tekstlaboratoriet vid Universitetet i Oslo. <tekstlab.uio.no/glossa2/iyl> (februari 2024).

Ordbog over Dansk Talesprog. Carsten Hansen (red.). <odt.hum.ku.dk/om/talesprogtilordbog> (februari 2024).

Risamálheildin. Árni Magnússon-institutet för isländska studier. <malheildir.arnastofnun.is> (februari 2024).

Samtalsorðabók. Helga Hilmisdóttir (red.). Árni Magnússon-institutet för isländska studier. <samtalsordabok.arnastofnun.is> (april 2024).

Annan litteratur

Árnason, Kristján (1997): Hljóðfræðiathuganir Jóns Ófeigssonar. I: *Orð og tunga* 3, 71–78.

Eyþórsson, Aðalsteinn (2005): Hver er kjarni orðaforðans? I: *Orð og tunga* 7, 9–20.

Fjeld, Ruth Vatvedt (2008): Talespråksforskningens betydning for leksikografien. I: Janne Bondi Johannessen & Kristin Hagen (red.): *Språk i Oslo: ny forskning omkring talespråk*. Oslo: Novus. 15–28.

Hansen, Carsten. (2015): Beskrivelsesproget i *Ordbog over Dansk Talesprog*. I: *LexicoNordica* 22, 57–75.

Hilmisdóttir, Helga (2016): *Nú* in Icelandic conversation. I: Peter Auer & Yael Maschler (red.): *NU/NÅ. A family of discourse markers across the languages of Europe and beyond.* Berlin/Boston: de Gruyter. 409–441.

Hilmisdóttir, Helga (2021): Talspråkskorpusar, diskurspartiklar och lexikografi. I: *LexicoNordica* 28, 79–100.

Hilmisdóttir, Helga (2023): Pragmatiska markörer i samtal – en webbaserad ordbok för isländskt talspråk. I: Louise Holmer et al. (red.): *Nordiska studier i lexikografi* 16. Rapport från 16:e konferensen om lexikografi i Norden, Lund 27–29 april 2022. 127–139.

Hilmisdóttir, Helga & Ásta Svavarsdóttir (2023): Icelandic dialect classification. I: *Dialectologia Special issue* 11, 147–175.

Hilmisdóttir, Helga & Elizabeth Peterson (2024): Language contact and language change: Impact on the languages of the Nordic countries. I: Elizabeth Peterson & Kristy Beers Fägersten (red.): *English in the Nordic Countries. Connection, Tensions, and Everyday Realities.* London: Routledge. 84–103.

Ingólfsson, Gunnlaugur (1997): Mállýskuorð. I: *Orð og tunga* 3, 21–24.

Jóhannsson, Ellert & Þórdís Úlfarsdóttir. Ordbog over moderne islandsk – udvikling og tilføjelser. I: *LexicoNordica* 31 (denna volym).

Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir (2011): ISLEX – en flersproget nordisk ordbog: I: Birgit Eaker, Lennart Larsson & Anki Mattisson (red.): *Nordiska studier i lexikografi* 11. Rapport från Konferensen om lexikografi i Norden, Lund 24–27 maj 2011. Lund: Skrifter utgivna av Nordiska föreningen för lexikografi 12. 353–366.

Kristinsson, Ari Páll (2021): Et indblik i islandsk deskriptiv og præskriptiv leksikografi: Importord i to nye digitale islandske ordbøger. I: *LexicoNordica* 28, 121–138.

Kvaran, Guðrún (2007): Undersøgelse af afløsningsord i de nordiske sprog. Indledning. I: Guðrún Kvaran (red.): *Udenlandske eller hjemlige ord? En undersøgelse af sprogene i Norden*. Oslo: Novus. 9–18.

Linell, Per (1982): *The written language bias in linguistics.* Linköping: University of Linköping.

Ottesjö, Cajsa (2005): Iallafall som diskursmarkör. I: Jan Anward & Bengt Nordberg (red.): *Samtal och grammatik – Studier i svenskt samtalsspråk*. Stockholm: Studentlitteratur. 201–229.

Óskarsson, Veturliði (1997): Tæk og miður tæk í Blöndalsorðabók. I: *Orð og tunga* 3, 25–34.

Selback, Bente & Terje Svardal (2021): LIA-korpuset som ressurs i revisjonen av tre ordbøker. I: Kristin Hagen, Gjert Kristoffersen, Øysten A. Vangsnes, Tor A. Åfarli (red.). *Språk i arkiva. Ny forskning om eldre talemål frå LIA-projektet*. Oslo: Novus forlag. 87–107.

Svavarsdóttir, Ásta & Halldóra Jónsdóttir (2020): Kontroversielle ord i purismens land. I: *LexicoNordica* 27, 127–136.

Svensén, Bo (2004): *Handbok i lexikografi: Ordböcker i teori och praktik.* Stockholm: Norsteds akademiska förlag.

Helga Hilmisdóttir
Forskningsdocent, fil.dr.
Árni Magnússon-institutet för isländska studier
Edda, Argrímsgata 5
IS-107 Reykjavík
helga.hilmisdottir@arnastofnun.is

Ordbog over moderne islandsk – udvikling og tilføjelser

Ellert Þór Jóhannsson & Þórdís Úlfarsdóttir

This article examines the lexicographic material found in *Dictionary of Contemporary Icelandic* (OMI). We focus on four processes for acquiring new material: 1) corpus data, 2) selective excerption, 3) user feedback, and 4) editorial additions. The result shows that from the time of inception of OMI in 2013 about 7,600 headwords have been added, of which 6,400 originate in corpus data. Other additions include related usage examples and collocations. The paper highlights the strategies needed to keep the vocabulary of OMI up to date.

1. Indledning

Íslensk nútímamálsorðabók (i dansk oversættelse *Ordbog over moderne islandsk*, herefter OMI) er et leksikografisk projekt ved Árni Magnússon-instituttet for islandske studier, redigeret af Halldóra Jónsdóttir og Þórdís Úlfarsdóttir, der har til formål at beskrive ordforrådet i det nutidige islandske sprog. OMI blev etableret i 2013 og er kun publiceret online. Ordbogen er den eneste nuværende leksikografiske ressource med udgangspunkt i moderne islandsk som bliver regelmæssigt opdateret (se Jónsdóttir & Úlfarsdóttir 2019).

OMI blev etableret ud fra arbejdet med flersprogede ordbøger på Árni Magnússon-instituttet for islandske studier. Ideen var at den nye ordbog skulle genspejle islandsk nutidssprog, især skriftsproget, ganske nøje og inkludere relevante opdateringer, nye ord og andre informationer, dvs. kollokationer og eksempler. Den skulle være brugervenlig med en nemt overskuelig grænseflade og være tilgængelig gratis online.

I forvejen fandtes der en anden islandsk ensproget ordbog, *Íslensk orðabók*, der udkom første gang på tryk i 1963 og i revideret udgave i 1983, 2002 og 2007. Femte og sidste udgave blev publiceret i 2010 og er tilgængelig mod betaling online, men bliver ikke længere opdateret. I dag er de fleste ordbogsbrugere blevet vant til webordbøger, som de også helst vil have gratis adgang til. Udviklingen af OMI var en reaktion på en efterspørgsel i samfundet efter en levende ordbog der beskriver den aktuelle sprogbrug. Samtidig skulle den være nemt tilgængelig og enkel i brug.

I denne artikel vil vi redegøre for materialet som ordbogen bygger på med fokus på dette materiales udvikling, og vi vil diskutere forskellige tilføjelser til ordbogen og motivationen bag dem. Vi vil forsøge at svare på følgende spørgsmål: Hvordan har man besluttet hvilke ord der skal medtages i ordbogen? Hvilke metoder sikrer at ordbogen bedst afspejler det aktuelle ordforråd? Hvordan bliver nye ord en del af ordbogen, og hvilke processer er blevet brugt ved supplering af lemmaer? Artiklen er struktureret på følgende måde: I afsnit 2 giver vi et overblik over baggrund og historik bag OMI hvor vi redegør for OMI's forgænger, ISLEX-projektet, og dets udvikling i perioden frem til 2011. I afsnit 3 og 4 fokuserer vi på forløbet fra OMI's etablering i 2013 og senere tilføjelser til ordbogen. Vi præsenterer her fire processer som vi har defineret som grundlag for tilføjelse af lemmaer. I afsnit 5 er der opsummerende bemærkninger.

2. Baggrund og historik

OMI's oprindelse går tilbage til ISLEX, et større leksikografisk projekt der går ud på at lave en flersproget ordbog med islandsk som kildesprog og andre sprog i Norden som målsprog (jf. Jónsdóttir & Úlfarsdóttir 2011, Úlfarsdóttir 2013, 2014). I sin nuværende form har ISLEX ækvivalenter på dansk, svensk, bokmål, nynorsk, fær-

øsk og finsk. ISLEX blev etableret i 2004, og første version blev publiceret online i 2011.

2.1. Nordisk ordbogsbase og starten af ISLEX

For at kunne redegøre for ordforrådet i OMI er det nødvendigt at se nærmere på ISLEX og det materiale som dette projekt har som fundament. ISLEX bygger på en lemmaliste der kaldes *Norrænn orðabókarstofn* (Nordisk ordbogsbase, dvs. basisgrundlag for nordiske ordbøger) som blev udarbejdet lidt før år 2000 og er baseret på forskellige tekstkilder. Arbejdet startede som et islandsk-norsk initiativ om at lave en liste over ord som kunne danne grundlag for en ny ordbog der skulle dække det almensproglige ordforråd (jf. Bjarnadóttir 1998). Man brugte materiale fra et udgivet ordbogsværk (Jónsson 1994) samt supplerende materiale efter behov fra bl.a. det store citatarkiv *ROH* og andre tekstsamlinger (jf. Bjarnadóttir 1998). Arbejdet resulterede i en bruttoliste på 200.000 lemmaer, kategoriseret efter formelle og semantiske kriterier, og rangeret efter lemmapotentiale. Efter manuel gennemgang og kritisk sortering blev ca. 46.000 ord valgt som materiale til ISLEX.

Oprindelig indlæsning af lemmaer i 2004 talte 46.440. I 2004-2005 havde man forberedt materialet til videre behandling. Under det tidlige redaktionelle arbejde blev lemmalisten revideret, og mange lemmaer og opslagsformer blev slettet. Samtidig startede man gradvis med at tilføje ord som man mente manglede.

ISLEX blev gennemført i årene 2006-2011 hvor redaktørerne for alle målsprog arbejdede med den samme islandske lemmaliste. I første omgang inkluderede ISLEX målsprogene dansk, nynorsk, bokmål og svensk. Ordbogen blev publiceret online i 2011. Efterfølgende er lemmalisten blevet udvidet med flere ord, og målsprogene færøsk og finsk er blevet tilføjet.

2.2. Etablering af ISLEX og senere tilføjelser

Nordisk ordbogsbase var for det meste kun en liste over lemmaer, men den indeholdt dog også en grov og ufyldestgørende semantisk klassificering (191 kategorier for omkring 15 % af ordene). For at lave en ordbog ud af basen var det nødvendigt at udvikle en klar mikrostruktur og tilføje supplerende oplysninger. Den islandske redaktion sørgede for at lemmaerne blev delt op i betydninger hvor der var grund til det, og forklaringer på islandsk blev tilføjet ved hvert eneste opslag. Det skulle være til nytte ved det efterfølgende redaktionelle arbejde med målsprogene (se yderligere i Jónsdóttir & Úlfarsdóttir 2011). Disse forklaringer dannede også senere grundlag for orddefinitioner i OMI.

Yderligere leksikografiske oplysninger blev tilføjet under det redaktionelle arbejde, og ordbogsartiklerne blev udvidet med brugseksempler og forskellige kollokationer og fraser. Til det sidstnævnte har man især brugt Jón Hilmar Jónssons ordbogsværker fra 1994 og 2002.

Man opdagede tidligt at der stadigvæk manglede ord i den oprindelige indlæsning fra den nordiske ordbogsbase, især inden for nogle specifikke semantiske felter, fx samfundsrelaterede ord. Det resulterede i målrettede tilføjelser af ord eller typer af ord. I det følgende gives nogle eksempler på ord der blev tilføjet under det redaktionelle arbejde i årene 2006-2007: *hælisleitandi* 'asylsøger', *móðurmálskennsla* 'modersmålsundervisning', *ráðningarsamningur* 'ansættelseskontrakt', *upprunaland* 'oprindelsesland' og *vegabréfaskoðun* 'paskontrol'. Kilderne for disse ord inkluderede websider fra multikulturelle foreninger og samfundsorganisationer rundt omkring i Island.

Der var også mangel på lande- og indbyggernavne. For at bøde på det brugte man en liste der blev kompileret af det islandske sprogcentrum *Íslensk málstöð*, som senere blev en del af Árni Magnússon-instituttet. I tabel 1 vises eksempler på den slags ord.

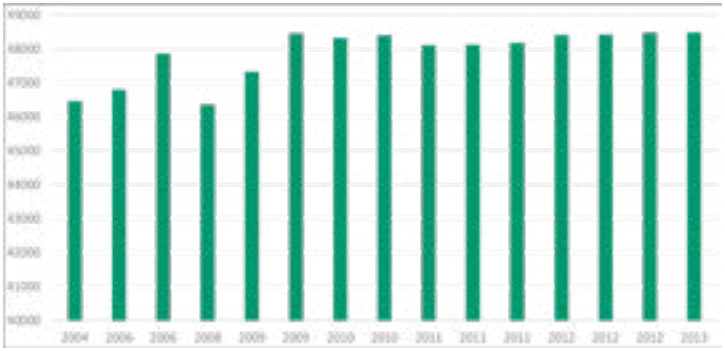
Land	**Adjektiv**	**Indbygger**
Alsír	alsírskur	Alsíringur
Argentína	argentínskur	Argentínumaður
Ástralía	ástralskur	Ástrali
Belgía	belgískur	Belgi
Brasilía	brasilískur	Brasilíumaður
Danmörk	danskur	Dani

Tabel 1: Udsnit af liste over lande med tilhørende adjektiver og indbyggernavne i ISLEX.

En liste som den man kan se i tabel 1, kan være et meget nyttigt værktøj idet det giver mulighed for at tilføje manglende oplysninger inden for et begrænset domæne.

Da man begyndte at arbejde med islandske sprogkorpusser, fik man et nyt værktøj for at finde frem til manglende ord, og det var klart at en del ganske almindelige ord ikke var inkluderet i den nordiske ordbogsbase og derfor også manglede i en fyldestgørende leksikografisk beskrivelse af sproget. Et eksempel er adverbier med endelsen *-lega* (svarer til da. *-lig*). Dette er en meget aktiv orddannelsesmetode i islandsk. Ordtypen optræder i mange tekstgenrer og forekommer derfor også tit i korpusdata. En række af denne type ord blev ved en målrettet indsats tilføjet til ISLEX, fx *afbrigðilega* 'unormalt' el. 'usædvanlig(t)', *andstyggilega* 'frygtelig(t)', 'forfærdelig(t)', *ársfjórðungslega* 'kvartalsvis(t)' og *feimnislega* 'genert'.

Alle disse tilføjelser som er diskuteret her, dvs. ordforråd som vedrører samfundet, indbyggernavne og adverbier, blev i lighed med mange andre tilføjet til ISLEX og blev derved en del af det oprindelige ordstof i OMI nogle år senere. Udviklingen af ordforrådet i ISLEX vises i figur 1:



Figur 1: Grafik der viser udviklingen af antal lemmaer i ISLEX. Der er generelt 1-3 datapunkter for hvert år.

Som man kan se af figur 1, blev der tilføjet mange ord, men der blev også foretaget en del sletninger. Kurven bevæger sig ikke lineært opad, men går tilbage de gange redaktørerne har ryddet op i lemmalisten.

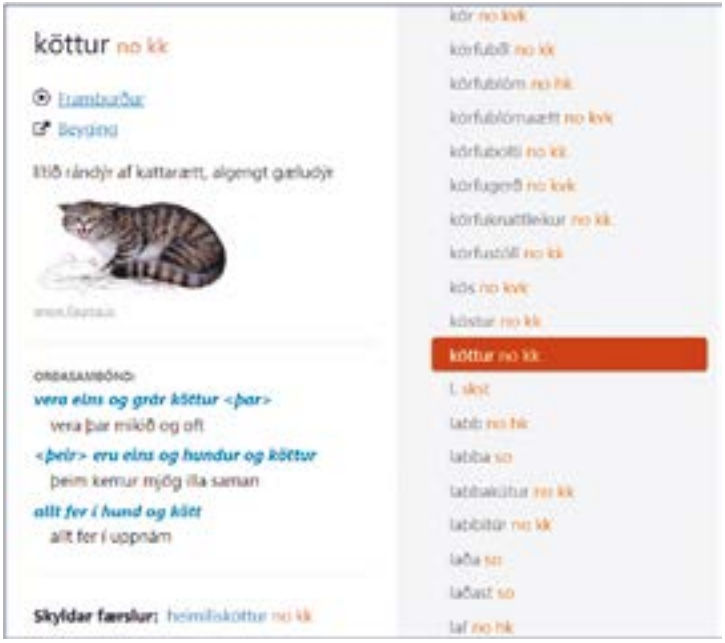
2.3. Sletninger fra ISLEX

Mellem 2006 og 2008 blev der slettet en del ord fra ISLEX. Da man var gået i gang med det redaktionelle arbejde, konstaterede man nemlig at den oprindelige ordbogsbase indeholdt en del ord som var forældede og/eller meget sjældne. Lemmalisten blev gennemlæst og revideret. Ordlister over mulige sletningskandidater blev genereret og efterfølgende diskuteret til redaktionsmøder. Mange af de slettede ord var sammensatte ord som man syntes var unødvendige at medtage som selvstændige lemmaer i en mellemstor ordbog. De inkluderede ord med forstærkende forled: *aðalhvatamaður* 'hovedinitiativtager', *hávísindalegur* 'højvidenskabelig', *fjallbrattur* 'stejl som et bjerg', men også andre sammensætninger, fx *kennslusjónvarp* 'undervisningsfjernsyn' og *prjónapeysa* 'strikkesweater', som redaktionen vurderede ikke hørte til

i en ordbog af denne type, da de enten er gennemskuelige eller sjældne. Tilsvarende nogle eksempler på verber der blev slettet: *þurrsalta* 'tørsalte', *þvermóast* 'nægte at gøre', *væma* 'få kvalme', *útdjöfla* 'udskælde' og *utanbókarlæra* 'udenadslære'. Der var kun få belæg på disse ord, og de er heller ikke en del af det almensproglige ordforråd.

3. OMI: En selvstændig ensproget ordbog

I afsnit 2 har vi diskuteret ISLEX og hvordan dette projekts materiale er blevet bearbejdet igennem årene. I forbindelse med OMI er ISLEX relevant indtil året 2013 hvor OMI blev etableret som en ny selvstændig ordbog. Første version var blot en kopi af ISLEX hvor man havde fjernet alle målsprogene (jf. Jónsdóttir & Úlfars-



Figur 2: Et typisk opslag i OMI, *köttur* 'kat'.

dóttir 2020). Dengang i 2013 var ISLEX kommet op på 48.500 ord, så det var det ordforråd man brugte i den første version af OMI. Efter den tid har disse to ordbøger kørt på hvert sit spor selvom de stadigvæk har tætte forbindelser, med fælles redaktører og tit identiske opdateringer. OMI er nu den eneste islandske ensprogede ordbog der bliver regelmæssigt opdateret, og den betragtes nu som standard referencekilde over nutidens islandske sprog.

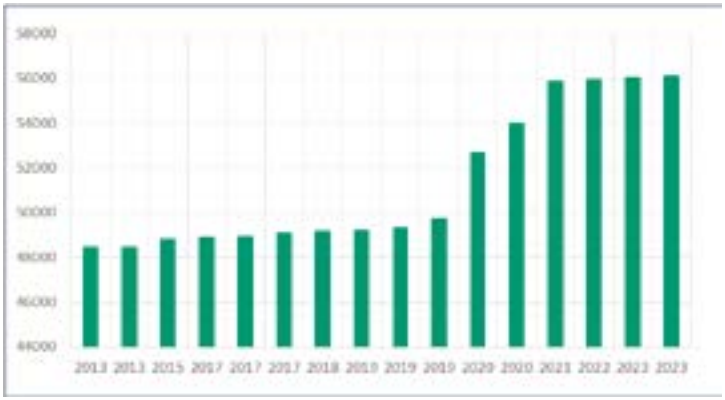
OMI er "født" med lemmastrukturen fra ISLEX. I figur 2 vises der et typisk opslag fra OMI. Figuren illustrerer hvordan de leksikografiske oplysninger er gengivet i ordbogen. Til venstre har man opslagsordet sammen med ordklasse. Dernæst har man en klikbar knap "Framburður" ('udtale') hvor man kan høre en lydgengivelse af det relevante ord. Man har også et link til en anden webressource, BÍN, hvor man kan se alle bøjningsformer (16 for substantiver). Nogle artikler har et billede ligesom den man kan se i figur 2. Man har også en del kollokationer og fraser med tilhørende forklaringer. Nederst er der en automatisk genereret liste over beslægtede ord, her bl.a. *heimilisköttur* 'huskat'. Til højre i artiklen har man et ordhjul med lemmaer i alfabetisk rækkefølge.

4. Processer ved tilføjelser i OMI

En vigtig del af det redaktionelle arbejde i en ordbog over samtidens sprog er at den konstant skal opdateres med nye ord og andre leksikografiske oplysninger. Motivationen for at tilføje nye lemmaer til OMI er at opretholde ordbogen som afbalanceret og repræsentativ for islandsk nutidssprog.

I figur 3 kan man se hvordan antallet af lemmaer har ændret sig. Datapunkterne er lidt tilfældige fordi ordbogen bliver opdateret et antal gange om året, men ikke på bestemte datoer. Når redaktørerne skønner at de har forberedt en del lemmaer til publicering, bliver ordbogen opdateret med de nye ord. På den måde

er ordbogen vokset gradvis med ca. 7.600 nye ord siden den blev etableret.



Figur 3: Overblik over udviklingen af antal lemmaer i OMI.

Figur 3 viser udviklingen fra 2013. Man kan se at der sker en gradvis stigning i antallet af ord frem til 2019. Mellem 2019 og 2021 sker der en markant stigning, og efter 2021 bliver stigningen igen mere dæmpet.

Vi var interesseret i at se nærmere på disse nye lemmaer, og hvordan man har fundet frem til dem. Hvis man ser på ordbogsdataene i historisk blik og undersøger de tilføjelser der er kommet til, kan man inddele de nye lemmaer i nogle overordnede kategorier efter hvordan man har fundet frem til dem. Efter at vi havde studeret materialet, kunne vi identificere fire processer som bidrager til tilføjelser af nye ord i OMI: 1) nye korpusdata, 2) målrettet excerpering, 3) brugerbidrag, 4) redaktionelle forbedringer. I de følgende afsnit vil vi redegøre for disse fire processer og den påvirkning de har haft på indholdet i OMI.

4.1. Nye korpusdata

Integreringen af nye korpusdata er afgørende for at holde OMI op-

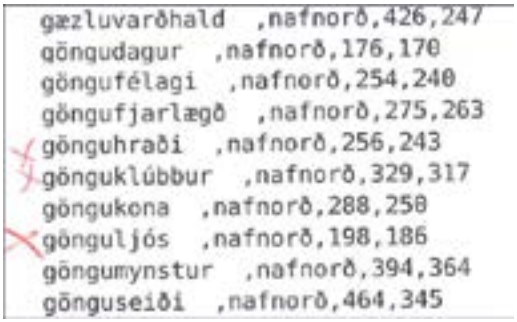
dateret og for at bevare ordbogens aktualitet i forhold til hvordan sproget bliver brugt. Indførelsen af friske sprogdata hjælper med til at sikre ordbogens relevans.

OMI bliver opdateret med nye ord nogle gange om året, fra forskellige kilder. I 2019-2021 blev der foretaget et særligt initiativ for at udvide OMI's lemmaliste ud fra korpusdata. Ord blev hentet fra *Risamálheild* [Gigaword Corpus] (Rmh), et nyere tekstkorpus (se Steingrímsson et al. 2018 og Barkarson et al. 2022). Dette korpus er udviklet hos Árni Magnússon-instituttet for islandske studier, ligesom OMI og ISLEX. En frekvensbaseret liste blev brugt, hvor de tre hovedordklasser, substantiver, adjektiver og verber, blev behandlet hver for sig. Listerne blev sammenlignet med OMI's lemmaliste, hvor de ord der allerede var i ordbogen, blev sorteret fra.

Dataene fra Rmh er dog ikke uden problemer. Der er en vis ubalance i korpusdataene, idet visse genrer fylder meget i tekstkilderne (jf. diskussion hos Sigurðsson & Steingrímsson 2024). En forholdsvis stor andel administrative tekster og reklameagtigt materiale kan give skævhed i frekvenslisterne som man skal være opmærksom på når korpusmateriale skal indgå i ordbogsarbejde. Ud over det er ordene i nogle tilfælde i en anden grundform end i OMI og bliver derfor ikke fanget af sammenligningsprocessen, fx Rmh *perluvinur* 'perleven', OMI *perluvinir* 'perlevenner'. Almindelige tastefejl eller stavefejl i Rmh kan også forekomme, fx *auðævi* i stedet for *auðæfi* 'rigdom'. I andre tilfælde kan der opstå variantformer fordi retskrivningsregler er blevet revideret, fx *verzlun* ved siden af *verslun* 'handel'. Ud over sådanne tilfælde er der også en del "støj" i korpusdata som skal sorteres fra, fx på grund af fejl i lemmatisering. Derfor er det klart at korpusdata skal bruges med kritisk omtanke i leksikografisk sammenhæng.

Der var imidlertid ikke en åbenlys og etableret procedure for hvordan ord fra Rmh bedst kunne integreres i OMI. Resultatet blev en excerperingsprocedure som inkluderede en blanding af maskinel og menneskelig indsats. I første omgang blev der genere-

ret frekvenslister. Disse lister blev udskrevet og siden gennemgået manuelt af ordbogsredaktørerne. Ved gennemlæsning blev ordene vurderet ud fra subjektive kriterier, med udgangspunkt i redaktørernes sprogfornemmelse og om ordet hører til i en almensproglig ordbog. Her kigger redaktørerne ikke kun på frekvens, men også på andre faktorer, dvs. hvor bekendt ordet er, og i hvilken slags tekster det forekommer. Godkendte ord blev markeret med rødt x (jf. figur 4).



Figur 4: Et udvalg fra en udskrevet liste med potentielle lemmaer genereret fra korpusdata.

Disse ord blev derefter tilføjet til OMI's lemmaliste og behandlet i stil med andre lemmaer i selve ordbogen.

Det er også interessant at se nærmere på tilføjelserne til OMI's lemmaliste i lyset af semantiske felter. OMI bruger et semantisk klassificeringssystem hvor alle ord er markeret som hørende til én eller flere af over 1.500 betydningskategorier. Dette system var oprindelig udviklet til ISLEX, men blev også en del af OMI.

OMI		
Semantisk felt	**Tilføjelser**	**Antal ord i alt**
mad	171	1217
medicin	113	1168
rejse	124	531

skole/uddannelse	154	584
samfund	68	467
IT	75	428
musik	177	538
administration	138	279

Tabel 2: Antal lemmaer i OMI inden for otte udvalgte semantiske felter.

I tabel 2 vises kun et udvalg af tilføjelser til nogle af de større semantiske felter. Når ordene blev tilføjet, var det dog ikke ud fra kriterier baseret på semantiske oplysninger, men samtidig med at ordene bliver en del af OMI, bliver de markeret som hørende til en eller flere semantiske felter. Man kan se at nogle af kategorierne voksede med op til 25 % ved tilføjelsen af korpusdata. Nogle ord falder ind under mere end én kategori, fx hører *heimsbikarmót* 'VM-turnering' til både <sport> og <samling>, og *samnemandi* 'medstuderende' hører til både <skole> og <sociale relationer>. I alt blev ca. 6.400 ord tilføjet fra korpusdata.

4.2. Målrettet excerpering

Selvom de mest markante tilføjelser til OMI stammer fra Rmh, er der en del ord der er blevet tilføjet efter andre metoder, både før og efter korpusindsatsen 2019-2021. På grund af overvægt af visse genrer i korpusmaterialet fanger den mere automatiske proces ikke alle de ord man skønner hører til i et ordbogsværk som OMI. I afsnit 2.2. har vi vist nogle eksempler på målrettet excerpering i forbindelse med supplering af materiale til ISLEX. Man har også udført lignende tilføjelser i forbindelse med OMI for at udfylde huller i lemmalisten. I figur 5 ses et eksempel på hvordan ord er blevet excerperet fra en brochure fra det islandske rigshospital vedrørende hjerteproblemer og stødbehandling (*Rafvending* 2021).


LANDSPÍTALI

FRÆÐSLUEFNI

Rafvending

Rafvending er gerð hjá einstaklingum með gáttatif eða gáttaflökt til að koma hjartslætti þeirra í reglulegan takt aftur. Einkenni um óreglulegan hjartslátt geta verið hraður og óreglulegur púls, svimi, minna úthald, máttleysi, mæði og brjóstverkur. Hjartsláttaróregla getur komið barði í hvíld og við áreynslu.

Ef annan við einn til tvær sólarhringar eru síðan hjartsláttaróregla byrjaði er hægt að rafvenda.

Blóðþynningarmeðferð

Ef hjartsláttaróregla hefur varað lengur en tvo sólarhringa er gefið blóðþynningar lyf í 3-4 vikur áður en rafvending er gerð til að minnka hættu á myndun blóðtappa. Til eru mismunandi tegundir blóðþynningalyfja en læknir tekur ákvörðun um meðferð og veitir upplýsingar um notkun lyfsins.

Figur 5: En excerperet side fra en brochure om hjerteproblemer.

Som man kan se på figur 5, har man ved en manuel gennemgang af materialet markeret ord som kunne være kandidater til OMI. Ordene er derefter tjekket mod lemmalisten i OMI og tilføjet hvis de manglede. Ved gennemgangen af denne bestemte udgivelse er der tilføjet en række ord i OMI, fx *rafvending* 'stødbehandling', *gáttatif* 'hjerteflimmer', *gáttaflökt* 'hjerteflagren', *hjartsláttaróregla* 'hjertearytmi' og *blóðþynningarmeðferð* 'blodfortyndingsbehandling'. Ordene overstreget med lilla, *brjóstverkur* 'brystsmerte' og *blóðþynningarlyf* 'blodfortyndende medicin', viste sig allerede at være i ordbogen.

En lignende excerpering fra en brochure om vulkansk luftforurening resulterede også i en del nye ord. Efter at en række jordskælv og vulkanudbrud begyndte på Reykjanes-halvøen i marts 2021, dukkede ordforråd relateret til dette op i medier og samfundsdiskussioner. For det meste var ordene allerede i ord-

bogen, men der var grund til at tilføje flere ord. Derfor blev der samlet et tekstmateriale hvor mange af de relevante ord forekom. Et eksempel er en brochure med titlen *Hætta á heilsutjóni vegna loftmengunar frá eldgosum* 'Fare for helbredet på grund af luftforurening fra vulkanudbrud' der blev udgivet i fællesskab af seks institutioner, bl.a. det islandske meteorologiske institut og sundhedsministeriet (*Hætta* … 2022). Brochuren indeholder fx ord som *brennisteinsdíoxíð* 'svovldioxid', *loftgæði* 'luftkvalitet', *bráðarugl* 'delirium' og *skyndidauði* 'akut død'.

Målrettet excerpering afspejler en bevidsthed om behovet for at finde frem til forskellige typer ord som måske ikke er repræsenteret i de tilgængelige korpusdata. Eventuelle skævheder i korpusdata har betydning for spørgsmålet om lemmasammensætningen, og det er nødvendigt at gøre en supplerende indsats for at sikre at terminologi og ordforråd inden for specifikke, vigtige emner er fyldestgørende repræsenteret i ordbogen. Her er det dog nødvendigt at vælge omhyggeligt hvilke termer der er relevante for almindelige brugere og ikke medtage ord der kan være for specialiserede, og som kun bruges i meget begrænset omfang.

4.3. Brugerbidrag

Ordbogsbrugere påpeger af og til at visse ord mangler i en ordbog. I nogle tilfælde resulterer dette i at der tilføjes ord. For eksempel har oversættere mellem islandsk og dansk fremsat mange forslag til lemmaer som de ønskede blev tilføjet til ISLEX, og alle disse tilføjelser blev også inkluderet i OMI. Eksempler på disse ord er *gagnaver* 'datacenter', *fjarfundur* 'videokonference' og *geðrof* 'psykose'. Ikke mange dinosaurer har fået et islandsk navn, men arterne *snareðla* 'velociraptor' og *þórseðla* 'brontosaurus' er undtagelser og er blevet optaget i OMI efter brugerhenvendelser. Der har også været andre forslag til lemmaer, af meget forskellig art, som fx *hipsumhaps* 'hip som hap', *heilsuójöfnuður* 'helbredsulighed' og *menningarnám* 'kulturel appropriation'.

Brugerkommentarer retter sig tit mod betydninger eller definitioner, frem for mod tilføjelser af ord. Et eksempel er en brugerhenvendelse vedrørende lemmaet *trilljón*: "der mangler betydningen 1.000.000.000.000.000.000". Dette har resulteret i at man har tilføjet betydningen 'tallet ét med 18 nuller' til den oprindelige ordbogsartikel som kun havde betydningen 'meget stor mængde'.

Brugerne bidrager ofte med tilføjelser og forslag til ordbogen, og inkludering af dem bringer ordbogen tættere på de personer der anvender materialet. En kritisk gennemgang og evt. ændringer på grund af brugerhenvendelser sikrer at ordbogen er inkluderende og repræsenterer det aktuelle sprog der anvendes af forskellige brugergrupper.

4.4. Redaktionelle tilføjelser

Nogle gange hvor redaktørerne tilføjer et enkelt ord, muligvis efter excerpering, bliver der efterfølgende tilføjet flere ord af samme type, fx verber på *-era* (tit låneord fra dansk), fx *flambera*, *formúlera*, *gratínera*, *impónera* og *stílísera*. Andre eksempler er substantiver på *-sjón* (låneord fra dansk/engelsk), fx *tradisjón*, *sessjón*, *frústrasjón* og *pródúksjón*, og IT-relaterede ord, fx *snjallúr* 'smartur', *skalanlegur* 'skalerbar', *vélþýðing* 'maskinoversættelse' og *tveggja þátta auðkenning* 'tofaktorgodkendelse'. Et enkelt ord eller et bestemt semantisk domæne kan derfor resultere i mange flere ord hvis det sætter gang i en frugtbar tankeproces hos redaktørerne.

Tit opstår der nye ord ved arbejdet på tosprogede ordbøger. Når redaktører på ISLEX eller andre tosprogede ordbogsprojekter (fx LEXÍA) støder på ord som de synes savner en islandsk ækvivalent, bliver den ofte identificeret og tilføjet. Det gælder især mange neologismer som måske ikke har en lang historie i sproget og ikke er blevet en del af det tidligere ordbogsmateriale, fx *skortsala* 'blankosalg', *auðsöfnun* 'kapitalakkumulation', *hámhorf* 'bingewatching', *spjallmenni* 'chatbot', *ljóslestur* 'OCR-behandling', *ofmettaður* 'overmættet' og *kynhlutlaus* 'kønsneutral'.

Disse eksempler på redaktionelle forbedringer illustrerer at det kan være vigtigt at leksikografer tilpasser de eksisterende data. Denne form for redaktionelt arbejde mindsker risikoen for at ordbogen muligvis overser nogle relevante ord, og arbejdet spiller en afgørende rolle for at nydannede ord hurtigt bliver optaget i ordbogen.

4.5. Andre tilføjelser

Ud over tilføjelser af enkelte opslagsord er ordbogens indhold blevet forøget på andet sæt. En opslagskategori som er blevet udvidet, er flerordslemmaer, dvs. et lemma der inkluderer faste fraser, dvs. mere end ét ord som allerede findes i ordbogen. I en række tilfælde danner den slags fraser en selvstændig betydningsenhed som formelt ikke kan ændres. I tabel 3 kan vi se eksempler på adverbielle flerordslemmaer som er blevet tilføjet fra 2013.

Flerordslemma	**Dansk oversættelse**
ef að líkum lætur	'sandsynligvis'
eða þar um bil	'eller deromkring'
ef á þarf að halda	'om nødvendigt'
eins og stendur	'for tiden'
ekki alls kostar	'ikke helt'
fyrir alla muni	'for enhver pris'
fyrir langa löngu	'for længe siden'

Tabel 3: Eksempler på flerordslemmaer tilføjet fra 2013.

De fleste flerordslemmaer er adverbielle fraser der har status som adverbier i sproget. De forekommer tit i tekster, men kan være svære at finde frem til i ordbøger fordi det er uklart hvilket ord man skal slå op under (se fx Jónsson 2005:25-26, 2015). I OMI behandles flerordslemmaer på samme måde som étordslemmaer

og kan søges både som tekststreng i sin helhed eller som individuelle dele. Hvis søgningen giver flere resultater, får man en liste af mulige kandidater og kan klikke på det relevante lemma. Flerordslemmaer er en væsentlig del af sproget, og det er vigtigt at man nemt kan finde frem til dem i leksikografiske ressourcer som OMI.

Derudover er der tilføjet materiale der ikke resulterer i flere lemmaer, men som figurerer i andre dele af artiklerne i ordbogen, fx fraser, kollokationer og brugseksempler. Når et nyt ord bliver etableret, tilføjes der nogle gange også fraser eller kollokationer til det nye opslag. Tit bliver også brugseksempler tilføjet til nyopprettede ordbogsartikler.

4.6. Sammenfatning om ændringer i antal opslagsord

Vores undersøgelse har vist at antallet af lemmaer er vokset med 7.600 fra OMI's etablering i 2013, fra omkring 48.500 til omkring 56.100, dvs. ca. 15 %. Hvis man inkluderer ISLEX-delen, viser det sig at der i alt er blevet tilføjet omkring 17.000 ord siden den oprindelige nordiske ordbogsbase blev konstrueret i slutningen af halvfemserne. I samme periode er omkring 7.000 ord blevet slettet.

Når man ser nærmere på perioden fra 2013 og disse 7.600 nye ord, fremgår det at 6.400 ord stammer fra korpusdata, især gennem den særlige indsats der fandt sted i årene 2019-2021. Omkring 1.200 ord stammer fra andre kilder som vi har identificeret, dvs. målrettet excerpering, brugerbidrag og redaktionelle tilføjelser. Det er ikke altid nemt at skelne imellem disse underkategorier fordi ordene ikke bliver konsekvent markeret i OMI's database efter hvilken metode de er blevet tilføjet. Nogle gange er det dog klart, og vi har givet nogle eksempler på forskellige typer tilføjelser.

En prominent kategori som hører under redaktionelle forbedringer, er flerordslemmaer. Antallet af dem er steget med en tred-

jedel fra ca. 600 til nuværende 900 i alt. Vi kan også se fra databasen at der er ca. 915 nye fraser og kollokationer der er kommet til i de seneste ti år, og de er nu omkring 10.000 i alt, hvilket nogenlunde er det som man forholdsmæssigt kunne forvente hvis man sammenligner med stigningen i antal lemmaer. Brugseksempler i ordbogsartiklerne som primært kommer fra korpusdata, er også steget omkring 15 %, hvilket er proportionelt i forhold til stigningen i antal lemmaer, nemlig fra omkring 29.500 til 34.000 i alt.

5. Opsummerende bemærkninger

I denne artikel har vi diskuteret ordforrådet i OMI og vist hvordan ordbogen er blevet til, og hvordan det leksikografiske materiale er blevet udvidet og forbedret i løbet af de 10-11 år der er gået siden ordbogen blev etableret. Vi har illustreret hvordan OMI (og dens slægtning ISLEX) er blevet udviklet igennem årene med forskellige tilføjelser og ændringer. Vi kan opsummere hvordan vi har svaret spørgsmålene i indledningen, om hvordan man har besluttet hvilke ord hører til i ordbogen, hvilke metoder der sikrer at ordbogen bedst afspejler det aktuelle ordforråd, hvordan nye ord bliver en del af ordbogen, og hvilke processer der er blevet brugt ved supplering af lemmaer.

Ideen med OMI var at den skulle være en mellemstor ordbog der fanger ordforrådet i samtidssproget. Størstedelen af lemmaerne i OMI er arvet fra ISLEX hvor de går tilbage til Nordisk ordbogsbase som blev kompileret i slutningen af sidste århundrede med det formål at belyse moderne islandsk. Samtidig har det været nødvendigt at ordbogen fortsat kan afspejle ord der bruges i samfundet og deres betydninger. Lemmalisten skal derfor revideres og nye ord tilføjes i takt med at sprogbrugernes behov ændres. Det kræver at ordbogens redaktion er opmærksom på de faktorer der

kan være relevant for ordforrådet og er klar til at reagere med revideringer og suppleringer af ordbogsmaterialet.

Når vi ser på de metoder der er blevet anvendt, fremgår det klart at den menneskelige indsats spiller den største rolle for at sikre ordbogens relevans, hvor redaktørerne vurderer ordene ud fra deres formelle karakteristika og skønnet frekvens. Mange beslutninger når det gælder nye ord, er baseret på subjektive indtryk og diskussioner af individuelle ord eller grupper af ord ved redaktionelle møder.

Korpusdata er vigtigt og kan bidrage med afgørende oplysninger om nye ord og betydninger. I lyset af semantiske oplysninger kan man se at mange af de nye ord hører til relativt få semantiske felter. Det kunne indikere at visse semantiske kategorier er mere produktive end andre og derfor burde få større opmærksomhed af OMI's redaktører. Forskellen i det semantiske indhold af de tilføjede ord kan også tyde på at der muligvis er grund til at gå målrettet efter suppleringsmateriale til underrepræsenterede semantiske kategorier.

Vi har identificeret fire processer der er blevet brugt ved supplering af lemmaer og redegjort for de metoder man har brugt i forbindelse med dem for at sikre at ordbogen bedst kan afspejle det aktuelle ordforråd. Af disse processer er brug af korpusdata det som er mest afgørende. Andre processer er dog også vigtige for at modvirke skævheder i korpusdata og til at tilpasse materialet til samfundsmæssig udvikling og for at reagere på brugerrelaterede behov. Vi har forklaret motivationen bag tilføjelserne og diskuteret relevante problemstillinger som er i spil når supplerende ordforråd vurderes og evt. nye ord bliver en del af ordbogen.

OMI stræber efter at registrere og belyse det nutidige islandske ordforråd. Ordbogen har etableret sig som et værdifuldt referenceværktøj for alle der er interesseret i det islandske sprog, og den er en meget populær ressource. Antallet af brugere har øgedes markant år fra år, og tallene fra en seks måneders periode fra

august 2023 til februar 2024 viser at der er over 234.000 unikke brugere som har besøgt ordbogen over 670.000 gange. Vedvarende arbejde med at forbedre og udvikle ordbogen så den bedst repræsenterer sproget, fortsætter, og redaktørerne udvikler processer for at holde ordbogen relevant med opdateringer og tilføjelser af nyt materiale. I den nærmeste fremtid har man planer om at bruge nyudviklede sprogteknologiske værktøjer som en hjælp til automatisk excerpering af digitalt materiale. Man er også i gang med at lave et korpus over udgivne bøger fra det 20. århundrede som repræsenterer tekstgenrer der ikke er så prominente i de nuværende korpusressourcer. Begge disse initiativer kommer sikkert til at resultere i flere tilføjelser til OMI de næste måneder eller år.

OMI er en forholdsvis ung ordbog. Den har dog på kort tid etableret sig som en vigtig islandsk sprogresurse. Med konstant revidering og udvikling af nye metoder til at fange ordforrådet i det levende sprog kan man sikre at ordbogen bliver relevant mange år fremover.

Litteratur

Ordbøger, korpusser og digitale resurser

BÍN = *Beygingarlýsing íslensks nútímamáls*. Kristín Bjarnadóttir (red.). Árni Magnússon-instituttet for islandske studier. <bin.arnastofnun.is> (marts 2024).

ISLEX. Þórdís Úlfarsdóttir (hovedred.). Árni Magnússon-instituttet for islandske studier. <islex.is>, <islex.dk> (marts 2024).

Íslensk orðabók (2010). Mörður Árnason & Laufey Leifsdóttir (red.). 5. útgáfa. [1. udg (1963), 2. udg. (1983), 3. udg. (2002), 4. udg. (2007)]. Reykjavík: Forlagið.

Jónsson, Jón Hilmar (1994): *Orðastaður*. Reykjavík: Mál og menning.

Jónsson, Jón Hilmar (2002): *Orðaheimur*. Reykjavík: JPV-útgáfa.

LEXÍA. Þórdís Úlfarsdóttir (hovedred.), Rósa Elín Davíðsdóttir (red. fransk). Árni Magnússon-instituttet for islandske studier. <lexia.hi.is> (marts 2024).

OMI = *Íslensk nútímamálsorðabók* [Ordbog over moderne islandsk]. Þórdís Úlfarsdóttir & Halldóra Jónsdóttir (red.). Árni Magnússon-instituttet for islandske studier. <islenskordabok.is> (marts 2024).

Rmh. = *Risamálheildin* [Icelandic Gigaword corpus]. Stofnun Árna Magnússonar í íslenskum fræðum. <malheildir.arnastofnun.is> (februar 2024).

ROH = *Ritmálssafn Orðabókar Háskólans* [Det leksikografiske instituts historiske citatarkiv]. <ritmalssafn.arnastofnun.is> (maj 2024).

Anden litteratur

Barkarson, Starkaður, Steinþór Steingrímsson & Hildur Hafsteinsdóttir (2022): Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus. I: Nicoletta Calzolari et al. (eds.): *Proceedings of the Language Resources and Evaluation Conference*, *Marseille*, *France*. Marseille: European Language Resources Association. 2371-2381.

Bjarnadóttir, Kristín (1998): *Norræna verkefnið*. Upubliceret rapport.

Bjarnadóttir, Kristín (2012): The Database of Modern Icelandic Inflection. I: Guy Pauw et al. (eds.): *LREC 2012 Proceedings of the workshop of Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 – AfLaT 2012*. Istanbul: European Language Resources Association. 13-18.

Bjarnadóttir, Kristín, Kristín Ingibjörg Hlynsdóttir & Steinþór Steingrímsson (2019): DIM: The Database of Icelandic Morphology. I: *Proceedings of the 22nd Nordic Conference on*

Computational Linguistics, *(NoDaLiDa 2019, Turku, Finland)*. Linköping Electronic Conference Proceedings, No. 167, NEALT Proceedings Series, No. 42. Linköping: Linköping University Electronic Press. 146-154.

Hætta á heilsutjóni vegna loftmengunar frá eldgosum. Leiðbeiningar fyrir almenning [Fare for helbredskonsekvenser på grund af vulkanudbrud. Vejledning for almenheden] (2022). Reykjavík: Sóttvarnalæknir, Umhverfisstofnun, Veðurstofa et al. <almannavarnir.is/frettir/baeklingur-radleggingar-vegna-gasmengunar/>.

Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir (2011): ISLEX – en flersproget nordisk ordbog. I: Birgit Eaker, Lennart Larsson & Anki Mattisson (red.): *Nordiska studier i lexikografi* 11. Lund: Nordisk förening för lexikografi. 353-366.

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

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

Jónsson, Jón Hilmar (2005): Aðgangur og efnisskipan í íslensk-erlendum orðabókum – vandi og valkostir. I: *Orð og tunga* 7, 21-40.

Jónsson, Jón Hilmar (2015): Flerordslemmaer. Form og funksjoner. I: Caroline Sandström et al. (red.): *Perspektiv på lexikografi, grammatik och språkpolitik i Norden* (Vol. 39). Helsinki: Institutet för de inhemska språken. 116-133.

Sigurðsson, Einar & Steinþór Steingrímsson (2024): Representativeness and biases in Icelandic corpora. *LexicoNordica* 31 (dette bind).

Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson & Jón Guðnason (2018): Risamálheild: A Very Large Icelandic Text Corpus. I: Nicoletta Calzolari et al. (eds.): *Proceedings of LREC 2018, Eleventh International Conference on Language Resources and Evaluation.* Myazaki: European Language Resources Association. 4361-4366.

Úlfarsdóttir, Þórdís (2013): ISLEX – norræn margmála orðabók. I: *Orð og tunga* 15, 41-71.

Úlfarsdóttir, Þórdís (2014): ISLEX – a Multilingual Web Dictionary. I: Nicoletta Calzolari et al. (eds.): *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation.* Reykjavík: European Language Resources Association. 2820-2825.

Ellert Þór Jóhannsson
forskningslektor
Árni Magnússon-instituttet for islandske studier
Edda, Arngrímsgötu 5
IS-107 Reykjavík
ellert.thor.johannsson@arnastofnun.is

Þórdís Úlfarsdóttir
ordbogsredaktør
Árni Magnússon-instituttet for islandske studier
Edda, Arngrímsgötu 5
IS-107 Reykjavík
thordis.ulfarsdottir@arnastofnun.is

Fra 'sandheden om sproget' til et opgør med stereotyper: ikke-korpusbaserede metoder til leksikografiske beskrivelser af kontroversielle ord

Sanni Nimb

In the last decade, the lexicographic community has begun to realize that the description of sensitive lemmas concerning, e.g., gender, sexuality, and ethnicity cannot be based on the usual corpus linguistic methods due to the risk of biases and stereotypes in corpora. To be able to assess the extent to which sensitive lemmas are derogatory, offensive, or politically incorrect, we instead suggest a method where the lemmas are annotated by a group of language users representing different genders and ages. Exactly which lemmas are considered sensitive by (sometimes only parts of) the language community is also difficult to determine. This article presents an approach where information from existing usage labels in a dictionary is combined with thesaurus information.

1. Indledning

Da korpusleksikografiske metoder blev taget i anvendelse for mere end 30 år siden, blev de i de første mange år anset som meget velegnede til at undgå subjektive og holdningsprægede ordbogsoplysninger. I de senere år er der opstået en større bevidsthed om at dette ikke altid er tilfældet. Denne artikel diskuterer udfordringerne ved at basere beskrivelser af kontroversielle ord, det vil sige ord relateret til sensitive emner som fx køn, seksualitet, etnicitet og handicap, på korpuslingvistiske metoder – udfordringer der skyldes omfanget af stereotyper og bias i tekstkorpora, næsten uanset hvor balanceret de er sammensat. Artiklen fremlægger først en metode til at indkredse kontroversielle ord, idet også ord der kun

har potentiale til at blive kontroversielle, inkluderes. Den giver efterfølgende et bud på hvordan man kan sikre at beskrivelserne af de udvalgte ord ikke alene baseres på leksikografens egen intuition når stereotype data fra korpus forkastes. Den nødvendige indsigt i ordenes nedsættende, krænkende eller politisk ukorrekte konnotationer kan opnås ved at lade en gruppe lingvister opmærke ordene systematisk. Det er en vigtig pointe at de pågældende lingvister repræsenterer forskellige aldersgrupper og køn, fordi der ofte er stor variation i holdninger til ord inden for de pågældende semantiske områder. De opmærkede data kan både danne udgangspunkt for et særligt 'tabu'-leksikon til brug i NLP (natural language processing) og til angivelse af sprogbrugsoplysninger i en ordbog.

Indledningsvis (afsnit 2) beskriver jeg hvordan fremkomsten af korpusleksikografiske metoder blev modtaget i ordbogsmiljøer i 1980'erne og 90'erne, og hvordan det bl.a. influerede på redigeringen af *Den Danske Ordbog* (DDO; 2003-05) i 1990'erne og 00'erne. I afsnit 3 diskuterer jeg de i dag anerkendte problemer ved at bruge korpuslingvistiske metoder til beskrivelsen af sensitive ord. I afsnit 4 diskuterer jeg mulige strategier før jeg i afsnit 5 og 6 fremlægger metoder til at indkredse, henholdsvis annotere det sensitive ordforråd.

2. Korpusleksikografiens fremkomst og 'sandheden om sproget'

I ældre danske ordbøger, hvor beskrivelserne var baseret på få indsamlede belæg og leksikografens egen intuition, skal man ikke lede længe før man finder eksempler på holdningsprægede beskrivelser af fx køn og seksualitet. I *Moths Ordbog*, der beskriver dansk i årene omkring 1700, skinner et negativt syn på kvinder i samtidens mandsdominerede samfund fx tydeligt igennem. Man finder således en stor overvægt af negative kollokationer og udtryk

ved lemmaet *Kvinde* ("kvinde er Satans redskab", "Ond kvinde er verre end pest, end skærsîld", "Mand mâe ei troe kvinder, ikke den allerbeste") og også ved lemmaet *Kvind-folk* ("kvindfolk er altid galne", "Kvindfolk er ei at troe" og "alt ont kommer fra kvindfolk"). Også i *Ordbog over det danske Sprog* (dansk sprog i perioden 1700-1950), der blev redigeret 200 år senere i perioden 1918-56, skinner negative holdninger vedrørende køn og seksualitet af og til igennem. En overført negativ betydning af lemmaet *Højttaler*, 'højrøstet, meget talende kvinde', nævner fx *hustru* i beskrivelsen: "(jarg., spøg.) om (meget ell. højrøstet talende) kvinde, især hustru", og en betydning af lemmaet *kæresteri* ("m. h. t. personer af samme køn") dokumenteres med følgende citat: "en vis Art af unaturligt Kæresteri mellem Kvinder, hvor det just ikke kommer til perverse Aktus, men hvor der dog kan erkendes en usmagelig Intimitet i Forholdet".

Med korpusleksikografiens fremkomst i løbet af 1980'erne og 90'erne fik man mulighed for at undgå subjektive og holdningsprægede beskrivelser. Williams (2003) beskriver det ligefrem som et revolutionerende metodeskift inden for leksikografien:

> A revolution in dictionary making came with the development of corpus linguistics, built on the contextualist view of meaning, and its transfer to lexicographical practice through the COBUILD dictionaries.

Man opfattede resultater af korpusundersøgelser som objektive data og sandheden om sproget, uanset en samtidig erkendelse af at de statistiske resultater ofte afspejlede negative og stereotype holdninger til fx køn og minoriteter. DDO, der blev redigeret i årene 1992-2004, var en af de første fuldt ud korpusbaserede ordbøger, og man redigerede ordbogen ud fra netop denne holdning. Scheuer (1995; forfatteren var it-redaktør ved projektet) analyserer fx bigrammer fra korpuslingvistiske undersøgelser i *Den Danske Ord-*

bogs korpus (40 millioner ord, bestående af tekster fra perioden 1983-1992) og konkluderer at de ofte er meget stereotype. Han konstaterer at "[d]er gives dette billede af kønnene: manden er et åndsvæsen […] Kvinden er en kødvare. Og det billede skal reproduceres i en deskriptiv ordbog", og at "[d]en deskriptive ordbog er et vigtigt historisk dokument, og DDO portrætterer samfundet og samfundsideologien som den ser ud i tiåret 1983-92" (Scheuer 1995:255). Eksempler på stereotype bigrammer som påpeges af Scheuer, er *hans kolleger*, *hans kontor*, *hans ideer*, *hans firma* over for *hendes hånd*, *hendes hår*, *hendes mor*, *hendes latter*. De stereotype bigrammer udgjorde materiale til kollokationerne i den første, trykte udgave af DDO, kollokationer der stadig findes i nutidens online-udgave af ordbogen. Man finder fx under opslagsordet *kvinde* disse kollokationer: *gifte kvinder*, *enlig kvinde*, *en smuk kvinde*, *kvinder og børn*, og under opslagsordet *mand* i stedet disse: *klog mand*, *rig mand*, *stærke mænd*, *rigtige mænd*. Fjeld (2015) påpegede mange stereotype citater i DDO, jf. fx citatet "[kvinderne] er eftergivende, indordner sig og accepterer at arbejde og leve på andres præmisser" under opslagsordet *eftergivende*.

De stereotype bigrammer og citater skyldtes ikke at DDO-korpusset var specielt ensidigt opbygget. Tværtimod lagde man i DDO-projektet meget stor vægt på at gøre det så velafbalanceret som muligt. Teksterne omfattede både aviser, ugeblade og magasiner, litteratur, private tekster i form af dagbøger og endda også talesprog. Scheuer (1995:246) skriver om indsamlingen af tekster til korpusset: "Der blev altså sat visse minimumskrav til spredningen – i praksis blev spredningen enorm". Der var også en vis balance i skribenternes køn, idet 1/3 var kvinder, og bigrammer målt på kun tekster af kvinder var i øvrigt også i lige så høj grad stereotype som dem beregnet på det samlede korpus (egen undersøgelse i 1996 da jeg var ansat som redaktør ved DDO; ikke publiceret). Dog var en væsentlig andel af teksterne avistekster. Da avistekster netop er den genre der indeholder flest stereotyper iføl-

ge Müller-Spitzer & Rüdiger (2022), var dette måske med til at øge antallet af stereotype bigrammer og citater.

3. Korpusleksikografiens særlige udfordringer: kontroversielle ord

I det seneste årti er der kommet et andet syn på reproduktion af de stereotyper der udledes statistisk af store tekstsamlinger, end det der lå bag redigeringen af DDO i 1990'erne. Inden for fagområdet leksikografi bliver emnet diskuteret bredt, fx i Fjeld (2015), hvor stereotyper i nordiske ordbøger undersøges, og i Petersson & Sköldberg (2020), hvor beskrivelsen af nogle udvalgte kontroversielle ord fremlægges og diskuteres, herunder *hora* ('hore'), *rödskinn* ('rødhud') og *bög* ('bøsse'). Müller-Spitzer & Rüdiger (2022) fremlægger en undersøgelse af stereotyper i tre genremæssigt forskellige korpora, nemlig et bestående af fiktionstekster, et bestående af avistekster og et bestående af ugeblade/magasiner; en undersøgelse der viser at kønsstereotyper er særligt udbredte i avistekster. Også uden for det leksikografiske område diskuteres problemet med bias og stereotyper i tekster; således konstaterer Huyssteen & Tiberius (2023) at den nyeste udvikling inden for AI og fremkomsten af sprogmodeller som fx OpenAI's GPT-4 og ChatGPT har gjort emnet højaktuelt. Håndteringen af stereotyper i store datamængder medfører i det hele taget mange etiske dilemmaer for firmaer bag sprogteknologiske produkter. Google har valgt at påtage sig et socialt ansvar ved ikke at reproducere bias og stereotyper og beskriver dette som et af deres vigtigste principper, idet de samtidig erkender at det ikke er nemt at afgøre hvad der er stereotypt og kontroversielt, og at det varierer afhængigt af kultur og samfund, se Google's AI Principles, princip 2. Huyssteen & Tiberius (2023) påpeger at problemet ikke kun er reproduktionen af bias og statistisk baserede stereotyper, men at der også i automa-

tisk sprogbehandling og NLP er et behov for at indkredse hvilke ord der i det hele taget er tabubelagte og kontroversielle. Lister over kontroversielle ord er påkrævet til fx automatisk identifikation af krænkende sprog på sociale medier. Oplysning om hvorvidt ord bruges bevidst for at vække anstød eller måske utilsigtet kan opfattes krænkende fordi det er tabubelagt i dele af sprogsamfundet, er et nødvendigt supplement til de polaritetsoplysninger i sentimentleksikoner der anvendes til sentimentanalyse (automatisk genkendelse af negativt, henholdsvis positivt ladede ord). *Det Danske Sentimentleksikon* (Nimb et al. 2022) indeholder også kun polaritetsværdier og ikke oplysninger om hvor kontroversielle ordene er, fx er *gebyrgrib*, *miljøsvin*, *negermusik* og *perker* beskrevet helt ens (stærkeste negative værdi). Information om hvor kontroversielle ordene er (*negermusik* og *perker* er i høj grad, *miljøsvin* og *gebyrgrib* er ikke), ville kunne tilføjes automatisk når DDO-lemmaer er opmærket som beskrevet i denne artikel, idet ord i sentimentleksikonet er koblet direkte til DDO på betydningsniveau. Huyssteen & Tiberius (2023) nævner to eksempler på lister over kontroversielle ord der anvendes i NLP, et italiensk og et japansk.

Der var allerede visse indvendinger mod at anse korpuslingvistiske metoder som absolut objektive i 1990'erne og årene omkring år 2000. Hidalgo-Tenorio (2000) nævner fx at det jo stadig er redaktører der ud af mange forekomster i et korpus udvælger netop de belæg på en betydning som vedkommende anser som bedst egnede til at illustrere den normale sprogbrug. I redaktionen bag den sydafrikanske ordbog *Woordeboek van die Afrikaanse Taal* (WAT, the Bureau of the WAT[1]) baserede man midt i 1990'erne det leksikografiske arbejde på andre principper end CoBuild- og DDO-projekternes når det kontroversielle ordforråd skulle beskrives (Harteveld & van Niekerk 1995 og 1996). The Bureau of

1 Bureau of the WAT (journals.co.za/publisher/botw) er et leksikografisk institut beliggende i Stellenbosch og grundlagt i 1926. Dets hovedopgave er udarbejdelsen af en omfattende betydningsordbog for afrikaans, *Woordeboek van die Afrikaanse Taal*.

the WAT så det som en samfundspligt i tiden efter apartheids ophør i Sydafrika at undlade at reproducere stereotyper. De redaktionsregler der blev formuleret til WAT ud fra denne holdning, er efter min mening blevet meget aktuelle i nutidens leksikografiske arbejde. The Bureau of the WAT udtrykte et direkte ønske om at spille en aktiv rolle i arbejdet med at fremme ligestilling i Sydafrika og dermed udvise forståelse for "a problem which caused great pain, indignation and interpersonal alienation in South Africa" (Harteveld & van Niekerk 1996:393). Hverken i den digitale eller trykte udgave af ordbogen måtte der ved "Insulting and Sensitive Lexical Items" – sensitive eller kontroversielle ord – optræde kollokationer, redaktionelle eksempler eller citater der afspejlede en negativ holdning til befolkningsgrupper. Et eksempel som *you cannot trust a black person with the building process* var fx ikke acceptabelt. I den trykte udgave af ordbogen var man endnu mere restriktiv. Der var hverken kollokationer, redaktionelle eksempler, antonymer, referencer, citater eller andre brugseksempler ved kontroversielle lemmaer, og der måtte ikke bringes krænkende synonymer ("hurtful synonyms"). Racistiske lemmaer blev decideret udeladt i den trykte udgave af ordbogen.

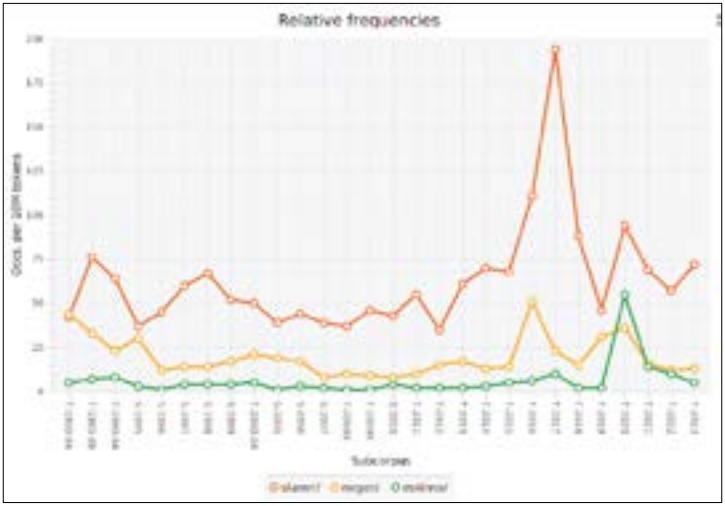
4. Kontroversielle ord: problemstillinger og mulige strategier

I dag har ordbogsredaktioner efter min mening pligt til at forholde sig til problemet med bias og stereotyper i store tekstmængder når ordbogsarbejdet er korpusbaseret. I DDO-projektet arbejdes der fx målrettet med at nedgradere synligheden af ældre stereotype citater fra det oprindelige DDO-korpus og erstatte dem med nyere, mere neutrale citater i den direkte visning (jf. Jensen et al. 2018, Trap-Jensen 2020). Nye redaktionsregler for valørangivelser ved DDO-lemmaer er også under udarbejdelse, bl.a. baseret på erfa-

ringer med redigering af et mindre antal kontroversielle ord relateret til både etnicitet, seksualitet og køn.

I arbejdet med kontroversielle ord indgår to problemstillinger. Den første er selve afgrænsningen af det relevante ordforråd. Petersson & Sköldberg (2020) påpeger at en del af problematikken vedrørende kontroversielle og diskriminerende ord er at de er svære at indkredse. Forfatterne påpeger også at man ikke kan nøjes med at se på de ord der allerede er beskrevet som nedsættende i ordbøger, fordi sproget forandrer sig hurtigt, og fordi ord der hidtil har været neutrale, pludselig bliver diskriminerende eller kontroversielle, fx på grund af påvirkning fra andre sprogsamfund. Fordi vi ved at kontroversielle ord, måske især de ord der ikke er enighed om i sprogsamfundet, typisk debatteres i medierne, kunne et særligt mønster i korpusfrekvens over tid, en såkaldt ordprofil, måske afsløre de ord der pludselig bliver kontroversielle. I perioden hvor et kontroversielt ord debatteres, vil dets frekvens være stærkt stigende i et aviskorpus, herefter vil det formentlig få en lavere frekvens end det havde før mediedebatten startede. Ved hjælp af DDO-redaktionens korpusværktøjer kan dette undersøges i DDO's korpus Bakspejlet (kun delvis offentligt tilgængeligt; bestående af mere end 1,3 milliarder ord fra primært aviser, se også Appel, Sørensen & Jensen, dette bind). Der er en tendens til at hypotesen er korrekt når man ser på ordene *neger*, *slaver* og *eskimo*, som vi med sikkerhed ved har været debatteret offentligt, og hvor i hvert fald *eskimo* som noget relativt nyt nu af mange opfattes som politisk ukorrekt (se figur 1). Men da også ikke-kontroversielle ord af andre grunde vil kunne udvise samme ordprofil i et korpus, kan metoden kun bruges til at få bekræftet en mistanke om at et kontroversielt ord rent faktisk også *er* blevet debatteret på et givet tidspunkt.

Brugerhenvendelser er til gengæld en direkte kilde til indkredsning af potentielt kontroversielle ord. Under den igangværende fjerde bølge af feminisme er det DDO-redaktionens erfaring



Figur 1: Ordprofiler for strengene *slaver*, *neger* og *eskimo* 1930-2023 i korpusset Bakspejlet.

at mange brugere vælger at henvende sig til redaktionen som en form for aktivistisk handling, i tråd med hvad Laura Na Blankholm skriver i dette indlæg: "I den fjerde bølge er internettet og sociale medier en central kampplads" (dagbladet Information 6. oktober 2017). Et godt eksempel på dette er mange samtidige henvendelser til redaktionen i 2019 vedrørende lemmaet *hudfarve*. En bruger skrev fx: "Hvordan kan definitionen på ordet 'hudfarvet' være ferskenfarvet/nordeuropæisk, når der er mange, mange mennesker som ikke er lyse i huden (også her i Danmark)". Et andet eksempel er denne henvendelse vedrørende ordet *grønlænderstiv*; igen blot en af mange samtidige i 2020: "Det er fuldstændig lige meget om ordet findes eller ej. I er med til og dele ordet ud til verden mens vi kæmper her for at få det fjernet! Fjern det fra jeres ordbog." Men selvom brugerhenvendelser bidrager med oplysninger om nye kontroversielle ord, ville det være langt at foretrække at være på forkant, så svaret oftere kunne være at vi allerede har observeret problemet og har noteret ved opslagsordet at beskrivelsen eventuelt bør ændres.

Den anden problematik er leksikografens vurdering af om lemmaet anvendes bevidst nedsættende eller måske kan opfattes som politisk ukorrekt eller krænkende af dele af sprogsamfundet uden at det var tilsigtet – og i hvilken grad det kan. Petersson & Sköldberg (2020) påpeger at det her er næsten umuligt ikke at basere vurderingen på egen sprogfornemmelse, også selv om man undersøger andet materiale som fx slangordbøger, og at det derfor ikke er uden risiko. Brugerhenvendelser til DDO's redaktion viser med al tydelighed at der er stor variation i holdninger til ord inden for køn, ligestilling, seksualitet og etnicitet, en variation der naturligvis også må eksistere blandt leksikografer. Et meget illustrerende eksempel for problematikken er lemmaet *neger*s beskrivelse i DDO. I øjeblikket har ordet sprogbrugsmarkeringen 'oftest nedsættende' for at afspejle at der ifølge redaktørernes opfattelse er tale om divergerende holdninger i sprogsamfundet. Om det evt. er aldersbetinget, er ikke specificeret; den præcise viden om dette kan ikke udledes af korpus. Sprogbrugsoplysningen har ført til to næsten samtidige brugerhenvendelser til DDO-redaktionen. Den ene bruger skriver: "[O]m ordet neger brugt om sorte og afrikanere, skriver i: "oftest nedsættende", men i dagens sprogbrug er det altid nedsættende, medmindre man er meget uoplyst". Den anden bruger skriver: "Jeg ser under opslaget om ordet 'neger', at sprogbrugen ifølge jer oftest er nedsættende. Dette undrer mig, for jeg har aldrig selv brugt ordet nedsættende eller hørt andre bruge ordet nedsættende. Det er et helt almindeligt objektivt ord, som beskriver det, det gør ifølge jeres egen ordbog. En afrikaner. En sort". Med andre ord tilfredsstiller sprogbrugsmarkøren 'oftest nedsættende' tydeligvis hverken den ene eller den anden gruppe. Her står man som redaktør i en vanskelig situation. Det er en nærmest umulig opgave dels at definere mere teoretisk hvordan et tekstkorpus skal opbygges, så det fyldestgørende repræsenterer de forskellige holdninger, dels i praksis at opnå rettigheder til at indsamle tekster til et sådant korpus, fx fra sociale medier. For at

sikre en videnskabelig metode foreslår jeg at man i stedet på en systematisk måde indsamler oplysninger om holdninger blandt flere annotører som – meget vigtigt for metoden – er af forskellig alder og køn, og lader den opnåede viden danne udgangspunkt for den leksikografiske formidling. Jeg vil i afsnit 6 give et bud på hvordan det kan gribes an, men først vil jeg i næste afsnit beskrive en metode der angår den første problematik, nemlig indkredsning af selve ordforrådet.

5. Udpegning af kontroversielle lemmaer

Metoden tager udgangspunkt i ord der allerede er beskrevet som nedsættende i andre ordbøger, i mit eksempel DDO. De lemmaer der er markeret som nedsættende i DDO, udgør ca. 1,5 % af ordbogen, og interessant nok er denne procentdel stabil over tid; den gælder både for det ordforråd der blev redigeret til den trykte DDO i årene 1994-2004, og for det der er redigeret i årene 2007-2024, og som løbende publiceres i DDO online. Valørangivelserne er 'nedsættende', 'nedsættende eller spøgende', 'ofte nedsættende', 'især nedsættende' og 'stærkt nedsættende'. Langt fra alle nedsættende ord i DDO er det vi forstår ved kontroversielle. *Lov om Ligebehandlingsnævnet* nævner i § 1 de områder som nævnet behandler: "Ligebehandlingsnævnet behandler klager over forskelsbehandling på grund af køn, race, hudfarve, religion eller tro, politisk anskuelse, seksuel orientering, alder, handicap eller national, social eller etnisk oprindelse"[2] (2016). Ud fra dette (og idet der ses bort fra ord vedrørende politisk anskuelse) har vi indledningsvis grupperet alle nedsættende ord i den trykte DDO og opstillet fire overordnede kontroversielle kategorier. Den første er 'køn, seksu-

2 Petersson & Sköldberg (2020) anvender i øvrigt, uafhængigt af os, en helt parallel metode på baggrund af lov om diskrimination i Sverige fra 2009 og lignende information på hjemmesiden for den svenske ligestillingsombudsmand.

alitet', den anden er 'handicap, udseende, alder'. Derudover har vi kategorien 'etnisk oprindelse, race, hudfarve, religion/tro' og endelig kategorien 'social oprindelse' (en kategori som vi dog endnu ikke har set så meget på).

Fordi man, som Petersson & Sköldberg (2020) fastslår, ikke kan nøjes med kun at se på ord der allerede er beskrevet som nedsættende, skal listen udvides med de nærsynonymer der ifølge DDO er neutrale, men som har potentiale til at blive nedsættende baseret på deres betydning og ordets eller orddelenes karakter. Da DDO's ordforråd er emneinddelt i *Den Danske Begrebsordbog* (Nimb et al. 2014) og koblet direkte sammen med den på betydningsniveau, kan bruttolisten af kontroversielle ord, både de i forvejen nedsættende og deres nærsynonymer, forholdsvis nemt indsamles på en liste (samme metode anvendtes til at udarbejde *Det Danske Sentimentleksikon* (Nimb et al. 2022)). Begrebsordbogen indeholder 888 navngivne afsnit, herunder også afsnit som 'Fremmede, udlandet', 'Ligestilling', 'Sygdom', 'Sex og begær'. Fordi oplysningen om nedsættende valør er overført automatisk fra DDO til Begrebsordbogens ordforråd (i forenklet form 'neds.'), kan vi nemt fokusere på de grupper i et afsnit der har mange nedsættende ord, se figur 2. Både de i forvejen nedsættende ord og de ord i samme gruppe der vurderes til også at være eller kunne blive kontroversielle og diskriminerende, medtages på listen.

Ved hjælp af denne metode har vi foreløbig indkredset 1.200 kontroversielle og potentielt kontroversielle ord og udtryk, flest af typen 'køn, seksualitet' og 'etnisk oprindelse, hudfarve, national oprindelse' (ca. 400 ord i hver kategori), men også mange vedrørende 'handicap, udseende, alder' (ca. 350 ord, fx *krykhusar*, *klumpfodet*, *dværg*, *hele den pukkelryggede (familie)*; *friluftsgebis*, *hængepatter*, *blåøjethed*, *brunette*; *aldersbyrde*, *gammelmandsbarn*, *bessemor*). Gruppen der vedrører 'social oprindelse', indeholder indtil videre kun 60 ord, fx *bonderøv*, *andedam*, *provinsiel*; gruppen skal suppleres med værdiladede ord der vedrører både lav og

Figur 2: Fire forskellige afsnit der er inddraget fra Begrebsordbogen i indkredsningen af kontroversielle ord: 'Menneskets udseende', 'Indbygger, befolkning', 'Fremmede, udlandet' og 'Tyk'. Nogle af ordene har i forvejen nedsættende valør i DDO (fx *neger* og *flæskebjerg*), andre har ikke (fx *farvet* og *fedtklump*), men bør vurderes og evt. revideres fordi de tilhører samme semantiske felt.

høj social status. Ord der vedrører religion, er ikke endeligt indkredset. Alle fire hovedkategorier bliver løbende udvidet med ord når nye tilfælde opdages, fx på grund af brugerhenvendelser.

Med henblik på at opnå mindre grupper af meget beslægtede ord underinddeles de fire hovedkategorier yderligere. De 400 ord i gruppen 'etnisk oprindelse, hudfarve, national oprindelse' inddeles fx i disse otte underkategorier:

1. ord der afspejler forældet teori om menneskeracer (*mulat*, *halvblods*, *negroid*)
2. ord der afspejler at ikke-vestlig kultur anskues fra et vestligt synspunkt (*naturfolk*, *indianerhyl*)
3. (fremmed) person der er beskrevet ud fra medfødt udseende eller spisevaner (el. anden kropslig egenskab) (*spaghetti* ('italiener'), *den hvide mand*)
4. fremmed folk/folkeslag der er beskrevet med en ikke-national/ikke-geografisk betegnelse (*kaffer*, *sigøjner*, *lap*/*laplænder*)
5. ord hvor etnicitet er del af et (negativt) ord eller udtryk med anden betydning (*grønlænderstiv*, *negerbolle*, *bande som en tyrk*, *du milde kineser*, *fransk horeunge*, *squaw* (i den overførte betydning 'hustru'))
6. ord vedrørende indvandring i Danmark (*perker*, *perkerdansk*)
7. ord vedrørende kolonitiden (*koloniherre*, *slaveopstand*)
8. evt. værdiladet, ikke-officiel betegnelse for statsborger (*thaipige*, *kineserinde*, *kartoffeltysker*)

På denne måde kan tæt beslægtede beskrivelser i DDO nemmere sammenlignes ud fra en diskriminationsvinkel og en ligestillingsbetragtning. Man ser fx hurtigt at nogle af ordene ikke har nogen sprogbrugsoplysning i DDO, selv om de i dag kan opfattes nedsættende eller krænkende. Nogle eksempler er *burkabil* og *squaw*. Ligeledes har *gringo*, *lap* og *laplænder* samt udtrykket *bande som en tyrk* ('bande kraftigt og ofte') ingen valørangivelse. Substantivet *lapp* er både markeret som stærkt nedsættende og gammeldags i *Svensk Ordbok*, og den danske ækvivalent bør derfor undersøges nærmere. Man opdager også at der kan være forskel på valøroplysninger ved ord der i øvrigt ligner hinanden, fx er *flæskebjerg* ('overvægtig person') 'nedsættende', mens *fedtklump* i samme betydning blot er 'uformelt'. Og *abekat* ('person der opfører sig vildt

og uciviliseret eller tåbeligt og naragtigt'), er 'nedsættende eller spøgende', mens *barbar* ('rå person der stammer fra et fremmed, uciviliseret land') blot er markeret 'især historisk'.

Nyligt indsatte, mere detaljerede valøroplysninger ved prøveord i DDO kan også være meget forskellige inden for undergrupperne (som nævnt arbejdes der i disse år med at opstille nye redaktionsregler). I tabel 1 ses eksempler på sprogbrugsoplysninger for underkategorien 'ord der afspejler at ikke-vestlig kultur anskues fra et vestligt synspunkt':

heksedoktor	person der menes at besidde magisk-religiøse kræfter til bl.a. at helbrede sygdomme, beskytte mod fremmed magi og spå om fremtiden – kendes især fra stammesamfund SPROGBRUG forældet | nu oftest nedsættende
indianer	person tilhørende et af de folk som sammen med inuit og aleuter udgør den oprindelige befolkning i Nord- og Sydamerika – som alternativ bruges nu ofte udtrykket *oprindelig/indfødt amerikaner*, idet *indianer* kan opfattes stødende
indianerblod	**have indianerblod i årerne**: tilhøre eller være efterkommer af den oprindelige befolkning i Amerika (med undtagelse af det nordligste Nordamerika) SPROGBRUG kan virke stødende
indianerhyl	meget kraftigt og vildt hyl SPROGBRUG ordet kan være problematisk fordi det afspejler en nu forældet opfattelse af Amerikas indfødte befolkning som vild og blodtørstig, som fx fremstillet i ældre westernfilm
naturfolk	folk der lever i nær kontakt med og i umiddelbar afhængighed af naturen under anvendelse af enkel teknologi SPROGBRUG denne brug kan være problematisk da den indebærer en forsimplet modstilling af natur og kultur

Tabel 1: Sprogbrugsoplysninger for ord i underkategorien 'ord der afspejler at ikke-vestlig kultur anskues fra et vestligt synspunkt'.

De opstillede lister og undergrupper kan danne udgangspunkt for det videre arbejde frem mod ændrede redaktionsregler. Den store udfordring er at opnå viden om ordenes valør, idet korpusundersøgelser suppleret med introspektion som beskrevet ovenfor

ikke nødvendigvis er repræsentativt for de holdninger der findes i sprogsamfundet.

6. Opmærkning af kontroversielle ord

For at sikre detaljeret viden om holdninger til sensitive ord, herunder den variation der ofte er i holdninger blandt sprogbrugerne, kan man vælge at bede en række lingvister af forskellig alder og køn opmærke ordene. Man kan fx tage udgangspunkt i de oplysningstyper med et tilhørende lukket inventar af værdier der foreslås af Huyssteen & Tiberius (2023). Oplysningstyperne er fastlagt i et samarbejde mellem en sydafrikansk og en hollandsk leksikografisk institution og baseret på en tilbundsgående analyse af problemstillingen med kontroversielle ord. Begge lande har stor erfaring med at arbejde med leksikografiske beskrivelser af det kontroversielle og sensitive ordforråd. I Holland har man fx gennem en lang årrække haft tradition for at udgive særlige ordbøger der beskriver det tabubelagte ordforråd, og situationen i Sydafrika er omtalt ovenfor. I store træk er de nødvendige oplysningstyper ifølge de to forfattere følgende (udfyldt med mine egne intuitive bud på danske eksempler). To oplysninger anvendes til at beskrive hvor ofte og i hvor høj grad ordet er kontroversielt:

- I hvor høj grad er lemmaet tabubelagt (<taboo Value>)? *Nigger* er fx i højeste grad, *thaipige* i lav grad.
- Hvor prototypisk for lemmaet er den kontroversielle betydning (<taboo Prototypicality>)? Hvis ordet uanset kontekst altid er kontroversielt, er svaret 'altid' (fx *nigger* og *neger*), andre værdier er 'ofte' (fx *indianer*), 'undertiden' (fx *slave*) og 'sjældent' (fx *handicappet*).

Hensigt og virkning ved anvendelse af ordene undersøges desuden ved hjælp af tre oplysningstyper:

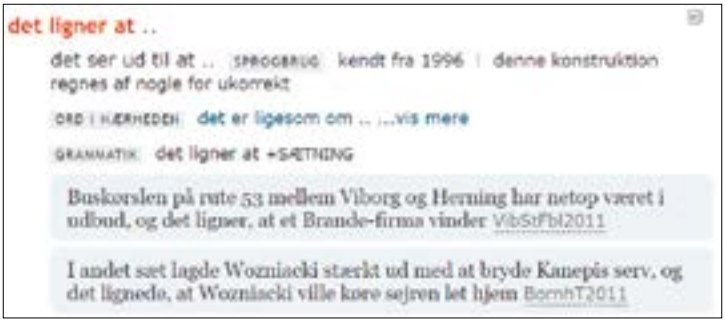
- Bruges lemmaet (i denne betydning) ofte eller mest som skældsord (<speechAct>)? I tilfældene *bøsserøv* og *nigger* er svaret ja.
- I hvor høj grad har den talende en bevidst hensigt med at bruge krænkende sprog når lemmaet i denne betydning anvendes <illocution>? *Sortsmudsker*, *negermusik* og *betonlebbe* er fx bevidst nedsættende.
- Opfatter modtageren dette lemma som stærkt krænkende, lidt krænkende, politisk ukorrekt eller ingen af delene (<perlocution>)? *Indianer* og *uland* opfattes fx ofte som politisk ukorrekt, *mulatpige* som krænkende.

Endelig er der mulighed for at angive et eventuelt neutralt synonym (<orthophemism>), fx *roma* som alternativ til *sigøjner*.

Også ord der ikke umiddelbart er kontroversielle, men som har potentiale til at blive det, fx fordi første- eller sidsteleddet er kontroversielt, skal opmærkes af lingvisterne og vil formentlig få så lave værdier at de blot sættes på en observationsliste. Et eksempel kunne være substantivet *slaverom* på grund af førsteleddet *slave-* og sprogbetegnelsen *eskimoisk* på grund af *eskimo*s kontroversielle karakter.

Idealet er at yngre og ældre sprogbrugere af forskelligt køn er repræsenteret blandt annotørerne, og at det blot er deres intuitive holdninger til de enkelte ord der skal anføres, så arbejdet ikke er for tidskrævende. Når intuitive holdninger indsamles blandt en større og til en vis grad repræsentativ gruppe personer, sikrer man bedre at variationen i holdninger i sprogsamfundet afspejles (jf. Ipsos MORI (2021), hvor fokusgrupper sammensat af personer af forskelligt køn, alder og etnicitet bl.a. danner udgangspunkt for sproglige holdningsanalyser). De ord der får varierende oplys-

ninger på tværs af annotørgruppen, er selvfølgelig særligt udfordrende at beskrive i en ordbog. Man kan vælge at lade variationen udgøre en del af beskrivelsen ud fra den betragtning at det udgør en vigtig viden om ordet der bør videreformidles. Plank (2022) argumenterer fx for at variation blandt sprogbrugere og annotører er meget vigtig information der bør inkluderes i annoterede data frem for den hidtidige praksis i NLP (og inden for leksikografien) hvor lingvister og leksikografer forhandler sig til enighed eller beslutter sig for kun at anvende én af værdierne. I NLP opererer man traditionelt med såkaldte guldstandarder med kun én værdi som datagrundlag. I leksikografi kan oplysninger altid modificeres (fx i form af 'ofte(st)' i sprogbrugsmarkøren 'oftest nedsættende' ved *neger* i DDO), men det fremgår sjældent tydeligt hvori variationen består, og det kan, som beskrevet ovenfor, skabe problemer. Variation med hensyn til grammatisk korrekthed fremgår fx mere tydeligt i DDO. Ved udtrykket *det ligner at* .. anføres det fx at "denne konstruktion regnes af nogle for ukorrekt", se figur 3. Her er det underforstået at *nogle* refererer til personer der er uddannet inden for danskfaget.



Figur 3: Udtrykket *det ligner at* .. har i DDO oplysningen: "denne konstruktion regnes af nogle for ukorrekt".

Man kan også vælge en mere enkel løsning, fx blot at angive en advarsel ved ordet eller at tolke krænkende virkning som et ud-

slag af at brugen i dag er blevet 'gammeldags', vel at mærke i de tilfælde hvor det er de ældre annotører der opfatter ordet som uproblematisk, mens de yngre annotører undgår at bruge ordet og opfatter det som krænkende (i den pågældende betydning). *Bokmålsordboka* angiver fx blot at *eskimo* er en "foreldet betegnelse", og *Svensk Ordbok* angiver blot at *flicka* i betydningen 'kvinna' er "något ålderdomligt". Hvis det er svært at udlede ud fra korpusundersøgelser (idet skribenternes alder ikke kendes), bør grundlaget for denne formidling ideelt set også bygge på indsamlede data fra en aldersmæssigt bredt sammensat gruppe. I princippet bør opmærkningsarbejdet også gentages forholdsvis ofte sammenlignet med andet nødvendigt revideringsarbejde i et ordbogsprojekt, fx udgiver BBC (BBC blog (2010)) opdaterede sproglige vejledninger til medarbejderne vedrørende kontroversielle ord i deres udsendelser hvert fjerde-femte år. Det bør derfor udformes som en ikke alt for tidskrævende annotationsopgave, hvor alene intuition danner grundlag for den enkelte annotørs opmærkning, men hvor antallet af annotører og spredning i alder og køn derimod er meget afgørende.

7. Konklusion

Der er mange leksikografiske udfordringer i arbejdet med at undersøge og beskrive kontroversielle ord i en erkendelse af at de tekstkorpora, der i øvrigt danner et rigtig godt udgangspunkt for deskriptivt leksikografisk arbejde, per se er biased og indeholder mange stereotyper. Problematikken er højaktuel på grund af den øgede opmærksomhed på ligestilling i samfundet og fremkomsten af AI og sprogmodeller der bygger på statistiske beregninger på store tekstmængder. I det leksikografiske arbejde er stereotype bigrammer og citater ud fra vores erfaring nemmest at håndtere redaktionelt, hvorimod angivelsen af sprogbrugsmarkører

ved kontroversielle ord indebærer flere udfordringer. Ordene er svære at indkredse i ordbog og korpus, og det er svært at tildele sprogbrugsmarkører fordi der kan være stor variation i holdningen til ordene blandt både sprogbrugere og ordbogsredaktører. Leksikografer har ekspertisen til at indsamle og beskrive det sensitive ordforråd i detaljer, og man kan med fordel anvende annotationsmetoder der kendes fra arbejdet med sprogdata i NLP. Man opnår et videnskabeligt datagrundlag hvis både ældre og yngre informanter af forskelligt køn involveres i opmærkning af ordforrådet, og det kan også anvendes til at formidle interessant viden om variation blandt sprogbrugerne. For at være på forkant med sprogudviklingen i samfundet bør også ord der blot har potentiale til at blive kontroversielle, opmærkes. Udfordringerne er at annotering er tidskrævende, især når mange skal involveres, og at det hurtigt forældes fordi sprogbrugen måske særligt i disse år er i hastig udvikling inden for området.

Litteratur

Ordbøger, korpusser og digitale resurser

Bakspejlet = DDO's interne korpus, Det Danske Sprog og Litteraturselskab (2024).

BBC blog (2010): New edition of BBC's Editorial Guidelines. <bbc.co.uk/blogs/theeditors/2010/10/new_edition_of_bbc_editorial_g.html> (archived page).

Begrebsordbogen = *Den Danske Begrebsordbog*.

Bokmålsordboka. Språkrådet og Universitetet i Bergen. <ordbokene.no/nob/bm> (april 2024).

Den Danske Begrebsordbog (2014). Sanni Nimb, Henrik Lorentzen, Liisa Theilgaard & Thomas Troelsgård. København/Odense: Det Danske Sprog- og Litteraturselskab og Syddansk Universitetsforlag.

Den Danske Ordbogs korpus = Ole Norling-Christensen & Jørg Asmussen (1998): The Corpus of The Danish Dictionary. I: *Lexikos* 8. doi:10.5788/8-1-955.

COBUILD (1987) = *Collins COBUILD English Language Dictionary*. Editor in Chief: John Sinclair, Managing Editor: Patrick Hanks. London/Glasgow: Collins.

DDO (2003-05) = *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab og Gyldendal.

DDO online = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ddo> (april 2024).

Det Danske Sentimentleksikon = <github.com/dsldk/danish-sentiment-lexicon> (april 2024).

Google's AI Principles = <ai.google/responsibility/principles> (april 2024).

Lov om Ligebehandlingsnævnet (2016, LBK nr. 1230 af 02/10/2016). Bekendtgørelse af lov om Ligebehandlingsnævnet. Beskæftigelsesministeriet. <retsinformation.dk/eli/lta/2016/1230>.

Moths Ordbog. Det Danske Sprog- og Litteraturselskab. <mothsordbog.dk> (april 2024).

Ordbog over det danske Sprog. Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ods> (april 2024).

Svensk Ordbok = *Svensk Ordbok utgiven av Svenska Akademien*, Göteborgs universitet. <gu.se/svenska-spraket/svensk-ordbok> (april 2024).

Woordeboek van die Afrikaanse Taal (WAT). <woordeboek.co.za> (april 2024).

Anden litteratur

Appel, Kirsten, Nathalie Hau Sørensen & Jonas Jensen (2024): Jagten på hverdagssproget – brugen af tekster fra internetfora i arbejdet med Den Danske Ordbog. I: *LexicoNordica* 31 (dette bind).

Fjeld, Ruth Vatvedt (2015): Om ordbokseksempler og stereotypisering av kjønn i noen nordiske ordbøker. I: Caroline Sandström, Ilse Cantell, Eija-Riitta Grönros, Pirkko Niolijärvi & Eivor Sommadahl (red.): *Perspektiv på lexikografi, grammatik och språkpolitik i Norden*. Helsingfors: Institutet för de inhemska språken. 35-65.

Harteveld, Pieter & Angelique E. van Niekerk (1995): Policy for the Treatment of Insulting and Sensitive Lexical Items in the *Woordeboek van die Afrikaanse Taal*. I: *Lexikos* 5. doi:10.5788/5-1-1068.

Harteveld, Pieter & Angelique E. van Niekerk (1996): Policy for the Treatment of Insulting and sensitive Lexical Items in the Woordeboek van die Afrikaanse Taal. I: Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström & Catarina Röjder Papmehl (red.): *Euralex '96 Proceedings I-II*. Gøteborg: Göteborg University, Department of Swedish. Part 1, 381-393.

Hidalgo-Tenorio, Encarnación (2000): Gender, Sex and Stereotyping in the Collins COBUILD English Language Dictionary. I: *Australian Journal of Linguistics 20(2)*, 211-230. doi:10.1080/07268600020006076.

Huyssteen, Gerhard B. van & Carole Tiberius (2023): Towards a lexical database of Dutch taboo language. I: Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček & Simon Krek (eds.): *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. 53-74. <elex.link/elex2023/wp-content/uploads/elex2023_proceedings.pdf>.

Ipsos MORI (2021): *Public attitudes towards offensive language on TV and radio: Summary Report*. <ipsos.com/en-uk/public-attitudes-towards-offensive-language-tv-and-radio> (april 2024).

Jensen, Jonas, Henrik Lorentzen, Sanni Nimb, Mette-Marie Møller Svendsen & Lars Trap-Jensen (2018): Thaipiger, muskelhunde og fulde svenskere: nedsættende ord, stereotyper og ligestilling i *Den Danske Ordbog*. I: Ásta Svavarsdóttir, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdís Úlfarsdóttir (red.): *Nordiske Studier i Leksikografi* 14. Reykjavik: Nordisk Forening for Leksikografi. 141-151.

Müller-Spitzer, Carolin & Jan Oliver Rüdiger (2022): The influence of the corpus on the representation of gender stereotypes in the dictionary. A case study of corpus-based dictionaries of German. I: Annette Klosa-Kückelhaus, Stefan Engelberg, Christine Möhrs & Petra Storjohann (eds.): *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany*. Mannheim: IDS-Verlag. 129-141. <euralex.org/wp-content/uploads/2022/09/EURALEX2022_Proceedings.pdf>.

Nimb, Sanni, Nikolai Hartvig Sørensen & Thomas Troelsgård (2018): From Standalone Thesaurus to Integrated Related Words in The Danish Dictionary. I: Jaka Čibej, Vojko Gorjanc, Iztok Kosem & Simon Krek (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts. 916-923. <euralex.org/publications/from-standalone-thesaurus-to-integrated-related-words-in-the-danish-dictionary>.

Nimb, Sanni, Sussi Olsen, Bolette Pedersen & Thomas Troelsgård (2022): A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. I: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, Stelios Piperidis (eds.): *Proceedings of the Thirteenth Language Resources and Evaluation Conference: LREC2022*. Marseille: European Language Resources Association. 2826-2832. <aclanthology.org/2022.lrec-1.302>.

Petersson, Stellan & Emma Sköldberg (2020): To discriminate between discrimination and inclusion: a lexicographer's dilemma. I: Zoe Gavriilidou, Maria Mitsiaki & Asimakis Fliatouras (eds.): *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress*. Alexandroupolis: Democritus University of Thrace. 381-386. <euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_ProceedingsBook-Vol1.pdf>.

Plank, Barbara (2022): The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. I: Yoav Goldberg, Zornitsa Kozareva & Yue Zhang (eds.): *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. EMNLP 2022*. Abu Dhabi: Association for Computational Linguistics. <aclanthology.org/2022.emnlp-main.731.pdf>.

Scheuer, Jann (1995): Hans hustru og hendes bryster. I: Mette Kunøe & Erik Vive Larsen: *5. Møde om Udforskningen af Dansk Sprog. Aarhus Universitet 13.-14. oktober 1994* (MUDS 5). Aarhus: Aarhus Universitetsforlag. 246-256.

Trap-Jensen, Lars (2020): Inklusion eller mindretalsdiktatur? Om politisk korrekthed, minoritetshensyn og leksikografisk deskriptivisme i *Den Danske Ordbog*. I: *LexicoNordica* 27, 137-156.

Williams, Geoffrey (2003): From meaning to words and back: Corpus linguistics and specialised lexicography. I: ASp 39-40 Groupe d'étude et de recherche en anglais de spécialité. 91-106. doi:10.4000/asp.1320.

Sanni Nimb
seniorredaktør, ph.d.
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
sn@dsl.dk

Kvensk revitalisering, normering og leksikografi

Anna-Kaisa Räisänen, Aili Eriksen, Thomas Brevik Kjærstad & Trond Trosterud

An online dictionary and language technology for Kven language have been developed in a collaborative project between the Kven Institute and Giellatekno at UiT The Arctic University of Norway. Sources for the material are older texts and learning material that are available for the public. Several hundred hours of spoken material in old recordings, however, have still not been used. Utilising this material will be a priority in our future work.

1. Innledning

Denne artikkelen presenterer *Nettidigisanat Kvääni-norja-kvääni-nettisanakirja* (Eriksen et al. 2013–2024), ei elektronisk kvensk-norsk-kvensk ordbok, heretter referert til som KNKO. Den kvensk-norske og den norsk-kvenske delordboka refererer vi til som henholdsvis KNO og NKO. KNKO er ei ordbok i *Neahtta-digisánit*-familien av ordbøker (Johnson, Antonsen & Trosterud 2013). Artikkelen forklarer motivasjonen bak ordboka og drøfter rolla ordboka spiller i revitaliseringsarbeidet for kvensk språk. Videre drøfter den rolla ordboka spiller for normering, og i hvilken grad ordboka er i stand til å fylle denne rolla. Først kommer et bakgrunnskapittel som gir en oversikt over kvensk revitalisering og kvensk språknormering. Deretter drøftes motivasjonen bak det kvenske ordboksarbeidet og utviklinga av og statusen for ordbøkene. Ordbøkene har spilt ei viktig rolle både for stabilisering av kvensk rettskriving og for utviklinga av ordforrådet, og vi gjør greie for hvordan den har fylt denne rolla. Kapittel 4 tar for seg utfordringer i kvensk språkdokumentasjon og i normering av

språkvariasjon, og det drøfter tilgjengelige ressurser for det leksikografiske arbeidet. Kapittelet drøfter også framtidsplanene for de kvenske ordbøkene.

2. Bakgrunn

2.1. Situasjonen for kvensk språk

I likhet med meänkieli er kvensk et resultat av en to tusen år lang ekspansjon av østersjøfinsk, fra området sør for Finskebukta til vest for Torneelva og helt nord til Ishavet. Etter at Finland i 1809 gikk fra å være en integrert del i det svenske riket til å bli et storfyrstedømme i Russland blei språkspørsmålet en viktig del av finsk politikk. Det dype skillet mellom øst- og vestfinske dialekter blei planmessig oppheva i et kompromiss som førte til et nytt standardspråk. Reinhold von Beckers *Finsk grammatik* fra 1824 viste hvordan et kompromiss mellom bøyingsverket i de østlige og vestlige dialektene kunne se ut, og Elias Lönnrots ordbok *Suomalais-Ruotsalainen Sanakirja* fra 1880 la grunnlaget for et moderne finsk ordforråd (se Trosterud (1996) for ei drøfting av denne prosessen). Både kvensk og meänkieli (sistnevnte blir snakka vest for Torneelva fra Haparanda opp til Gällivare) sto utenfor hele denne prosessen, og den viktigste kilden til skriftlig finsk de kvensk- og meänkielispråklige talerne blei eksponert for, var Bibelen og det religiøse språket, dvs. den delen av finsk som hadde vært i bruk fra 1500-tallet og ikke var en del av standardiseringa på 1800-tallet. Utviklinga i det finske talemålet, *yleispuhekieli*, nådde heller ikke den kvenske befolkninga i særlig stor grad, og finske og kvenske dialekter har utvikla seg i ulike retninger. De kvenske dialektene har innen finsk dialektologi blitt klassifisert som nordfinske dialekter. Nordfinske dialekter har hatt langvarig kontakt med samisk og skandinaviske språk, og denne kontakten har vært spesielt sterk

for de kvenske dialektene. Kvensk var på begynnelsen av 1900-tallet det dominerende språket i de kvenske bygdene i Nord-Troms og Finnmark. I perioden fra 1900 fram til midten av 1900-tallet gikk antallet språkbrukere kraftig ned. Fornorskingstiltaka som blei satt i gang på slutten av 1800-tallet, spesielt gjennom skolen, endra språkbruken i de kvenske områda. Moderniseringa av samfunnet, krigen og evakueringa forsterka denne tendensen. Mange familier som fortsatt brukte kvensk før krigen, skifta språk hjemme i etterkrigsåra. De var redd for at flerspråklighet kunne være skadelig, eller at barn skulle bli stigmatisert hvis de var kvensktalende og hadde kvensk identitet. I løpet av 1950- og 1960-åra gikk de aller fleste kvener over til å snakke kun norsk med barna sine, noe som gjorde at kvensk i stor grad opphørte som hjemmespråk (Lane 2010; Räisänen 2014).

Bevisstheten om kvensk kultur og tilhørighet fikk et oppsving igjen på 1970-tallet, og de første kvenske organisasjonene blei danna på 1980-tallet. Samtidig etablerte kvenene et samarbeid med tornedalingene i Nord-Sverige. Mange kvener har røtter i Tornedalen og et sterk slektskap med tornedalingene. Dette samarbeidet leda til ei utvikling som skulle sørge for at kvenene kunne ta vare på de lokale dialektene sine i stedet for å «forfinske» språket sitt ved å bruke det standardiserte finske språket. Et av hovedmåla til Norske kveners forbund blei på 1980-tallet å få kvensk anerkjent som et eget språk og å styrke språkets status i samfunnet (Sundelin 2008). Kongen i statsråd vedtok i 2005 at kvensk skulle anerkjennes som et eget språk i Norge. Kvensk er et av tre offisielle minoritetsspråk i Norge og har vern etter språkloven og den europeiske pakten om regions- og minoritetsspråk del II.

En undersøkelse drøfta i Rasmussen (2004–2005) viser at det på begynnelsen av 2000-tallet var 5640 personer som var i stand til å forstå kvensk i Norge. En stor del av disse, rundt en tredel, snakka også samisk og bodde i kommunene Kautokeino, Karasjok og Tana. En viktig grunn til dette er at den samiskspråklige befolk-

ninga har tett kontakt over landegrensene og med befolkninga som snakker nordfinske dialekter. En annen grunn er at flerspråkligheten har blitt bedre bevart i trespråklige områder enn i tospråklige. I dag tilhører de fleste kvensktalende den eldste generasjonen, og kvensk brukes aktivt muntlig kun noen få steder i Nord-Norge, hovedsakelig i Porsanger og Varanger. Kvensk er lite synlig innafor alle samfunnsområder og -institusjoner, og kan derfor defineres som et kritisk trua språk (Laakso et al. 2016:134–137).

2.2. Kvensk revitalisering

Kvenske lærere og språkaktivister begynte å utvikle kvenske læremidler på 1980-tallet. Da de begynte å undervise i finsk, var kvensk fremdeles et levende muntlig språk blant den eldste generasjonen i mange kvenske bygder. Noen hadde også lært å lese og skrive på finsk. Lærere i Porsanger fikk tilbakemelding fra kvensktalende om at de ønska å videreføre nettopp den lokale språkforma og ville at barna skulle lære sin lokale varietet og ikke finsk. Derfor blei det første kvenske skriftspråket utvikla etter lærernes og elevenes behov (Huss 1999). Mange kvener ville snu språkskiftet fra kvensk til norsk og revitalisere de lokale kvenske varietetene. Målet blei «å oppnå funksjonell tospråklighet (norsk – kvensk), også blant yngre generasjoner i de kvenske kjerneområdene», som det sto i revitaliseringsplanen i 2004 (Norske Kveners Forbund & Norsk senter for kvænsk språk og kultur 2004).

Språkundervisning og kulturprosjekter for barn og unge var en sentral del av denne revitaliseringsplanen, og utdanningsinstitusjoner og kvenske fagmiljøer spilte i tida etterpå ei essensiell rolle i formidling og videreføring av språket til kvenene som ikke hadde lært det hjemme. Det har vært viktig å skape nye språkdomener som gir nye språkbrukere muligheter til å snakke og skrive språket. Samtidig blei utvikling av ulike typer læremateriell en forutsetning for å kunne ha undervisning i kvensk. Det var like-

vel ingen ordboksprosjekter inkludert i planene på dette tidspunktet (Kainun institutti – Kvensk institutt 2007).

Etter at kvensk fikk status som et nasjonalt minoritetsspråk i 2005, begynte arbeidet med å normere språket. Den nye statusen førte til at mange kvenske institusjoner blei etablert. Kvensk institutt fikk nasjonalt ansvar for kvensk språk og kultur. Halti kvenkultursenter blei oppretta for å jobbe regionalt i Nord-Troms, og etter hvert har fem kvenske språksentre i Alta, Kvænangen, Porsanger, Storfjord og Vadsø kommet til for å styrke språket. Kvensk institutt har også administrert normeringsorgana for kvensk. Alle de kvenske språksentera har jobba med kvensk undervisning, men de har også jobba systematisk for å synliggjøre språket i offentligheta. En viktig del av dette har vært å oversette tekster fra norsk til kvensk for ulike offentlige institusjoner og frivillige organisasjoner.

Nye språkkurs og økt tekstproduksjon på kvensk ga et større behov for ordbøker. I denne tidlige fasen av revitaliseringa var det først viktig å skape et felles normert skriftspråk. Kvenske institusjoner satsa derfor i starten mer på språknormering og tekstproduksjon enn på ordboksarbeid. Språk- og kulturprosjekter for barn og ungdom blei også høgt prioritert.

2.3. Normering av kvensk skriftspråk

I kjølvannet av at kvensk blei anerkjent som et eget språk, begynte også et mer systematisk organisert arbeid med det kvenske skriftspråket. Kvensk institutt blei stifta i 2006 og har siden hatt ansvar for å drive språknormeringsarbeid gjennom to språkorgan: Kväänin kieliraati (Kvensk språkråd) og Kväänin kielitinka (Kvensk språkting). Kvensk språkråd var fra 2007 til 2010 et fagorgan som forberedte saker om ulike strategier for normering. Kvensk språkting har siden etableringa bestått av medlemmer som representerer ulike kvenske dialektområder. Språktinget er vedtaksorgan i normeringsspørsmål for kvensk (Larsen, Paavalniemi &

Söderholm 2012). Etter at Kvensk språkråd blei nedlagt i 2010, har Kvensk institutt hatt ansvar for saksbehandlinga for språktinget.

Da de kvenske språknormeringsorgana begynte sitt arbeid i 2007, måtte de ta stilling til ortografiske prinsipp og språklig variasjon. Språktingets mandat var å utvikle et standardisert kvensk språk, men de bestemte seg først for å utvikle en læreboknormal til undervisning i kvensk (Keränen 2018:9; Söderholm 2017:29). Kvensk språkting vedtok at kvensk språk skulle ha de samme ortografiske prinsippa som finsk. Dette var begrunna med likhetene i kvensk og finsk fonologi. Disse ortografiske prinsippa var også brukt av kvensklærer og språkaktivist Terje Aronsen og forfatter Alf Nilsen-Børsskog. Kvenske stedsnavn var også registrert med samme ortografi. For å kunne ta stilling til variasjonen i de kvenske dialektene forberedte Kvensk språkråd ulike strategier der de tok hensyn til både dialektene, meänkieli og det finske skriftspråket. Kvensk språkting vedtok at det kvenske skriftspråket skulle bygges på et kompromiss mellom de ulike kvenske dialektene og samtidig ligge så nær meänkieli som mulig (Kvensk språkting 2008). Da språktinget seinere skulle drøfte Söderholms normative grammatikker (Söderholm 2014, 2017), kom spørsmålet om hvordan variasjonen i kvensk skulle beskrives i grammatikken, opp på nytt. Söderholm beskriver variasjonen mellom de tre sentrale kvenske dialektområda, nemlig Porsanger, elvedalsområda og Varanger (Söderholm 2014:24, 2017:30).

I og med at *Kvensk grammatikk* (Söderholm 2017) bygger på Språktingets vedtak, er den også den viktigste beskrivelsen av normert kvensk. Selv om grammatikken beskriver variasjonen i kvensk språk, blir porsangervarieteten presentert som hovedform, mens de andre varietetene blir presentert som sideformer. Söderholm (2017:31) begrunner dette valget med at porsangervarieteten var den første som blei analysert i arbeidet med grammatikken. De fleste eksempelsetningene i grammatikken er fra Porsanger, og andre varieteter er i mindre grad synlige (Söderholm 2017:30).

Seinere har Isaksen (2018) arbeida med en kvensk skolegrammatikk basert på porsangervarieteten. Kvensk språkting gikk inn for at skolegrammatikken skulle utgis også på de to andre varietetene, men på grunn av manglende ressurser blei det ikke gjort.

Da den første kvenske romanen, *Kuosuvaaran takana* (av Alf Nilsen-Børsskog), blei gitt ut i 2004, fantes det kun noen få tekster på kvensk fra før av. Terje Aronsen (1984, 1986) hadde skrevet læremateriell og noen faglitterære tekster på porsangervarieteten, og Olav Beddari (1987) ga ut en lesebok med fortellinger på sin egen varietet fra Sør-Varanger. Da Eira Söderholm (2009) og Agnes Eriksen (2014–2017) begynte å utvikle læremateriell, brukte de kun porsangervarieteten. Også forfattere som skriver på kvensk, har stort sett brukt porsangervarieteten i sine utgivelser, for eksempel Eriksen (2011) og Räisänen (2020), og de kvenske mediene *Ruijan Kaiku* og *NRK Kvääni* bruker også porsangervarieteten. Etter Språktingets vedtak om å tillate variasjon i kvenske dialekter laga Eira Söderholm parallelle versjoner på tre ulike kvenske varieteter av kursmateriellet som var utvikla for studenter ved UiT Norges arktiske universitet (heretter UiT). De tre varietetene var porsanger-, elvedals- og varangervarieteten.

Til tross for vedtaka som blei gjort om variasjon i skriftspråket (jf. Larsen, Paavalniemi & Söderholm 2012), har porsangervarieteten blitt dominerende i nesten alle kvenske tekster. For eksempel lanserte UiT sitt nye læremateriell *Meiđän joukko* (UiT 2019) kun på porsangervarieteten. *Meiđän joukko* har ikke blitt utgitt med parallelle tekster på de andre varietetene på lik linje med Söderholms *Aikamatka* (2009). Dette kan igjen ha ført til at porsangervarieteten har blitt dominerende i bruk, i og med at UiT er den eneste høgere utdanningsinstitusjonen som underviser i kvensk. Kvensk institutt har også oversatt mest fra norsk til porsangervarieteten og i mindre grad brukt varanger- og elvedalsvarietetene. Disse to varietetene har blitt brukt i tekster som har omhandla områda der varietetene har vært talespråk. For eksempel har Halti

kvenkultursenter publisert noen oversettelser og andre tekster på elvedalsvarieteten. De som har bestilt oversettelser fra Kvensk institutt, ønsker iblant å få oversettelsen på den varieteten som ligger nærmest deres lokale kvenske språklige identitet.

3. Kvensk leksikografi

Ordbøker er sentrale virkemidler i språkdokumentasjon, men de er også viktige for å bevare språk. Leksikografisk arbeid har derfor flere funksjoner i en minoritetsspråkkontekst (Ogilvie 2011:392). Ikke minst er det å utarbeide ordbøker for trua språk ofte et avgjørende steg i prosessen med å hjelpe nye språkbrukere til å ta tilbake sine forfedres språk. Ordbøker formidler grunnleggende kunnskap om språk og er av stor betydning for undervisning i trua og lite brukte språk. Dette ser vi også i forbindelse med revitaliseringa av kvensk.

Initiativet til å utvikle ei ordbok for kvensk kom fra styret ved Kvensk institutt i 2010 da de tok opp behovet for ei slik ordbok med Kvensk språkråd (Kvensk språkråd 2010). Da Kvensk språkråd behandla dette initiativet, påpekte de at kvenske dialekter var godt dokumentert av finske språkforskere. Språkrådet mente at det omfattende materialet (over 400 timer lydopptak) som var arkivert i *Suomen kielen nauhoitearkisto – Finska bandarkivet* i Helsingfors, kunne danne et godt grunnlag for ei kvensk-norsk ordbok. De anbefalte at dette materialet kunne brukes sammen med Alf Nilsen-Børsskogs romaner og Terje Aronsens ordlister som en begynnelse på ordboksarbeidet. Kvensk institutt fikk ansvaret for å planlegge arbeidet med ordboka og etablerte et samarbeid med Giellatekno (Senter for samisk språkteknologi ved UiT) og Halti kvenkultursenter. Disse samarbeidspartnerne satte i gang et prosjekt for kvensk språkteknologi i 2013 for å øke bruken av det nyetablerte skriftspråket og bidra til kvensk revitalisering (Haavisto et al. 2014:183).

Prosjektet resulterte i opprettelsen av KNKO, som Kvensk institutt og Giellatekno har videreutvikla. Parallelt med dette arbeidet satte Kvensk institutt og Giellatekno i gang et tett samarbeid med kvenske fagmiljøer ved UiT for å utvikle en morfologisk analysator, stavekontroll og korpus for kvensk språk (Trosterud et al. 2017). Både den morfologiske analysatoren og mulighet for direkte lemmasøk i det kvenske korpuset KORP er innlemma i ordboka. Den tekniske infrastrukturen i prosjektet er utvikla av Giellatekno og Divvun[1] ved UiT. Det kvenske språkteknologiprosjektet hadde finansiering fra Kommunal- og moderniseringsdepartementet for til sammen to millioner kroner i perioden 2014–2017.

Trosterud (2019) viser at 62,4 % av oppslaga som blei gjort i KNKO i 2018, var norske lemma, og at bare 26,5 % av de kvenske oppslaga i KNO var oblike former (dvs. andre former enn lemma). Artikkelen konkluderer derfor med at KNKO ikke primært brukes til resepsjon av minoritetsspråktekst, men som pedagogisk ordbok og til en viss grad som produksjonsordbok. Dette er interessant ettersom det i en annen kontekst i Trosterud et al. (2017) kommer fram at den morfologiske analysatoren allerede i 2016 kunne ha vært til hjelp for dem som har ønska å lese tekster på kvensk. I Trosterud et al. (2017) presenteres også den kvenske grammatiske språkmodellen som ordboka er basert på.

3.1. Behovet for ei nettordbok for kvensk

Da den første versjonen av KNKO blei publisert i 2013, var et av måla at folk skulle kunne bruke ordboka for å lese tekster på kvensk. Alf Nilsen-Børsskog hadde allerede gitt ut flere romaner på kvensk, og den kvenske avisa Ruijan Kaiku hadde en plan om å produsere 20 % av sine tekster på kvensk. Det blei antatt at norsktalende kvener ville lese disse tekstene på kvensk, og at de hadde

1 Divvun er ei forskningsgruppe som utvikler samisk språkteknologi og praktiske hjelpemidler basert på denne språkteknologien i nært samarbeid med Giellatekno.

behov for å bruke KNO til dette. Kvensk hadde på det tidspunktet begynt å bli mer synlig i offentligheten i Norge, og etterspørselen etter ulike oversettelser økte. Noen kommuner hvor kvener historisk sett hadde vært en stor andel av befolkninga, ønska å gjøre kvensk språk og kvener mer synlige. Kvensk skriftspråk blei tatt i bruk i mange nye språkdomener, som for eksempel utstillingstekster og administrative tekster. Mange kvener ønska også å skrive vanlige hilsninger på kvensk, og det blei etterspurt tematiske ordbøker og fraseordbøker. Derfor økte spesielt behovet for å videreutvikle NKO for språkbrukere som ønska å oversette egne tekster og fraser fra norsk til kvensk.

Ordbøker kan som nevnt spille ei viktig rolle for språkdokumentasjon og språkrevitalisering. Ord og bruken av dem kan videreformidles på en effektiv måte med ordbøker. Skal det være mulig å skrive på et ungt skriftspråk, kreves det kontinuerlig utvikling av ordforråd. Derfor har det blitt prioritert at orda som Kvensk institutt har utvikla og samla inn i samarbeid med Kvensk språkting og lokale kvenforeninger, har blitt kontinuerlig innlemma i KNKO. Mens ansatte på Kvensk institutt har jobba med tekster av ulike slag, blant anna administrative tekster, nettsider, barnesanger, eventyr og religiøse tekster, har Kvensk språkting utvikla terminologi for offentlig skilting og organisasjonsnavn. I tillegg til at ordboka er med på å dokumentere språket, har det kvenske korpuset KORP blitt et viktig redskap for dokumentasjon. I ordboka finner språkbrukeren bare et svært begrensa utvalg av ordbetydninger og språklige kontekster. I korpuset finner man derimot hele setninger og tekster, og dermed langt flere bruksmåter for orda.

I forbindelse med trua språk er det særlig viktig å utvide ordbøker med nytt ordforråd som en del av revitaliseringa og normeringa. KNKO har fylt behovet for å dokumentere utviklinga av kvensk ordforråd. De viktigste brukergruppene for ordbøkene er studenter og oversettere, to grupper som har ulike behov. En som har kunnskap om og språkferdigheter i kvensk, trenger ikke

samme informasjon om valg av ord og bøyningsmønster som nybegynnere. Samtidig som ordboka gir den grunnleggende kunnskapen språkbrukerne etterspør, ser vi også at brukerne ikke finner all informasjonen de leter etter. Tilbakemeldinger som ordboksutviklerne stadig får, handler for det meste om ord som mangler i ordboka. Det etterspørres også informasjon om uttale av orda. I og med at KNO gir grunnleggende informasjon om hvordan kvenske ord skrives, fungerer den som bindeledd mellom Kvensk språkting og språkbrukerne. De som jobber med revitalisering og tekstproduksjon i kvenske institusjoner, har gitt positive tilbakemeldinger om at NKO fungerer sammen med KORP. Korpuset, ordboka og den morfologiske analysatoren letter oversettelsesarbeidet for kvenske institusjoner betraktelig.

3.2. Data og kilder for ordboksutvikling

Det er forska relativt lite på oppbygginga av og strukturen til det kvenske ordforrådet. Den grundigste analysen av det kvenske ordforrådet hadde fokus på terminologien for fisk og sjødyr (Andreassen 2003). I tillegg fins det to masteroppgaver om norske lånord i kvensk (Utvik 1996; Leskinen 2017) og en analyse av forskjeller i ordforrådet mellom kvensk og finsk (Eriksen 2018). Mange forskere har påpekt at de største forskjellene mellom kvensk og finsk ligger nettopp i ordforrådet (Andreassen 2003:18; Eskeland, Lindgren & Norman 2003). Vilkåra for kvensk terminologisk arbeid uavhengig av finsk blei drøfta i Trosterud & Skanke (2012). Der blei det pekt på at selv uten at det hadde blitt etablert en kvensk språkstandard uavhengig av finsk, ville det vært nødvendig å etablere en egen terminologi på mange fagområder. Selv om kvenske institusjoner har kommet godt i gang med å normere det kvenske språket og utvikle ny terminologi og nytt allmennspråklig ordforråd, er det behov for forsking både på det kvenske leksikonet og på mekanismene for å utvide det.

Ettersom det er gjort såpass lite forskning, har det vært viktig å bruke ordlistene som de kvensktalende har laga, som grunnlag for KNO. Ordboka blei først utarbeida med grunnlag i ei kvensk-norsk ordbok som var utvikla av Terje Aronsen (2010). Aronsen hadde opprinnelig utvikla ei ordliste til kvenskundervisninga, men han hadde også samla inn en del kvenske ord og utvikla ordforråd i forbindelse med oversettelsesarbeid ved Kvensk institutt. I arbeidet sitt brukte Aronsen kun porsangervarieteten, og dermed blei dette rådende for utviklinga av ordboka. Utviklinga av kvenske språkteknologiske verktøy tok derfor utgangspunkt i denne varieteten. Den var mest brukt i alle skriftlige kilder og danna grunnlag for Söderholms første kvenske grammatikk, *Kainun kielen grammatikki* (Söderholm 2014), men ordboksprosjektets uttalte mål var å inkludere elvedals- og varangervarietetene i videreutviklinga av den morfologiske analysatoren og KNO fra starten av (Haavisto et al. 2014:180–181). Arbeidet med å ta med alle de kvenske varietetene starta i 2015.

Antall oppslagsord i KNKO i mars 2024 var for KNO 15 649 og for NKO 14 304, ei stor utviding sammenligna med de 3600 oppslagsorda i Aronsen (2010). Det nye ordforrådet har kommet fra UiTs kvenske undervisningsmateriell *Aikamatka* (Söderholm 2009) og *Meiđän joukko* (UiT 2019), fra Agnes Eriksens *Kainuruija-kainu koulusanakirja* (2018) samt fra fraseordbøker (Kvääninuoret 2023; Larsen 2015) og ordlister som kvenske språkbrukere selv har laga. En god del av orda har kommet via Kvensk institutts arbeid, både i form av nyord fra oversettelsesarbeid og fra ordinnsamlingsarbeid som Kvensk institutt og Kvensk språkting har gjort i ulike kvenske dialektområder. Facebookgruppa *Kveenin sanat*, der kvenske språkbrukere diskuterer ord og betydninger, har også bidratt til innsamling av kvensk ordforråd til ordboka.

En del av det kvenske språkteknologiprosjektet har vært å dokumentere den skriftlige bruken av kvensk språk. Kvensk institutt har i samarbeid med Giellatekno og Divvun samla inn alle kven-

ske tekster som de har fått tillatelse til, i KORP. Korpuset er ikke stort (ca. 500 000 ordformer), men det har likevel blitt den mest sentrale kilden til kvensk leksikografisk arbeid. Korpuset inneholder oversettelser fra Kvensk institutt, men det har også noen medietekster og faglitterære tekster fra andre skribenter. Korpuset blei dessuten nylig utvida med romansamlinga til Alf Nilsen-Børsskog. Dette utgjør over 600 000 ord med skjønnlitterær tekst, og den totale størrelsen på korpuset er dermed over 1 115 000 ord per juli 2024. Kvensk institutt har nylig fått i stand en avtale med forlaga som publiserer på kvensk, om å kunne innlemme litteratur som blir skrevet på kvensk, i korpuset. Disse nye tekstene vil bli en viktig ressurs i ordboksarbeidet framover.

3.3. Status for ordbøkene

Trosterud (2019:189) dokumenterte i hvilken grad KNKO dekker det sentrale ordforrådet i kvensk og norsk. Dekningsgraden blei for kvensk målt etter KORP, mens den for norsk blei målt etter NoWaC (Norwegian Web as a Corpus), et korpus utvikla ved Universitetet i Oslo. Målinga blei gjentatt i 2024, og utviklinga i dekningsgrad vises i tabell 1.

Ordbok	**KNO**		**NKO**	
De ... vanligste	1000	10 000	1000	10 000
2019	79,9 %	28,8 %	66,0 %	32,2 %
2024	95,0 %	88,9 %	88,3 %	54,0 %

Tabell 1: Dekningsgrad for de 1000 og 10 000 vanligste orda i kvensk og norsk i KNO og NKO, i 2019 (Trosterud 2019) og 2024.

Som det går fram av tabellen, har det særlig for KNO vært ei stor forbedring, men også de vanligste norske orda er bedre representert i ordboka enn før. Når det gjelder sjeldnere ord, er dekningsgraden naturlig nok dårligere. Av de norske orda i frekvensspen-

net 20 000–30 000 i NoWaC-korpuset kjenner NKO for eksempel igjen bare 13,9 %.

Trosterud (2019:119ff.) inneholdt også ei sammenligning av ordbøkene fra norsk til henholdsvis nordsamisk, sørsamisk og kvensk. De 10 lemmaene som blei slått opp flest ganger i den norsk-nordsamiske ordboka, var sentrale verb, og de fem neste var funksjonsord (lista i frekvensrekkefølge): *skulle*, *ha*, *kunne*, *gå*, *se*, *bli*, *komme*, *få*, *dra*, *være*, *for*, *det*, *snakke*, *som*, *på*. Ei sannsynlig forklaring er at den sentrale bruken av ordboka var som produksjonsordbok for nordsamisk. I kontrast til dette utgjorde de 15 lemmaene med flest oppslag i NKO ei mer heterogen gruppe: *vel*, *jeg*, *hus*, *være*, *hei*, *gå*, *kaste*, *bli*, *kunne*, *du*, *se*, *eie*, *følge etter*, *ha*, *skulle*. Her er det også mange sentrale verb, men inntrykket er at produksjonsaspektet ikke var like dominerende. For første halvpart av 2024 har lista over de 15 lemmaene med flest oppslag i NKO endra seg noe. De er nå: *komme*, *kunne*, *være*, *gå*, *se*, *hus*, *hei*, *møte*, *snakke*, *til*, *jeg*, *få*, *som*, *måtte*, *skulle*. Det er fortsatt ett eller to substantiv med på lista, men sammenligna med den eldre lista er det flere sentrale verb, og det ser med andre ord ut som om bruken av ordboka som produksjonsordbok for kvensk er på vei opp. Merk også at i tillegg til pronomenet *jeg* inneholdt loggen like mange søk på *jeg* + verb (de vanligste var *jeg heter*, *jeg elsker deg*, *jeg er* og *jeg har*). Slik ordboka er strukturert nå, vil slike fraser ikke gi tilslag, men teknisk sett er det ingen ting i veien for å la den kvenske morfologiske generatoren generere relevante kvenske verbformer på grunnlag av norsk *pronomen + verb*.

Materialet i NKO er som nevnt delvis henta fra oversettelser av tekster fra offentlig sektor. Ordforrådet i slike tekster er likevel ikke nødvendigvis bedre representert enn ordforrådet i mer allmennspråklige tekster. Ordforrådet i (en oversettelse til bokmål av) oppsummeringskapittelet i St.meld. nr. 15 (2000–2001) *Nasjonale minoritetar i Noreg* er for eksempel ikke bedre representert enn ordforrådet i Ketil Melhus' novelle «Et lite stykke tid»

(de to tekstene har henholdsvis 72,1 og 75,6 % dekningsgrad i NKO).

4. Kvensk ordboksarbeid

4.1. Utfordringer og skeivheter

Kvensk språk er i samme situasjon som mange andre trua minoritetsspråk med et ungt skriftspråk og en ung litterær tradisjon. Datagrunnlaget for den kvenske ordboksutviklinga er helt annerledes enn datagrunnlaget for utvikling av ordbøker for språk med en lang og godt dokumentert skriftlig tradisjon.

Det er få som oversetter og skriver på kvensk, og dermed få som påvirker utviklinga av det nye ordforrådet. Hittil har korpuset langt på vei bestått av oversatte tekster. Selv om vi nå har lagt til en stor mengde originale tekster, utgjør disse fortsatt bare rundt halvparten av tekstene i korpuset. Korpustekstene har en direkte påvirkning på ordforrådet som blir lagt inn i ordboka, og er derfor viktige.

De fleste tekstene som Kvensk institutt og Halti kvenkultursenter oversetter, har vært bestilt av museer, lærebokforfattere og offentlig administrasjon, og inneholder ikke nødvendigvis det ordforrådet som trengs for å oppfylle behova til det kvenske språksamfunnet. Derfor er det mange kvener som vil bruke språket skriftlig, som tar kontakt med Kvensk institutt med ønske om at det utvikles nytt ordforråd om spesifikke tema som interesserer dem. Det kommer hele tida slike forespørsler om å utvikle ordforråd innafor et visst område. Antall språkbrukere som skriver på kvensk, påvirker også utviklinga av ordforrådet videre framover. Generasjonen som fortsatt lærte kvensk hjemme, fikk ikke mulighet til å lære kvensk på skolen – de blei til og med hindra i å bruke kvensk. Derfor har ikke de fleste kvenske morsmålstalere lært å

skrive verken på kvensk eller på finsk. I stedet har de hatt all sin skriftlige opplæring i og på norsk. Utviklinga av kvensk ordbok står dermed i en paradoksal situasjon. På den ene sida må det tas hensyn til krava til synliggjøring av kvensk språk i storsamfunnet, som er svært ressurskrevende. På den andre sida må behova til det kvenske språksamfunnet ivaretas. Det er en balansegang som vi prøver å gjennomføre til tross for små ressurser.

Kvenske dialekter er relativt godt dokumentert, men denne dokumentasjonen er lite systematisert, forska på og brukt i utvikling av ordbøker (Niiranen 2022; Lane et al. 2022). Det innsamla dialektmaterialet er dessuten bare i liten grad tilgjengelig for språkbrukerne. Kvensk institutt får av og til tematiske ordlister fra morsmålsbrukere som ikke nødvendigvis kjenner til kvensk ortografi eller hvordan orda brukes i kontekst. Dette viser at språket brukes i liten grad i det kvenske samfunnet sammenligna med situasjonen som er dokumentert i det eldre materialet beskrevet av Niiranen (2022). I ordboksutviklinga er det derfor vanskelig å vurdere hvor etablerte visse ord og betydninger er i det kvenske språksamfunnet sammenligna med hos individuelle språkbrukere, og om orda som er i ordlistene fra språkbrukerne, er deres egne forslag til nye ord.

I utviklinga av nytt ordforråd må det alltid vurderes hvordan og hvor mye man skal låne direkte fra andre språk, og i hvor stor grad man skal konstruere nye ord med utgangspunkt i kvensk. Ettersom kvensk er nær beslektet med meänkieli og finsk og har vært i nær kontakt med samisk og norsk, er det også et spørsmål om hvor mye man bør låne fra disse språkene framover. Norsk har vært et naturlig språk å låne fra fordi kvener befinner seg i det norske samfunnet. I kvensk fins det allerede en del skandinaviske låneord som er tilpassa strukturen i kvensk. Mange språkbrukere oppfatter dem likevel ikke som låneord, fordi de ikke er like gjennomsiktige som moderne norske låneord. Det finske skriftspråket kunne også vært et aktuelt språk å låne fra, fordi det allerede

har velutvikla terminologi innen politikk, kultur og samfunnsliv. Mange kvenske språkbrukere oppfatter imidlertid moderne finske ord som fremmede og synes det er mer naturlig å låne kjent terminologi fra norsk, selv om det også er kvener som finner det naturlig å låne fra finsk. En av grunnene til å ha et kvensk skriftspråk atskilt fra finsk er jo at kvenene skal kunne skrive på sitt eget språk. Meänkieli har lenge vært i samme situasjon som kvensk og har derfor vært relevant å sammenligne med. Språka har ei felles historie og samme kulturhistoriske kontekst med mye felles ordforråd som kan være til nytte i normeringa av kvensk ordforråd (Söderholm 2006:47–50). Når det mangler termer på kvensk, kan man også konstruere nye ord ut ifra kvenske ord som allerede eksisterer.

I tillegg til hvordan utviklinga av nytt ordforråd i KNKO skal dokumenteres, har det også vært et spørsmål om hvordan de tre varietetene i kvensk skal synliggjøres. Porsangervarieteten har som nevnt hittil vært prioritert, men språkbrukere som kan andre varieteter, kan også søke opp ord i ordboka på sin egen varietet med hjelp av den morfologiske analysatoren. Situasjonen speiler den språklige variasjonen i det kvenske korpuset, så hvis tekstproduksjonen innafor de andre språkvarietetene øker, blir det også større behov for å inkludere disse varietetene i ordboka. Kvensk institutt har lenge fått tilbakemeldinger fra språkbrukere om at ordboka bør synliggjøre alle varietetene. Å synliggjøre de lokale dialektene er viktig for mange kvensktalende, og mange opplever det som viktig at markører for tilhørighet til slekt og spesifikke kvenske tradisjonelle områder, og dermed egen språklig identitet, blir synlig også i ordboka.

4.2. Planer framover og manglende ressurser

Mangel på personale som kan arbeide med både leksikografi og språkteknologi, har vært ei stor utfordring for utviklinga av kven-

ske ordbøker. Det er i utgangspunktet vanskelig å rekruttere folk som har kunnskap i kvensk, og som i tillegg har kompetansen som kreves for å kunne jobbe med leksikografi og den tekniske infrastrukturen. Med tanke på de ressursene vi har hatt tilgjengelig til utvikling av ordboka, analysatoren og korpuset, så har arbeidet hatt god framgang. Bruken av ordbøkene har økt kraftig. Mens det i 2014 blei gjort i underkant av 1000 oppslag i KNKO, var antallet oppslag over 70 000 i 2018 (Trosterud 2019). I 2021 var tallet økt til over 156 000, og i 2023 var det nesten 250 000.

Disse talla forteller også indirekte om hvor mange som er interessert i å lære og bruke kvensk i dag (jf. også Lanes 2023). Stadig flere ønsker å ta kvensk språk tilbake. Antall studenter i kvensk språk har økt på UiT, og flere velger også kvensk som andrespråk i skolen (i Troms og Finnmark). Mange voksne ønsker også å lære kvensk, og det har vært en stor økning i deltakere på nybegynnerkursa de siste åra. I denne situasjonen er det enda viktigere å kunne oppfylle kvenske miljøers behov for språkrevitalisering med videreutvikling av KNKO. De tospråklige nettordbøkene, den morfologiske analysatoren og tekstkorpuset gjør det kvenske språket lettere tilgjengelig for dem som ønsker å ta tilbake språket.

I de neste åra skal det legges mer vekt på å jobbe parallelt med å systematisere oppbygginga av korpuset og å supplere det med eksisterende kvenske tekster som hittil mangler. Det kvenske korpuset er som nevnt nå utvida med Alf Nilsen-Børsskogs romaner, og framover vil dette materialet bli brukt til å dokumentere og analysere Nilsen-Børsskogs ordforråd, som så vil innlemmes i ordboka. En annen viktig kilde som skal brukes i utviklinga av ordboka framover, er lydopptaka i *Suomen kielen nauhoitearkisto – Finska bandarkivet* samt i andre arkiver (Niiranen 2022:186–187). I bandarkivet er det 419 timer med kvenske dialektopptak som dekker området fra Storfjord i vest til Sør-Varanger i øst.

5. Konklusjon

Som vi har vist, har kvensk ordboksutvikling vært tett knytta til normering, revitalisering og undervisning i kvensk språk. Seinere har også behova for oversetting av tekster vært styrende for ordboksutviklinga. Dette gjenspeiles også i ordforrådet i ordboka. Det at ordboka har vært knytta til den kvenske grammatiske språkmodellen, har gitt den ei normativ rolle. Målet med normeringsarbeidet har vært å styrke lokal språklig variasjon og den lokale språklige identiteten, men i praksis har porsangervarieteten så langt blitt prioritert.

I og med at språket står i ei så svak stilling, må den leksikografiske metoden tilpasses en minoritetsspråklig situasjon. Det mangler både tekst og skribenter som kunne skrive mer tekst og utvikle ordforrådet, samt leksikografer til å analysere tekstene. Som vist i del 3.2 fins det språklig dokumentasjon fra tida da kvensk var i aktiv bruk i kontinuerlige språksamfunn. Denne dokumentasjonen sammen med flere tekster i ulike sjangere skal i tas i bruk for å forbedre KNKO framover.

Status for KNO i dag er at 95,0 % av de 1000 vanligste og 88,9 % av de 10 000 vanligste orda er oppført i ordboka. Status for NKO er at 88,3 % av de 1000 vanligste og 54,0 % av de 10 000 vanligste orda er representert. Særlig for NKO er det ei forbedring sammenligna med funna i Trosterud (2019). For et større ordforråd er dekningsgraden likevel dårlig, naturlig nok for ei ordbok som omfatter 14 304 lemma. Rett nok vil prosentandelen være langt bedre i webgrensesnittet, der den morfologiske analysatoren deler opp norske sammensetninger og ordboka gir oversetting av enkeltorda. Ordbøkene er likevel langt unna å kunne representere ordforrådet i et språk til bruk i et moderne samfunn.

KNKO har blitt flittigere brukt de siste åra, etter hvert som antall kvenske språkkurs og kvenskstudenter har økt. Digitale språkressurser er uunnværlige for kvenske språkmiljøer som øn-

sker å ta tilbake språket sitt. Ordboka, korpuset og den morfologiske analysatoren er nødvendige også som språkdokumentasjon. Sammen gir de kunnskap om det kvenske språket og gjør det tilgjengelig for språkbrukere i en situasjon hvor språket ellers er trua. Derfor er det viktig å se KNKO, kvensk leksikografi og språkteknologi i sammenheng med revitaliseringa.

Litteratur

Ordbøker

Aronsen, Terje (2010): *Kvensk-norsk elektronisk ordbok*. Redigert av Verena Schall & Trond Trosterud. Universitetet i Tromsø.

Eriksen, Agnes (2018): *Kainu-ruija-kainu koulusanakirja = Kvensk-norsk-kvensk skoleordbok.* Lakselv: Porsanger kommune.

Eriksen, Aili, Mervi Haavisto, Mari Keränen, Tobias Kvalness, Thomas Kjærstad, Tove Reibo, Anna-Kaisa Räisänen, Verena Schall, Sindre Reino Trosterud & Trond Trosterud (2013–2024): *Nettidigisanat Kvääni-norja-kvääni-nettisanakirja*. Tromsø: UiT Norges arktiske universitet. <sanat.oahpa.no> (juli 2024).

KNKO = Eriksen et al. (2013–2024).

Kvääninuoret (2023): *Fraasikirja – Frasebok: Norsk-kvensk frasebok for ungdom*. Tromsø: Ruija forlag.

Larsen, Karin (2015): *Norsk-kvensk fraseordbok for pleie og omsorg*. Børselv: Kainun institutti – Kvensk institutt.

Lönnrot, Elias (1880): *Suomalais-Ruotsalainen Sanakirja – Finskt-Svenskt Lexikon I-II.* Helsinki: WSOY.

Neahttadigisánit (2013–2024) = *Ordboksportalen Neahttadigisánit*. <sanit.oahpa.no/more/> (juli 2024).

Pyykkö, Vappu (2008): *Sana-aitta. Ordliste til "Kuosuvaaran takana" og "Aittiruto" av Alf Nilsen-Børsskog. Utvalgt og bearbeidet av Vappu Inkeri Pyykkö*. Digital versjon: <www.kvenskinstitutt.no/wp-content/uploads/2020/05/Sana-aitta.pdf > (juli 2024).

Annen litteratur

Andreassen, Irene (2003): *Tainariksi kuttuthaan se steimpiitti täälä: en studie av kvenske fiske- og sjødyrnavn i Varanger, Porsanger og Alta*. Avhandling (dr.art.). Universitetet i Tromsø.

Aronsen, Terje (1984): *Meiđän kielelä 1: Opetusmateriaalii Pyssyjoven tialektilä*. Upublisert læremateriell.

Aronsen, Terje (1986): *Meiđän kielelä 2: Opetusmateriaalii Pyssyjoven tialektilä*. Upublisert læremateriell.

Becker, Reinhold von (1824): *Finsk grammatik*. Åbo: Bibel-Sällskapet.

Beddari, Olav (1987): *Niin saapi sanoa: pieni ruijansuo-malainen lukukirja*. Vadsø: Skoledirektøren i Finnmark.

Eriksen, Agnes (2011): *Kummitus ja Tähtipoika 1*. Ruija forlag.

Eriksen, Agnes (2014–2017): *Minun kieli: minun aaret 1–7*. Lakselv: Porsanger kommune.

Eriksen, Aili (2018): *Hiljemin ja huonet: riskisanoja porsanginkveenin ja suomen kielen välillä*. Masteroppgave i finsk språk og kultur. University of Jyväskylä. <jyx.jyu.fi/handle/123456789/57715> (juli 2024).

Eskeland, Tuula, Anna-Riitta Lindgren & Marjatta Norman (2003): Osima ja Baskabusk – monet suomet Norjassa. I: Hannele Jönsson-Korhola & Anna-Riitta Lindgren (red.): *Monena suomi maailmalla: suomalaisperäisiä kielivähemmistöjä*. Tietolipas, nr. 190. Helsinki: Suomalaisen kirjallisuuden seura.

Haavisto, Mervi, Kaisa Maliniemi, Leena Niiranen, Pirjo Paavalniemi, Tove Reibo & Trond Trosterud (2014): Kvensk ordbok

på nett – hvem har nytte av den? I: Ruth Vatvedt Fjeld & Marit Hovdenak (red.): *Nordiske Studier i Leksikografi 12*. Rapport fra Konferanse om leksikografi i Norden, Oslo 13.–16. august 2013. Oslo: Novus forlag. 176–192.

Huss, Leena (1999): *Reversing Language Shift in the Far North: Linguistic Revitalization in Northern Scandinavia and Finland.* Acta Universitatis Upsaliensis.

Isaksen, Ann-Mari (2018): *Kvensk grammatikk 5.–7. klasse.* Tromsø: Ruija forlag.

Johnson, Ryan, Lene Antonsen & Trond Trosterud (2013): Using finite state transducers for making efficient reading comprehension dictionaries. I: *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NoDaLiDa), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16, 59–71.

Kainun institutti – Kvensk institutt (2007): *Program for revitalisering av kvensk språk og kultur (2004). Rapport 2006–2007*. Kainun institutti – Kvensk institutt.

Keränen, Mari (2018): Language maintenance through corpus planning. I: *Acta Borealia* 35, 176–191.

KORP. UiT Norges arktiske universitets og det norske Sametingets kvenske tekstsamling. Versjon 3.4. 2024. <github.com/giellalt/corpus-fkv/tree/77d64f08054397639e6b7c207444e67b1d31f9b0/converted> (juli 2024).

Kvensk språkråd (2010): Sak 5/10 Eventuelt – Kvensk ordbok. Referat fra møte i Kvensk språkråd, 24.–25.3.2010. Kainun institutti – Kvensk institutt.

Kvensk språkting (2008): Møteprotokoll fra møte i Kvensk språkting, 18.04.2008. Kainun institutti – Kvensk institutt.

Laakso, Johanna, Anneli Sarhimaa, Sia Spiliopoulou Åkermark & Reetta Toivainen (2016): *Towards Openly Multilingual Policies and Practices: Assessing Minority Language Maintenance Across Europe*. Bristol: Multilingual Matters.

Lane, Pia (2010): "We did what we thought was best for our children": a nexus analysis of language shift in a Kven community. I: *International Journal of the Sociology of Language* 202, 63–78.

Lane, Pia, Kristin Hagen, Anders Nøklestad & Joel Pristley (2022): Creating a corpus for Kven, a minority language in Norway. I: Sjur Moshagen, Lene Antonsen & Øystein Vangsnes (red.): *Morfologi, målstrev og maskinar – Trond Trosterud {fyller | täyttää | deavdá | turns} 60!. Nordlyd* 46(1), 159–170.

Lanes, Laila (2023): Stor interesse for å lære kvensk: – En wow-faktor sier Hilde Skanke. NRK, 2. oktober 2023. <www.nrk.no/kvensk/okende-interesse-for-a-laere-kvensk-1.16577050> (juli 2024).

Larsen, Karin, Pirjo Paavalniemi & Eira Söderholm (2012): *Allmenn innføring i skriving av kvensk.* Børselv: Kvensk institutt.

Leskinen, Kaisa Grimsby (2017): *«Norjalainen sanoo beriket» (Nordmenn sier beriket). En korpusstudie av norske lånord i kvensk hos bugøynesværinger. Masteroppgave i nordisk språkvitenskap*. Trondheim: Norges teknisk-naturvitenskapelige universitet.

Niiranen, Leena (2022): Språkdokumentasjon innen fennistikken og kvensk. I: Sjur Moshagen, Lene Antonsen & Øystein Vangsnes (red.): *Morfologi, målstrev og maskinar – Trond Trosterud {fyller | täyttää | deavdá | turns} 60!. Nordlyd* 46(1), 181–192.

Nilsen-Børsskog, Alf (2004): *Kuosuvaaran takana.* Iđut forlag.

Norske Kveners Forbund & Norsk senter for kvænsk språk og kultur (2004): *Program for revitalisering av kvensk språk og kultur 2005–2007*.

NoWaC (Norwegian Web as Corpus) (2010): Universitetet i Oslo: Tekstlaboratoriet.

Ogilvie, Sarah (2011): Linguistics, lexicography, and the revitalization of endangered languages. I: *International Journal of Lexicography* 24(4), 389–404.

Rasmussen, Torkel (2004–2005): Hvor mange kan finsk og kvensk i Nord-Norge? I: *Arina. Nordisk tidsskrift for kvensk forskning* 1, 48–54.

Räisänen, Anna-Kaisa (2014): Minority Language Use in Kven Communities – Language Shift or Language Revitalization. I: Julia Sallabank & Peter Austin (eds.): *Endangered Languages: Beliefs and Ideologies in Language Documentation and Revitalization*. Oxford, U.K.: Oxford University Press. 97–108.

Räisänen, Anna-Kaisa (2020): *Linus ja Karhu riepu seilathaan*. Tromsø: Ruija forlag.

Sundelin, Egil (2008): Norske kveners forbund/Ruijan Kveeniliitto. Fra idé til virkelighet. I: *Ottar* 269 (2008), 12–18.

Suomen kielen nauhoitearkisto – Finska bandarkivet. Institutet för de inhemska språken. <www.sprakinstitutet.fi/sv/arkiv/dialekt-_och_namnarkiv/finska_bandarkivet> (august 2024).

Söderholm, Eira (2006): Planlegging av kvensk språk – utvikling av ordforråd. I: Anna-Riitta Lindgren, Einar Niemi, Marit Anne Hauan, Leena Niiranen & Trond Thuen (red.): *Kvener og skogfinner i fortid og nåtid. Rapport fra seminaret "Kvener og skogfinner i fortid og nåtid – identitetsforvaltning og strategier", Vadsø oktober 2005*. Speculum boreale nr. 9. Skriftserie fra instituttet for historie. 47–50.

Söderholm, Eira (2009): *Aikamatka*. <uit.no/Content/207399/cache=20231105100037/Aikamatka%201%E2%80%9310.PDF> (juli 2024).

Söderholm, Eira (2014): *Kainun kielen grammatikki*. Helsinki: Suomalaisen kirjallisuuden seura.

Söderholm, Eira (2017): *Kvensk grammatikk*. Oslo: Cappelen Damm Akademisk.

Trosterud, Sindre Reino, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto & Kaisa Maliniemi (2017): A morphological analyser for Kven. I: *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*.

St. Petersburg, Russia: Association for Computational Linguistics. 76–88. <aclanthology.org/W17-0608> (juli 2024).

Trosterud, Trond (1996): Den finske dialektkrigen (1820–1840). I: *Mål og makt* 26, 130–151.

Trosterud, Trond (2019): Kva bruker vi minoritetsspråksordbøker til? Ein studie av brukarloggane for tolv tospråklege ordbøker. I: *LexicoNordica* 26, 177–198.

Trosterud, Trond & Hilde Skanke (2012): Kvensk juridisk terminologi. I: *Terminologen* 1(1), 56–63.

UiT (2019) = Institutt for språk og kultur ved UiT Norges arktiske universitet: *Meiđän joukko: Grunnkurs i kvensk*. <kvensk.uit.no/> (juli 2024).

Utvik, Hanne-Elin (1996): *Norske ord i finsk språkdrakt. En studie av nyere skandinaviske substantivlån i kvensk / ruijafinsk tekstmateriale med hovedvekt på norske lån.* Hovedoppgave i finsk språk. Universitetet i Tromsø.

Anna-Kaisa Räisänen
fagansvarlig
Kainun institutti – Kvensk institutt
Østre Porsangerveien 4202
NO-9716 Børselv
anna-kaisa.raisanen@kvenskinstitutt.no

Thomas Brevik Kjærstad
prosjektmedarbeider
Kainun institutti – Kvensk institutt
Østre Porsangerveien 4202
NO-9716 Børselv
thomas@kvenskinstitutt.no

Aili Eriksen
språkmedarbeider
Kainun institutti – Kvensk institutt
Østre Porsangerveien 4202
NO-9716 Børselv
aili.eriksen@kvenskinstitutt.no

Trond Trosterud
professor
UiT Norges arktiske universitet
NO-9037 Tromsø
trond.trosterud@uit.no

Representativeness and biases in Icelandic corpora

Einar Freyr Sigurðsson & Steinþór Steingrímsson

All language data are inherently biased, as collection methods, availability of texts and recordings, and the views of the collectors will always affect the process and its results. We examine how bias is manifested in NLP tools trained on corpora and how that could be used to detect biases in Icelandic corpora. We also look at how sports coverage seems to exhibit something that we call male-by-default bias, as an example of a bias that might be hard to detect using automatic approaches. Finally, we suggest how metadata could be enriched to perform better analyses.

1. Introduction

In recent years there has been substantial growth in available language resources for use in Icelandic language technology and linguistic research.[1] The Icelandic Gigaword Corpus (IGC) is the largest of these resources. Its latest version comprises approximately 2.5 billion words. *Tímarit.is* is a collection of all major newspapers, magazines and periodicals published in Icelandic from the 19th century to the present day, digitized using OCR (Optical Character Recognition) and made available online. For both corpora, the focus is on quantity rather than on balance or on being representative of the language as a whole. By *representativeness* of a corpus, we refer to the relation between the corpus and the language it is being used to represent (Hunston 2008).

When we do research using corpora, we are faced with questions of potential biases regarding what is represented and to what

1 We would like to thank two anonymous reviewers and the editors for very helpful comments on this paper.

degree. Some things may be overrepresented, while others may be underrepresented. These biases can, however, often be difficult to detect. The data contained in the two corpora mentioned above 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 are 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. When language data are used for research, the researcher must be aware of these limitations.

In this paper, we propose two research questions:

1. How can we use existing corpora to find ingrained biases, such as gender biases?
2. What kind of metadata is needed to facilitate research on biases and representativeness?

We seek to answer these questions from the viewpoint of Icelandic corpora, discuss potential biases in these corpora with respect to representativeness, and discuss possible approaches for answering these questions.

We examine how bias is manifested in NLP (Natural Language Processing) tools trained on corpora and how that could be used to detect biases in Icelandic corpora. We also look at how sports coverage seems to exhibit something that we call male-by-default bias, as an example of a bias that might be hard to detect using automatic approaches. Furthermore, we suggest how metadata could be enriched in order to better analyse where and how biases and other specific types of artifacts present themselves in the data.

2. A note on biases

Before we go any further, it is crucial to state what we mean when we refer to biases. It may be helpful to look at dictionary definitions of the word *bias*. Among other things, *Merriam-Webster* mentions 'prejudice' and 'deviation' under its definitions of the noun *bias*. One definition of the noun in *Collins English Dictionary* talks about bias being "a tendency to prefer one person or thing to another, and to favour that person or thing", and includes the example "Bias against women permeates every level of the judicial system". Finally, *Cambridge Dictionary* talks about *bias* as "the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment" and also as "the fact of a collection of data containing more information that supports a particular opinion than you would expect to find if the collection had been made by chance".

The definitions picked out from these dictionaries are in line with our thinking of the term and what we are referring to when we use it in this paper, where *bias* entails that there is something skewed, where, for example, the sample does not represent the population.

Many different types of bias have been detected and some are well known and have a name of their own. The entry for *bias* in the *Cambridge Dictionary* gives three examples of biases, *publication bias* (negative results are not reported), *selection bias* (e.g., in social studies) and *survivorship bias* (the tendency for failed samples to be excluded from performance studies). Other commonly known biases include *confirmation bias* and *authority bias*. When we discuss various biases in this paper, we name them for the sake of clarity.

When we work with language, whether it is in the context of linguistic research, when building language technology tools or as

lexicographers, we are constantly faced with questions of what and whose language is represented.

Linguists sometimes give their own linguistic judgments when they are studying their native language. From their own acceptability judgments, they may overgeneralize and claim that a certain phenomenon or construction is ungrammatical in the language they speak. If the linguists do not consult other speakers, who might disagree with their judgments, we may witness something we can call *my-grammar-as-the-default bias*.

We might encounter something similar if we were, for example, working on enhancing a dictionary by finding and selecting words not found in previous versions. We might be biased towards accepting new words that we are familiar with rather than words we have not heard before. We could refer to this as *my-dialect-as-the-default bias*.

Furthermore, representativeness and biases are clearly a factor when we look at prescriptivism. Prescriptivist directions on language use often give examples of actual language use that is to be avoided, for one reason or another. When Böðvarsson (1992:21) states that the plural of *Ástrali* 'Australian' is *Ástralar* 'Australians', he also adds that the plural *Ástralir* is incorrect. He would probably not have mentioned this unless he knew or believed *Ástralir* is a form that people actually use. In the light of this, it is interesting that when we search for "Ástralir" and "Ástralar" in the IGC, we get 2,465 results for the "incorrect" form *Ástralir*, while we only get 1,888 hits for the "correct" form *Ástralar*, or 57% vs. 43%, respectively. This indicates that the prescriptively favoured variants do not necessarily represent the language or grammar of the majority of speakers – or how the language is actually spoken.

Prescriptivism often also invokes authority bias – if directions on how to speak properly do not conform with our own grammar, it can matter who it is that considers this or that to be improper. An interesting situation may currently be developing in Icelandic

discourse: In recent years, professor emeritus Eiríkur Rögnvaldsson, a well-known linguist in Iceland, has objected to how prescriptivist directions are often given and contradicted in various ways by actual language use – see, e.g., Rögnvaldsson (2022). On the one end we have the prescriptivist authorities that have been followed throughout the years, whereas on the other we have another authority objecting to saying that linguistic phenomena like "dative sickness" are incorrect or improper language.

Still, the question arises why we should be concerned about biases in corpora. For one thing, some biases we encounter can be harmful. Nonetheless, it is not always clear in what way. Blodgett et al. (2020) surveyed 146 papers on biases in NLP systems. It was generally unclear in these papers "what kinds of system behaviors are harmful, in what ways, to whom, and why" (Blodgett et al. 2020:5454). It is not immediately clear whether something like the my-grammar-as-the-default bias can cause harm, but it might be harmful if certain features of someone's grammar or dialect are stigmatized; individuals may, for example, become insecure when expressing themselves. Such biases could also "lead to feelings of invisibility and marginalization", to use Friðriksdóttir & Einarsson's (2024:7596) words when discussing potential harms caused by gender biases.

In fact, in some cases it is more obvious how gender biases could be harmful. Such biases have frequently been discussed in recent years in relation to language technology. Some of this research, like the aforementioned study by Friðriksdóttir & Einarsson (2024), focuses on Icelandic. Friðriksdóttir & Einarsson conducted an experiment in which they got large language models to predict the pronoun in a sentence on the form *he/she/they is/are a(n) <occupation term>*. The goal was to see whether large language models "merely echo the gender distribution within respective professions" or if "they exhibit biases aligned with their grammatical genders" (Friðriksdóttir & Einarsson 2024:7597). In this

context, they define gender bias "as the tendency of these models to generate or perpetuate gender stereotypes" (Friðriksdóttir & Einarsson 2024:7596). They furthermore state that a gender bias "can reinforce harmful societal norms, such as by influencing individuals' perceptions regarding the careers or roles accessible to them based on their gender" (ibid.).

We will not be discussing Friðriksdóttir & Einarsson's (2024) study further. However, we look at a different study on gender bias, by Sólmundsdóttir et al. (2021, 2022), in machine translation from English to Icelandic in section 3.2, and in section 3.3 we show examples of how we can detect biases and imbalances in the IGC and *Tímarit.is* that relate to gender or sex.

3. Detecting biases and imbalance in Icelandic corpora

The largest resources, both in terms of mere quantity and timespan, for studying written Icelandic are the IGC and *Tímarit.is*. Therefore, when examining potential biases and imbalances in Icelandic corpora and how we can detect them, we work with these two corpora. In this section we also discuss how biases present themselves in NLP tasks such as machine translation (MT) and word embeddings, and how such tasks can be used to help identifying biases in corpora.

3.1 Representativeness and balance

The IGC is an ongoing corpus project, with new versions of the corpus published on a regular basis. The first edition, published in 2017, contained 1.3 billion running words. The latest edition, published in 2022, has over 2.5 billion running words in eight subcorpora: journals, news, social media, parliamentary proceedings,

adjudications, laws, published books, and Wikipedia (see description in Barkarson et al. 2022). The corpus has become fundamental for linguistic research, dictionary work, and language technology for Icelandic. The data collection has focused on collecting recently published data from sources that allow for the data to be distributed with permissive licenses. All the data are collected in digital format from the source or publishers; no OCR is carried out when the corpus is compiled. This entails that the corpus comprises mostly recent texts, with the vast majority having been published after the year 2000. *Tímarit.is*, on the other hand, contains only OCR-processed texts. The accessibility also differs, the IGC is PoS-tagged and lemmatized, searchable through Korp, a KWIC-system for studying corpora, as well as being downloadable, while *Tímarit.is* is only available through a string-based search engine and is neither tagged nor lemmatized.

The goal of the IGC project is to build as large a corpus as possible of contemporary texts in a language spoken by less than 400 thousand people and, instead of emphasizing representativeness, the aim was to achieve as much coverage as possible and to provide extensive metadata so that users of the corpus can construct their own subcorpora as needed. Steingrímsson et al. (2018:4361) point out in the original IGC paper that trying to achieve representativeness in a text corpus can be problematic. First of all, what should it be representative of? It can be difficult to determine whose language and perspective is represented, and then again how accurately it is represented. And because it can be hard to determine where a variety of language ends and another begins, any corpus is virtually by definition biased to a greater or a lesser extent. In their discussion on balance and representativeness, Beelen et al. (2022) raise the question of what type of representativeness is desirable, and they describe how problematic it can be to achieve a balance of perspectives without downplaying any of them.

Finally, to know the strengths and limitations of the corpus

better, the user needs to understand which type of language the corpus represents. Rich metadata on the origins, classification, and analysis of the text are essential for reaching this understanding and to discern which biases can be expected, although experimentation is also needed to reveal them.

3.2 Machine translation

Machine translation (MT) systems rely on large amounts of data for training. Neural machine translation (NMT), which has been the dominant paradigm in MT since 2016–2017, needs large parallel corpora in order to achieve acceptable translation capability, while the more recent large language models (LLMs) are trained on billions of words of monolingual texts. In both cases, the texts are likely to reflect views and opinions of those who wrote or published the texts, and these may or may not be appropriate for the MT systems being trained.

Sólmundsdóttir et al. (2021, 2022) detected difference in gender use in Google translations from English to Icelandic. The authors point out that technology could maintain "societal inequalities and outdated views" (Sólmundsdóttir et al. 2022:3113) and come to the conclusion that "the results show a pattern which corresponds to certain societal ideas about gender and gender roles" (Sólmundsdóttir et al. 2021:199). Furthermore, a large amount of data will not necessarily guarantee its diversity, see, e.g., the discussion in Bender et al. (2021).

English does not exhibit gender on adjectives and past participles, whereas Icelandic does. It is therefore interesting to see how sentences like *I am <ADJECTIVE>* are translated. In their research, using Google Translate, Sólmundsdóttir et al. (2021, 2022) observed a peculiar gender bias when translating predicative sentences from English to Icelandic. They compiled a list of adjectives generally used to describe people and classified them in two categories: 1) words that describe personality traits, such as

strong, *weak*, *clever*, or *stupid*, and 2) words that describe appearance, such as *beautiful*, *ugly*, *fat*, or *thin*.

The authors found that adjectives describing positive personality traits appeared more often in the masculine form, while negative ones were more likely to appear in the feminine form. They examined 262 adjectives that describe people's personal traits. 156 of them were translated in the masculine, whereas 65 were translated in the feminine – with the rest being translated in the neuter, as uninflected adjectives or as syncretic for masculine and feminine (Sólmundsdóttir et al. 2021:189). Interestingly, 59% of the adjectives used in the masculine were considered by the authors to describe positive features, whereas only 23% of the adjectives used in the feminine were positive (Sólmundsdóttir et al. 2021:189). Two examples are shown below, where *strong* is translated as the masculine *sterkur* and *weak* as the feminine *veik*:

(1) English: I am *strong*.
Icelandic: Ég er *sterkur*. (masculine; positive)
(2) English: I am *weak*.
Icelandic: Ég er *veik*. (feminine; negative)
(Sólmundsdóttir et al. 2021:190)

The study also looked at 67 adjectives that describe people's look or appearance, such as *beautiful* and *handsome*. Of these, 31 adjectives were translated into the masculine gender, whereas 15 were translated into the feminine. Here, however, the ratio of positive adjectives is much higher among the feminine usage: 67% of the adjectives in the feminine were positive as opposed to 23% of the adjectives that were used in the masculine (Sólmundsdóttir et al. 2021:191–192).

Moreover, the research tested adjectives in predicative sentences that describe the speaker's ability to carry out certain tasks. Two examples are shown in the following:

(3) English: I am *good* at electrical work.
Icelandic: Ég er *góður* í rafmagnsvinnu. (masculine)
(4) English: I am *good* at cooking.
Icelandic: Ég er *dugleg* að elda. (feminine)
(Sólmundsdóttir et al. 2021:193)

In this part of the research, Sólmundsdóttir et al. focused on sentences that describe people's ability in craft and industry, on the one hand, and housekeeping, on the other. When adjectives in sentences like *I am good at electrical work* (ability in craft and industry) were translated, they were used in the masculine in 12 out of 15 cases. Furthermore, different adjectives seemed to be used, depending on the gender: *dugleg* for the feminine, *góður* for the masculine, even though they were used to translate the same English adjective, *good*. When adjectives in sentences describing housekeeping were translated, they were in the feminine in 18 out of 21 examples (Sólmundsdóttir et al. 2021:192–193).

The results for sentences containing adjectives which describe people's appearance (e.g., *beautiful*, *handsome*) are in some ways opposite to the results for adjectives that describe people's personality traits (e.g., *strong*, *weak*). The authors state that this shows a pattern which corresponds to societal ideas about gender. In the light of how MT systems are developed, this bias must reflect the data the systems are trained on. When the systems are faced with ambiguity, they generate the most likely translation with the likelihood derived from the training data. This is thus an example of an MT system perpetuating a societal bias presented in corpora used to build these systems.

LLMs trained on massive monolingual data sets in multiple languages rather than parallel corpora, have been shown to exhibit similar tendencies. Vanmassenhove (2024) ran a small experiment where she evaluated ChatGPT (based on GPT-3.5) in terms of translating ambiguous words with respect to gender from

English to Italian. Her findings indicate a strong male bias which becomes even more prevalent when asked to generate alternatives. She concludes with a call to raise awareness about these issues and taking proactive steps to address them.

Going back to our discussion on detecting bias in corpora, the case of MT shows that understanding what different parts of text corpora represent, and what biases we are likely to find in them, can be crucial when developing NLP systems which derive their model of the language from data, whether we want to mitigate gender bias or other undesirable artifacts.

3.3 Bias and imbalance detection

In previous subsections, we have discussed how biases and imbalances can have undesirable effects in different NLP tasks, where models based on text corpora are employed. In this subsection we look at how data in corpora can be imbalanced and biased with respect to gender, and how these biases are not necessarily easily detected.

3.3.1 *I am …* (counting linguistic phenomena)

Inspired by Sólmundsdóttir et al.'s (2021, 2022) work, we decided to look for examples that are somewhat similar to the ones they discuss (e.g., *I am good at housekeeping*).[2] We searched for sentences that start with *ég er* 'I am' immediately followed by a "strongly" inflected adjective or past participle[3] in masculine, feminine and neuter, which in turn is followed by an infinitival marker and

2 We have, however, not looked at the distribution of the gender of adjectives with respect to, e.g., positive and negative personality traits. We leave that for future research.

3 By excluding weak inflection we exclude various examples that are syncretic for feminine and masculine, and we also exclude many examples that are not applicable, such as *Ég er meira að segja …* 'I even am …'.

then a verb in the infinitive.[4] We show examples for each gender use below that we found using our search queries in the IGC.

(5)	Ég	er	*glaður*	að	hafa	gert	það.
	I	am	glad.MASC.SG	to	have	done	that
	'I'm glad I did it.'						
(6)	Ég	er	*búin*	að	reyna	allt.	
	I	am	done.FEM.SG	to	try	everything	
	'I have tried everything.'						
(7)	ég	er	*búið*	að	biðjast	afsökunar :)	
	I	am	done.NEUT.SG	to	ask	apology	
	'I have apologized.'						

With this search we find examples where the predicate consists of a single word – an adjective or a participle – as it is immediately followed by an infinitival clause; therefore, the adjective or participle could not be part of a larger noun phrase (which would determine its gender, as in *Ég er glaður nemandi* 'I'm a glad/happy student', where the masculine noun *nemandi* determines the grammatical gender of the adjective). This gives us examples where an adjectival or participial predicate agrees with *ég* 'I' (which itself does not show gender features). The gender of the adjective or participle then indicates the gender or sex of the speaker. We take the speaker in (5) to be a male speaker, the one in (6) to be a female speaker, while speakers in examples like (7) could be non-binary or genderqueer using neuter when referring to themselves.

We expected utterances by male speakers to be the most frequent out of the three as we expected male speakers to be the most

4 The search query for feminine gender was as follows: <sentence> [word = "Ég" %c] [word = "er" %c] [(pos = "l" | hattur = "þ") & lob = "s" & kyn = "v"] [word = "að" %c] [hattur = "n"]. By replacing kyn = "v" with kyn = "k" or kyn = "h" in the query, we get the search queries we used for the masculine and neuter, respectively.

dominant in the overall discussion. That was not the case, however.[5]

(8) a. Masculine: 31,731 results
b. Feminine: 75,263 results
c. Neuter: 742 results

The fact that there are twice as many examples of feminine agreement with *ég* 'I' comes as a surprise if we take these results at face value. Why are there so many utterances that suggest female speakers as opposed to male speakers? Does the IGC generally represent much more female speakers than male speakers? Or do female speakers use this construction more than male speakers? We could ask many other questions to get a clearer picture – showing that we need to take a closer look at what is behind these numbers.

	Feminine	**Masculine**
Journals	20	14
News	5,519	8,392
Social media	69,134	21,412
Parliamentary proceedings	494	1,772
Adjudications	4	16
Laws	5	1
Books	87	124
Wikipedia	0	0
All IGC	**75,263**	**31,731**

Table 1: Feminine vs. masculine agreement in the IGC.

5 It should be noted that the majority of the 742 results for the neuter do not reflect the gender of the speaker, as many of these results are utterances like *Ég er satt að segja* ... 'I am, truth be told, ...' where the gender feature on the adjective – in this case *satt* 'true' – does not come from the subject *ég* 'I'. If the subject were *hann* 'he' or *hún* 'she', the form of the adjective would still be neuter: *Hann/hún er satt að segja* ...

When we look at the distribution with respect to subcorpora within the IGC, see Table 1, we see a different picture than the one painted in (8) above. In four out of eight subcorpora, namely news, parliamentary proceedings, adjudications and books, the masculine is used in the majority of cases. There are no examples of what we looked for in the Wikipedia subcorpus, and only in three subcorpora, namely journals, social media and laws, is there more use of the feminine than the masculine form in the construction we are looking at. Furthermore, out of 75,263 examples of feminine-form use, 69,134 examples are from the social-media subcorpus.

3.3.2 Sports coverage with respect to sex

In a survey looking at sports coverage with respect to sex in Icelandic newspapers from 1 May 1999 to 30 April 2000, around 85% of the coverage was on men's sports, 7% on women's sports, and 8% on sports in general (*Nefnd um konur og fjölmiðla: álit og tillögur* 2001:20). To test whether this has changed, we might try some simple methods, like looking at the frequency of the use of words that indicate which sex is being discussed. Such a word pair is *kvennalandsliðið* 'the women's national team' and *karlalandsliðið* 'the men's national team'.

We looked for these two terms on *Tímarit.is* to see whether the survey's results would be reflected in the use of the two terms. The numbers for the decade 2010–2019 are 1653 (48%) search results for *kvennalandsliðið* vs. 1802 (52%) results for *karlalandsliðið* (these two word forms are each syncretic for nominative and accusative case and have a suffixed definite article; we did not search for results in the dative and the genitive case). Given the numbers in the report mentioned above, this would certainly indicate that there is more discussion on women in sports than before. However, the total results for the two terms are rather interesting, as shown in Table 2.

	Kvennalandsliðið	Karlalandsliðið
1950–1959	9 (100%)	0 (0%)
1960–1969	77 (76%)	24 (24%)
1970–1979	138 (69%)	62 (31%)
1980–1989	631 (73%)	234 (27%)
1990–1999	862 (71%)	354 (29%)
2000–2009	1541 (64%)	851 (36%)
2010–2019	1653 (48%)	1802 (52%)

Table 2: *Kvennalandsliðið* vs. *karlalandsliðið* on *Tímarit.is*.

In all the decades prior to 2010–2019, *kvennalandsliðið* is more frequent in the corpus on *Tímarit.is* than *karlalandsliðið*. Various news reports like the one in (9) shed a light on the possible reason for this (English translation ours).

(9) *Landsliðið í handbolta*
Landsliðið í handbolta sem leikur gegn Luxemborg hefur verið valið [...]
The national handball team
The national handball team, which plays against Luxemburg, has been chosen [...]
(*Þjóðviljinn*, 25 November 1975, p. 14)

This is the heading together with the first words of a very short article about the men's national team in handball. However, there is no mention of this being the men's team – it only says "Landsliðið í handbolta" 'the national handball team'. It cannot be inferred from the context elsewhere on the page that this is the men's national team. That is, 'the national team' without mentioning the sex seems to refer to men by default, something we can call *male-by-default bias*.

The results of our little search query on *Tímarit.is* are yet another example where we cannot simply take the results at face

value and they reveal a gender bias in the data. Even though the ratio for *kvennalandsliðið*, compared to *karlalandsliðið*, is lower for 2010–2019 than any other decade, this may in fact reflect a more balanced coverage of men and women in sports – especially if there is less of the male-as-default bias than in previous decades.

3.4 Word embeddings

Word embeddings are vector representations of words calculated from very large data sets. Using the popular Word2Vec model (Mikolov et al. 2013), each word is typically represented by a 300-dimensional vector of real numbers. These vectors capture semantic and syntactic information about the word, based on surrounding words in the training data. Figure 1 shows an example of how four words could be represented in vector space. By observing their position in relation to one another, we find that by simple arithmetic we can calculate semantic relationships; in this case *King – Man + Woman = Queen*. Similarly, we can find the closest equivalents for any word in a corpus by training word embeddings on the corpus and in a similar manner, calculate the distance between words in the vector space.

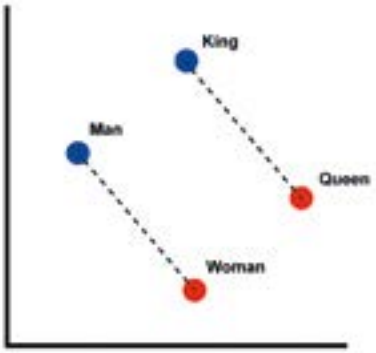


Figure 1: A simplified visual illustration of the relation between four words in vector space. Figure taken from Wikipedia (<en.wikipedia.org/wiki/Word2vec>).

Garg et al. (2018) show that word embeddings can be effective tools to study historical biases and stereotypes. They do this by relating measurements from embeddings trained on 100 years of text data to historical census and survey data, and they find that changes in the embeddings track with demographic and occupation shifts over time.

They use embeddings each trained on a decade of 20th century texts and inspect words from word lists they collate to represent gender and ethnicity, as well as lists of adjectives and occupations. Using the embeddings and word lists they measure the strength of association between a group and neutral words in several experiments. For example, they compute the average embedding distance between words that represent gender, on the one hand, and ethnicity, on the other, as well as words for occupations to estimate the strength of the embeddings to calculate sociological trends over time. They compare the results to historical surveys and show that the embeddings capture both gender and ethnic occupation percentages and consistently reflect historical changes. They also try to quantify ethnic and gender stereotypes by finding the top adjectives associated with different groups over time and the adjectives most biased towards Asians.

1910	1950	1990
irresponsible	disorganized	Inhibited
envious	outrageous	Passive
barbaric	pompous	Dissolute
aggressive	unstable	Haughty
transparent	effeminate	Complacent
monstrous	unprincipled	Forceful
hateful	venomous	Fixed
cruel	disobedient	Active
greedy	predatory	Sensitive
bizarre	boisterous	Hearty

Table 3: Adjectives most biased towards Asians in 1910, 1950 and 1990 in the experiment carried out by Garg et al. (2018).

Table 3 shows how adjectives biased towards Asians have changed over time in the data the embeddings have been trained on. While this may not completely reflect the general attitude of the time, it is at least an indication of how the discussion has changed.

Work such as this shows that machine learning approaches can be used to help us understand bias and representativeness in texts. It can give us some indication of what can be expected if we, or the tools we build, try to generalize from the corpus, but it can also be used for teaching us about the fluidity of the language and how biases may change over time or even between different text sources.

While the IGC mainly comprises texts from the 21st century, the news subcorpora contain information about sources of the texts and for some sources there is also further categorization of the texts. Some of the other sources have information about age and gender of authors, and the parliamentary speeches have party affiliation. Using the word embeddings approach could potentially help us identify biases in the corpus not only over time but also between text categories and in speeches of parliamentarians from different parties.

4. Which metadata are needed?

In the IGC, all information accompanying the texts that were collected is distributed as metadata. All eight subcorpora have publisher, date, and place of publishing, if available. All texts are tagged and lemmatized. For books, journals, and parliamentary proceedings, information on the author is also available. For parliamentary proceedings, gender, year of birth, age when the text was written, and political affiliation are also included. As an example of the usefulness of the metadata for linguistic research, we can mention the study by Stefánsdóttir & Ingason (2018, 2022) on the variable use of stylistic fronting in Icelandic in thousands

of parliament speeches given by parliamentarian Steingrímur J. Sigfússon, where they identify syntactic change across his lifespan related to status-associated factors.

4.1 Generating metadata using NLP tools

Enriching a corpus with more metadata that allow for more fine-grained research could be achieved by analysing the data using NLP tools. *Tímarit.is* does not have any categorization that can be used while searching the corpus, and while the IGC contains some categorization of texts, in particular news from some media outlets, more thorough categorization could be helpful, for example for studies such as the one conducted in section 3.3.2 above. An example of such a categorization could be to analyse news and classify them into fine-grained categories, such as:

(10) *News* → *Sports* → *Handball* → *Men* → ...

Sentiment analysis could be used to investigate which topics are being discussed in a positive or negative fashion, or even how people or groups of people are being discussed in the texts. To make the user able to gather information on certain individuals or groups of people, named entity recognition would make analysis easier.

4.2 Other corpora

For other corpora, for example web-scraped corpora and data in newspaper scanning projects (e.g., *Tímarit.is*), less metadata are available. On the other hand, a lot is often known about the data. A newspaper could be right-leaning or left-leaning, it could be funded by a certain industry or solely by subscriptions. It could be yellow press journalism or business oriented, etc. If we are seek-

ing balance of representation, should we then also take circulation into account? All these types of information can be important when the texts are analysed to identify biases or to understand what or who the texts are likely to represent. By providing these types of additional information, the corpus would make all such endeavours easier for researchers.

Furthermore, some newspapers are read by many, while others are read by few. Should these newspapers be regarded as equal or should researchers using these data use some sort of oversampling or undersampling approaches? Similar principles could apply to books. A book may be popular and be read for decades after publication while another book is only read by 20 people in the first few months after publication and then forgotten. While such information can be useful for some researchers, it has to be acknowledged that corpus publishers do have to prioritize and include what is most likely to be of use, especially in cases where adding the information requires time-consuming manual work.

5. Conclusion

In the introduction above, we outlined the following research questions:

1. Can we use existing corpora to find ingrained biases?
2. What kind of metadata is needed to facilitate research on biases and representativeness?

While we have not made an effort to give concrete and definite answers to these big questions, we have tried to shed a light on the topic with respect to Icelandic and Icelandic corpora.

Firstly, we have, based on discussion on research in machine translation, considered the importance of understanding what dif-

ferent parts of text corpora represent, and what biases we are likely to find in them. Furthermore, we discussed how word embeddings can play a role in detecting biases and how they change over time or differ between text sources. We also pointed out biases that are not necessarily easily detectable and need to be considered.

Secondly, we sketched up possible approaches to enrich the metadata with the goal in mind to facilitate bias and imbalance detection.

A demonstration of how biases are to various extent ingrained in all text corpora and how they can be detected, may help lexicographers, as well as language technologists, understand the limitations of corpora and perhaps facilitate richer data selection that reflects more diverse aspects of the language and language use.

It is a constant problem trying to strike the balance between different sources, quantity and quality of texts, and the perfect balance for one user may not suit all others. Introducing new text sources may introduce biases not prevalent in other sources and while larger corpora may give us more examples to work with, they may also amplify various biases. When enlarging corpora, giving users the tools to analyse them, descriptive metadata and perhaps some analyses may help the users find the balance that suits their needs.

References

Dictionaries and corpora

Cambridge Dictionary. <dictionary.cambridge.org> (April 2024).

Collins English Dictionary. <www.collinsdictionary.com> (April 2024).

Icelandic Gigaword Corpus = Barkarson, Starkaður, Steinþór Steingrímsson, Þórdís Dröfn Andrésdóttir, Hildur Hafsteins-

dóttir, Finnur Ágúst Ingimundarson & Árni Davíð Magnússon (2022): *Icelandic Gigaword Corpus (IGC-2022) – annotated version*. CLARIN-IS. <hdl.handle.net/20.500.12537/254> (April 2024).

Merriam-Webster.com. <www.merriam-webster.com> (April 2024).

Tímarit.is. Landsbókasafn Íslands – Háskólabókasafn. <timarit.is> (February 2024).

Other references

Barkarson, Starkaður, Steinþór Steingrímsson & Hildur Hafsteinsdóttir (2022): Evolving large text corpora: Four versions of the Icelandic Gigaword Corpus. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association. 2371–2381 <aclanthology.org/2022.lrec-1.254>.

Beelen, Kaspar, Jon Lawrence, Daniel C.S. Wilson & David Beavan (2022): Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. In: *Digital Scholarship in the Humanities* 38, 1–22. <doi.org/10.1093/llc/fqac037>.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell (2021): On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*. New York, NY: Association for Computing Machinery. 610–623. <doi.org/10.1145/3442188.3445922>.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III & Hanna Wallach (2020): Language (technology) is power: A critical survey of "bias" in NLP. In: Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Associ-

ation for Computational Linguistics. 5454–5476. <aclanthology.org/2020.acl-main.485>.

Böðvarsson, Árni (1992): *Íslenskt málfar*. Reykjavík: Almenna bókafélagið.

Friðriksdóttir, Steinunn Rut & Hafsteinn Einarsson (2024): Gendered Grammar or Ingrained Bias? Exploring Gender Bias in Icelandic Language Models. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, LREC-COLING 2024*. Torino: ELRA and ICCL. 7596–7610. <aclanthology.org/2024.lrec-main.671>.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky & James Zou (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. In: *Proceedings of the National Academy of Sciences* 115, E3635–E3644. <www.pnas.org/doi/abs/10.1073/pnas.1720347115>.

Hunston, Susan (2008): Collection strategies and design decisions. In: Anke Lüdeling & Merja Kytö (eds.): *Corpus Linguistics: An International Handbook*, Volume 1. Berlin: De Gruyter. 154–168.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado & Jeffrey Dean (2013): Efficient estimation of word representations in vector space. In: *1st International Conference on Learning Representations*. <arxiv.org/abs/1301.3781>.

Nefnd um konur og fjölmiðla: álit og tillögur (2001). Reykjavík: Menntamálaráðuneytið. <rafhladan.is/handle/10802/6120>.

Rögnvaldsson, Eiríkur (2022): *Alls konar íslenska. Hundrað þættir um íslenskt mál á 21. öld*. Reykjavík: Mál og menning.

Stefánsdóttir, Lilja Björk & Anton Karl Ingason (2018): A high definition study of syntactic lifespan change. In: *University of Pennsylvania Working Papers in Linguistics* 24, 169–178.

Stefánsdóttir, Lilja Björk & Anton Karl Ingason (2022): Einstaklingsbundin lífsleiðarbreyting: Þróun stílfærslu í þingræðum Steingríms J. Sigfússonar. In: *Íslenskt mál og almenn málfræði* 44, 151–178.

Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson & Jón Guðnason (2018): Risamálheild: A Very Large Icelandic Text Corpus. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation.* Miyazaki: European Language Resources Association. 4361–4366. <aclanthology.org/L18-1690>.

Sólmundsdóttir, Agnes, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir & Anton Karl Ingason (2021): Vondar vélþýðingar: Um kynjahalla í íslenskum þýðingum Google Translate. In: *Ritið* 3/21, 177–200. <doi.org/10.33112/ritid.21.3.7>.

Sólmundsdóttir, Agnes, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir & Anton Karl Ingason (2022): Mean machine translations: On gender bias in Icelandic machine translations. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association. 3113–3121. <aclanthology.org/2022.lrec-1.333>.

Vanmassenhove, Eva (2024): Gender bias in machine translation and the era of large language models. In: *arXiv* <arxiv.org/html/2401.10016v1>.

Einar Freyr Sigurðsson
Research Associate Professor, ph.d.
The Árni Magnússon Institute for Icelandic Studies
Edda, Sæmundargata 5
IS-107 Reykjavík
einar.freyr.sigurdsson@arnastofnun.is

Steinþór Steingrímsson
Research Assistant Professor, ph.d.
The Árni Magnússon Institute for Icelandic Studies
Edda, Sæmundargata 5
IS-107 Reykjavík
steinthor.steingrimsson@arnastofnun.is

ANMELDELSE

Nordnorsk ordbok, den største i sitt slag

Tor Erik Jenstad

Ove Arild Orvik: *Nordnorsk ordbok. Arven etter Hallfrid Christiansen*. Harstad: Utenfor Allfarvei Forlag 2023. 472 s. 449 kr.

1. Regionale ordbøker i Norge

Med utgjevinga av *Nordnorsk ordbok. Arven etter Hallfrid Chrisiansen* (heretter *Nordnorsk ordbok*) føyer Ove Arild Orvik seg inn i ein nokså tynn norsk tradisjon med det ein noko upresist kan kalle regionale ordbøker. Med dette meiner eg ordbøker som dekker ordforrådet i eit større eller mindre dialektområde ut over den einskilde bygda eller byen, oftast med vekt på det mest typiske eller særmerkte for området (og altså ikkje regionalt standardtalemål). Vi har ein rik og stadig aukande flora av større ordbøker og mindre ordsamlingar for einskilde bygde- eller bymål, men på regionnivået har ikkje aktiviteten vore særleg stor. Eit tidleg døme er *Ordbog over bygdemaalerne i Søndhordland* (Vidsteen 1900), og ein må vel kunne tillata seg å nemne *Trønderordboka* (Dalen & Jenstad 1997 og 2002). For Nord-Noreg sin del har vi eit utplukk i *1000 ord i nord* (Ellingsve 2012), og først og fremst eit viktig bidrag i *Håløygsk ordsamling* (Hveding 1968). Med sine 472 sider, der sjølve ordsamlinga utgjer det aller meste, er nok *Nordnorsk ordbok* den største ordboka av dette slaget i Noreg. Med «nordnorsk» er både her og i boka meint dei tre nordlegaste norske fylka (Nordland, Troms og Finnmark), både administrativt og som dialektområde.

Mindre kjent er det kanskje at dialektforskaren Hallfrid Christiansen (1886–1964) hadde planar om ei nordnorsk ordbok og etterlet seg eit stort materiale til dette. Christiansen er vel mest

kjent for doktoravhandlinga om Gimsøy-målet (Christiansen 1933) og for oversynsverket *Norske dialekter* (Christiansen 1946–48).

Christiansen registrerte mykje ordtilfang på samlarferdene sine om sommaren i perioden 1945–1962. Setlane vart overlatne til Universitetet i Oslo, og seinare er dei flytta til Språksamlingane ved Universitetet i Bergen. Ein kopi av det meste av tilfanget finst også ved NTNU Dragvoll (ein del av Noregs teknisk-naturvitskaplege universitet i Trondheim). Orvik har brukt dette tilfanget, og supplert med materiale frå Nord-Troms og ikkje minst Finnmark, dit Christiansen aldri kom. Tilleggstilfanget er oftast henta frå arkivet til *Norsk Ordbok*. I eit intervju i avisa Norsk Tidend i 2024 kjem det fram at Orvik også har kontrollert opplysningane mot enkeltpersonar og historielag i heile landsdelen (Knudsen 2024).

For ei stutt og grei innføring i Hallfrid Christiansens liv og virke sjå til dømes Karlsen (2011). Vil ein ha ei meir omfattande utgreiing, kan ein gå til Karlsen (1997).

2. Tilfanget etter Hallfrid Christiansen

I setelarkivet til Norsk Ordbok er setlane til Hallfrid Christiansen merkte med medarbeidarnummeret hennar, 622, som kan søkjast opp i Universitetet i Bergen sine samlingsdatabasar (UiB samlingar). Dette gjev 16 369 tilslag og omfattar oppskrifter frå ulike stader i Nord-Noreg, særleg Gimsøy. Ho har også gått gjennom ein del samlingar som andre har stått for. Dessutan er det ekserpt frå forfattarar som i stor grad nyttar nordnorsk ordtilfang, som Regine Normann, Lars Berg, Sigurd Sivertsen og Einar K. Aas / Peter Wessel Zapffe. Ei eldre samling frå E.G. Schytte (Schytte 1807) er også grundig ekserpert av Christiansen og inngår i talgrunnlaget her. Viktig å merke seg er også at fleire av delsamlingane til Christiansen har eigne medarbeidarnummer i setelbasen til *Norsk Ordbok*. Til dømes er nr. 2170 Christiansen-setlar frå Lødingen, nr. 2174

frå Vega og nr. 2300 frå Tromsøysund. Dette er altså medarbeidarnummer, og seier ingenting om talet på setlar. Men det blir i alle fall ein god del fleire setlar enn det ein finn under medarbeidarnummer 622. Uansett står Christiansen med dette fram som ein av dei største og viktigaste bidragsytarane til *Norsk Ordbok* sitt redigeringsgrunnlag. Ein vil rekne med høg kvalitet på oppskriftene til ein så framståande og dyktig granskar, spesielt i uttaleopplysningane. Sjølvsagt finst mange av orda også utanom Nord-Noreg. Somme kan vera heilt allmennorske, men det er også mange døme på ord som er nokså lokalt avgrensa.

3. Kort om forfattaren/redaktøren

Ove Arild Orvik er fødd i 1952 og oppvaksen på Otrøya i Romsdal, i tidlegare Midsund, no Molde kommune. Han har arbeidd som norsklektor i den vidaregåande skulen, men har gjeve ut fleire mindre ordsamlingar, både frå sitt opphavlege heimeområde på Romsdalskysten (Orvik 1989a og 1989b, Orvik 2006) og seinare frå Nord-Noreg, der yrkeslivet førte han til Sortland i Vesterålen (Orvik 1997, Orvik 2003). Han har vore særleg oppteken av ord og nemningar knytt til kystkultur og laga i 2017 heftet *Kystspråk*, der han jamfører kystrelaterte ord frå Aust-Island, Vesterålen og Romsdal (Orvik 2017). Det er såleis ein røynd populærvitskapleg leksikograf vi har å gjera med. *Nordnorsk ordbok* er hans mest omfattande prosjekt hittil, og med ein stad mellom 6000 og 7000 oppslagsord er dette også den største nordnorske ordsamlinga som er utgjeven. Dette rettferdiggjer bruken av «ordbok» i tittelen.

4. Presentasjon av boka

Boka tek til med eit forord der det er gjort greie for arbeidet, følgt av eit kort kapittel om Hallfrid Christiansen. Deretter får vi ei tri-

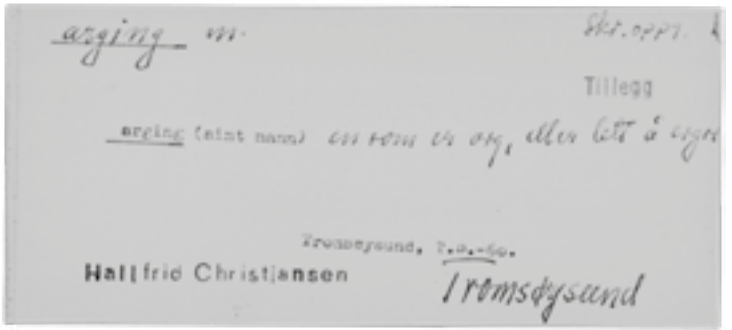
veleg skildring av ein dag i Hallfrid Christiansens samlarliv, slik han kunne ha arta seg på Meyergården på Mo i Rana i juli 1949. Ei stutt skisse over hovuddrag i nordnorske dialektar har det også vorte plass til, før innleiingsdelen blir avslutta med ei orientering om korleis ordlista er ordna, ei forklaring av notasjonen/lydskrifta og ei liste over forkortingar (som med fordel kunne vore ordna alfabetisk). Hovuddelen er ei alfabetisk ordliste frå s. 23 til s. 463. Heilt til slutt kjem ei litteraturliste, ei takkeliste til folk som har hjelpt til, og ein stutt biografi over forfattaren.

Oppbygginga av dei enkelte artiklane er det gjort greie for i innleiingsdelen s. 16. Oppslagsordet, som er på normert nynorsk, står med utheva skrift, følgt av ein klammerparentes med uttalen i ei modifisert Norvegia-lydskrift, og deretter opplysning om ordklasse. Artikkelkroppen inneheld ein eller fleire definisjonar, følgt av heimfesting. Denne er ofte til einskildkommunar, men kan også vera til større område, som Lofoten. Når ordet etter Orvik si vurdering er funne tilstrekkeleg vidt spreidd, får det merkinga «Vanleg». Det er ikkje sagt noko eksplisitt om kva kriterium som er brukt for dette. Bruksdøma står i kursiv, og som regel er dette slike som Christiansen har notert, men omsett til redaksjonsmålet nynorsk. Av og til er det også med sitat frå ei av dei skriftlege kjeldene som Christiansen sjølv har ekserpert frå og truleg har rekna som illustrerande for nordnorsk ordbruk (jf. del 2). Døma ser elles ut til å vera ført samla etter alle tydingane, og ikkje fordelt under kvar tyding, såleis til dømes i artikkelen *all*. Innimellom er det tilvisingar mellom (nær)synonym, til dømes frå *finntukt* til *hundtukt* (men ikkje andre vegen i dette tilfellet). Med to spalter på kvar side, høveleg store skrifttypar og tilstrekkeleg «luft» mellom oppslaga er denne delen lett og god å lesa, akkurat som resten av boka.

Vi kan ta med eit relativt enkelt døme som viser vegen frå Christiansen si oppskrift til *Nordnorsk ordbok*, artikkelen *arging*, m. Den ser slik ut (Norvegia-teikn for velar nasal er ikkje med her):

Arging [*ar`geng*] m sint person, menneske som er lett å terge (Tromsøysund): *Pass dæ førr den argengen!*

Om vi jamfører med setelen (jf. figur 1 nedanfor), ser vi at Christiansen har ordet frå Tromsøysund, innsendt av ein av informantane hennar («skr.oppg.»). Definisjonen er formulert noko om. Blant anna er «mann» bytt ut med «person». Dessutan ser vi at Orvik har lagt til eit døme på bruken.



Figur 1: Setel oppskriven av Hallfrid Christiansen for substantivet *arging* i setelarkivet til *Norsk Ordbok*.

5. Nokre kritiske merknader

Vurderinga her er først og fremst basert på ein snøgg gjennomgang av bokstaven *a*, som er det same strekket som meldaren no er i ferd med å redigere for eit prosjekt der ein er i gang med å revidere *Norsk Ordbok* frå A til H. Elles er det gjort meir tilfeldige nedslag rundt ikring i boka. *Nordnorsk ordbok* har i alt 230 oppslag på *a-*. På Christiansens medarbeidarnummer 622 i setelarkivet til *Norsk Ordbok* får ein tilslag på 489 setlar i det same alfabetstrekket. Sjølvsagt kan det vera mange setlar på same oppslagsord. Til dømes er det 25 setlar på *auga*, n. På andre sida kjem alt tilfanget som har andre medarbeidarnummer i *Norsk Ordbok*,

som nemnt ovanfor, og som gjer at det totale talet på aktuelle ord aukar.

Det er altså gjort eit utval, utan at eg kan sjå at dette er gjort eksplisitt greie for nokon stad i boka. Med det omfanget det er lagt opp til, gjev det seg sjølv at ein ikkje kan ta med slikt som uttalevariantar av allmenne ord som *absolutt* (to setlar hos Christiansen på nummer 622). Derimot kunne eg godt ha tenkt meg at til dømes *ammunisjon* hadde fått ei innførsle, ut frå uttalen «ammusjon». Men dette er vanskelege avvegingar. Døme på ord Christiansen har registrert, og som eg synest burde vore med, er *ardjarn*/*-jern* 'skjer på plogtypen ard', innbyggjarnemninga *alstværing* 'person frå øya Alsta' og sjeldne, men interessante ord som adjektivet *aur-dikut* 'fullt av myrhol med aur (grus)' og verbet *avkjølmast* 'bli forfrosen (så ein ikkje greier å gjera noko)'. I artikkelen *ambodty* '(liten) bruksgjenstand' burde det vore opplyst om uttalen «handbodty».

Sidan boka er dedikert til arven etter Hallfrid Christiansen, burde det kanskje gått tydelegare fram av dei enkelte artiklane kva som er frå henne, og kva som er henta frå andre. Til dømes har ikkje Christiansen opplysningar om ordet *agnkipe* i det heile, så denne artikkelen må byggje på andre kjelder (truleg mellom anna Hveding (1968)). Bruksdømet i denne artikkelen høyrer forresten til under usamansett *kipe*. Men her kjem naturlegvis plassomsyn sterkt inn. Artikkelen *arl* n 'uro, stadig og slitsom aktivitet' byggjer sannsynlegvis på Hveding (1968), han er oppgjeven som einaste kjelde her. Men Hveding sin eigen definisjon er 'uro, stendig rørsla, ofte under ståk', så denne er omarbeidd. Heimfesting til Trondenes er lagd til, medan Astafjord og Senja, som også finst i arkivet til *Norsk Ordbok*, ikkje er med. «Ofoten» hos Hveding har her vorte til «Lofoten». Bruksdømet hos Hveding, som er godt og illustrerande, er ikkje teke med. Christiansen har ikkje opplysningar om dette ordet i det heile. Såleis finst det ting å pirke på når ein går inn i detaljane. Kjeldeføringa ser ut til å vera berre delvis

gjennomført, og om det er eit system i det, er det ikkje gjort klårt greie for.

Det er elles ein lei ombrekkingsfeil under bokstaven *a*, truleg på grunn av problem med å skilje *akter-* og *atter-* i samansetningar. På s. 33 kjem *akterrom* inn etter *atterlot*. Etter uttalen burde normeringa truleg vore *atterrom*, og da er dette på rett plass. Men så kjem heile bolken frå og med *akterskot* til og med *atterlot* på nytt. Denne bolken står altså både på s. 25–33 og på s. 33–42. På s. 42 er det elles alfabetiseringsfeil, med rekkjefølgja *atti*, *attfor*, *atthål* og *attersigle*. Vidare er det to artiklar *all*, der den eine kjem før *ald-*. Det er delvis overlapping i tyding, og dei to artiklane burde vore samkøyrt, i tillegg til å stå på same alfabetiske plass.

Oppslagsformene verkar jamt over rette og rimelege og ser som nemnt ut til å vera i tråd med reglane for standardnynorsk. Enkelte kan sjølvsagt diskuterast, men da er det ved fleire høve Christiansen si eiga normering som er brukt. Det er tilfellet med til dømes *dabbedant* 'tullball, uvesentleg person', som eg ikkje skjønar skal ha denne oppslagsforma ut frå uttalen «dobbedant». Om ein først skulle avvike frå uttaleforma her, ville *dubbedant* vore meir naturleg. Eit anna døme er *bol* (med kakuminal *l*) 'svull', som neppe skal førast til *bolde*. Men også det er Christiansens oppslag. *Forbjælla* 'overraska, forvirra; skremt' er derimot eit merkeleg val av oppslagsform. Uttalen er med *a* og ustemt palatal *l*, og både Christiansen og *Norsk Ordbok* normerer til *forbjatla*. *Hogsken* burde hatt oppslagsform *hosken*, som Christiansen og *Norsk Ordbok*. *Gnallande* (forsterkande adverb) er for hardhendt normering av *gnalla* («gnalla hard») og bryt med Christiansens eiga normering. Avleiingar på *-an* er vanskelege å halde frå presens partisipp på *-ande*. Likevel må ein stusse på den ulike behandlinga av oppslaga *drikkande* 'drikkegilde' og *flødan* 'stigande sjø'.

Tilvisingar må sjølvsagt avgrensast på grunn av formatet. Men det kunne og burde vore noko fleire enn det er no. Til dømes burde ein kanskje vurdert å ha ei tilvising frå uttaleforma «dekedan»

til den elles korrekte oppslagsforma *dykdalb* 'påle i hamn'. Under artikkelen *gueridon* 'trestativ til å sette lys eller tranlampe på' burde det vore opplyst om uttaleforma «kjerridon». Kanskje dette heller burde vore oppslagsforma, med etymologisk opplysning om fransk *gueridon*. Dette ordet er elles belagt frå mange fleire stader i Nord-Noreg enn Nord-Rana, slik det er oppgjeve. Som vi har sett, finst det nyttige tilvisingar mellom (nær)synonym. Det kunne godt vore fleire, til dømes mellom *garnsnur* og *varplund* (begge: 'det å la rokken gå mot høgre').

6. Samla vurdering og konklusjon

Dei kritiske merknadene ovanfor må ikkje få skyggje for det faktum at vi har for oss eit stort og verdifullt produkt, og ein milestolpe i nordnorsk leksikografi. I eit så stort arbeid gjort av ein einskild person vil det alltid finnast ting å peike på. Generelt må ein seie at Orvik gjennomfører konsekvent og nøyaktig den systematikken han sjølv legg opp til, noko som viser at han er meir enn ein vanleg amatørleksikograf.

Sjølvsagt kan ein også halde fram at tilfanget finst i *Norsk Ordbok*, sjølv om det i mindre grad er komme med i strekket *a–h*, og mindre dess tidlegare i alfabetet ein er. Dette har likevel ein lett for å «drukne» i den store samanhengen. I *Nordnorsk ordbok* blir både det typiske og det meir spesifikt nordnorske ordforrådet løfta fram og gjort tilgjengeleg på ein grei og folkeleg måte. Dei litterære sitata gjev liv til framstillinga. Tidsmessig kan ordmaterialet grovt sett plasserast i det tradisjonelle bonde- og fiskarsamfunnet inn mot overgangen til industrisamfunnet. Samlinga avspeglar eit sjølvbergingssamfunn «prega av fiske og mykje vêr og vind», slik Orvik blir sitert i det nemnde intervjuet i avisa Norsk Tidend. Uttrykk for dette er interessante godbitar som *bubbelurvad* 'tønneband med not for å fange sjøsniglar som skal brukast til agn' og

fiskehentardag 'dagen/dagane då tørrfisken blei tatt ned frå hjell og selt'. For vêr, vind og straum er ordforrådet spesielt velutvikla, med presise nemningar som *dumlebåre* 'bølge utan spesiell retning i ly av ein odde' og *fjordræs* 'skarp landvind, oftast frå aust, som blæs ut fjordane'. La oss avslutte med det einestående og herlege *frygdesnelle* 'fuglen lauvsongar' frå Sømna. Det er ei frygd å bla i denne boka, og grunn til å takke Ove Arild Orvik for å ha fått fram i lyset dette tilfanget, som er ei skattkiste både for folk frå Nord-Noreg og for andre.

Litteratur

Ordbøker, korpus og digitale ressursar

Dalen, Arnold & Tor Erik Jenstad (1997): *Trønderordboka*. Trondheim: Tapir Forlag.

Dalen, Arnold & Tor Erik Jenstad (2002): *Trønderordboka*. 2. utgåve. Trondheim: Tapir Akademisk Forlag.

Ellingsve, Eli J. (2012): *1000 ord i nord*. Stonglandseidet: Nordkalottforlaget.

Hveding, Johan (1968): *Håløygsk ordsamling*. Bodø: Nordland Boktrykkeri AS.

Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet (1966–2016). Oslo: Det Norske Samlaget. <no2014.uib.no/perl/ordbok/no2014.cgi> (april 2024).

Orvik, Ove Arild (1989a): *Frænsk ordbok*. Elnesvågen. Eiga utgjeving.

Orvik, Ove Arild (1989b): *Midsundsk ordbok*. Elnesvågen. Eiga utgjeving.

Orvik, Ove Arild (1997): *Vesterålsk ordbok*. Sortland: Målmann Forlag.

Orvik, Ove Arild (2003): *Ord og dialekt i Vesterålen*. Sortland: Målmann Forlag.

Orvik, Ove Arild (2006): *Ta va fole! Ord og uttrykk frå Romsdal*. Målmann Forlag.

Orvik, Ove Arild (2017): *Kystspråk. Ord og uttrykk om hav, båt og fisk i Vesterålen, Romsdal og Aust-Island*. Eiga utgjeving.

Schytte, Erik Gerhard (1807): *Nogle faa rare norske Ord, efter den Dialekt, som i Lofoten's Fogderi i Nordlandene er brugelig*. Det Skandinaviske Litteraturselskabs skrifter 3. Særprent av *Håløygminne*. Svorkmo prenteverk, Svorkmo 1940.

UiB samlingar = Universitetet i Bergens samlingsdatabasar. <usd.uib.no/perl/search/search.cgi> (mars 2024).

Vidsteen, Chr. (1900): *Ordbog over Bygdemaalene i Søndhordland*. Bergen: John Griegs Bogtrykkeri.

Annan litteratur

Christiansen, Hallfrid (1933): *Gimsøymålet. Fonologi og orddannelse*. Oslo: Det Norske Videnskabs-Akademi.

Christiansen, Hallfrid (1946–48): *Norske dialekter*. Oslo: Tanum.

Karlsen, Knut E. (1997): *Hallfrid Christiansen: Nordnorsk språkforskar mellom to tradisjonar. Ei faghistorisk utgreiing med vekt på "Gimsøy-målet"*. Hovudoppgåve. Universitetet i Tromsø.

Karlsen, Knut E. (2011): Hallfrid Christiansen – språkforskaren frå Lofoten. I: *Lokalhistorisk magasin* 2/2011. Trondheim. 4–7.

Knudsen, Vemund N. (2024): 7000 nordnorske ord. I: *Norsk Tidend* 1/2024.

Tor Erik Jenstad
Forskar, dr.art.
Universitetet i Bergen, Språksamlingane
Antonie Løchens v 6
NO-7020 Trondheim
tor.jenstad@uib.no

MINDEORD

Minnesord om Lars Holm

Lars Svensson

En enastående språkman, lexikografihistoriker och textutgivare har gått bort. Lars Holms författarskap kännetecknas av stark självständighet, rik materialsamling, omfattande arkiv- och bibliotheksstudier, akribi, uppslagsrikedom, klokt omdöme, stor förtrogenhet med källor och vidsträckt kulturhistorisk beläsenhet. Därtill var Lars Holm en skicklig handskriftsläsare och en lysande stilist med spänst i språket. Han var en flitig skribent och hans forskargärning är värd all beundran, inte minst med tanke på att den delvis utfördes parallellt med en 40-årig lärartjänst.

2018 tilldelades Lars Holm Blomska stipendiet av Svenska Akademien. Priset tillfaller ”personer, vilka på ett mera framträdande sätt gjort sig förtjänta om svenska språkets rykt och ans”.

Utbildning och personlig bakgrund

Lars Holm föddes den 27 maj 1934 som den yngste sonen till Wilhelm och Tora Holm. Föräldrarna var båda folkskollärare. Efter studentexamen vid Malmö högre allmänna läroverk för gossar 1953 följde militärtjänst i Uppsala. Höstterminen 1954 skrevs han in vid Lunds universitet. 1957 blev han fil. kand., 1959 fil. mag. (litteraturhistoria med poetik, nordiska språk, tyska, teoretisk filosofi och pedagogik) och 1986 fil. dr. i nordiska språk vid Uppsala universitet.

Efter provårstjänstgöring arbetade han under 1960-talet som gymnasielärare i svenska och tyska och för Skolöverstyrelsen med bl.a. utveckling av undervisningsformen grupparbete, som han

var tidig med att införa. 1970–1977 tjänstgjorde han vid Lärarhögskolan och Katedralskolan i Uppsala. Efter hemkomsten till Skåne 1977 var han anställd som gymnasielärare i svenska och tyska till 1990 vid Linnéskolan i Hässleholm och från 1990 till pensioneringen 1998 vid Spyken i Lund.

Lars Holm hade från slutet av 70-talet utöver sin lärartjänstgöring dels ett flerårigt samarbete med Tor Hultman och Åke Pettersson vid Lärarhögskolan i Malmö med att ta fram bedömningsmallar för nationella prov i svenska, dels en fast forskarplats på Svenska Akademiens ordboksredaktion i Lund som han under en följd av år kontinuerligt besökte för studier i Swedbergs ordbok (Skara-handskriften från ca 1720).

Sin blivande hustru Karin träffade han hos gemensamma vänner 1957 och två år senare gifte de sig. Paret fick två barn: Martin och Lisa.

Swedbergforskare

Tidigt kom Lars Holm att intressera sig för språkmannen, lexikografen, psalmisten och Skarabiskopen Jesper Swedbergs outgivna[1] *Swensk Ordabok* från ca 1725, endast sporadiskt beaktad i språkvetenskaplig litteratur. Den blev hans främsta forskningsobjekt och följde honom genom hela livet. Ordboken, som annars bara föreligger i fem bevarade handskrifter, är den första som med förklaringar på latin redovisar såvitt möjligt den samtida svenskans ordförråd i sin helhet inklusive lågspråkliga ord som könsord, invektiv och svordomar.

Lars Holm påbörjade under Uppsalaåren sina doktorandstudier för professor Gun Widmark, som blev hans stöd och in-

1 Swedberg fick aldrig se sin ordbok tryckt. Det svensk-latinska registret i Petrus Schenbergs *Lexicon latino-svecanum* (1739) bygger dock helt på Swedbergs ordboksartiklar (som emellertid gallrats hårt).

spirerande handledare fram till disputationen. Han har med stor tacksamhet flerstädes framhållit att det var hennes förtjänst att han 1986 kunde framlägga en sammanläggningsavhandling med titeln *Jesper Swedbergs Swensk Ordabok – bakgrund och tillkomsthistoria*. Avhandlingen behandlar ordbokens tillkomsthistoria och Swedbergs fruktlösa försök att få den tryckt, de bevarade handskrifterna och ordbokens tryckta källor, föregångare och möjliga förebilder.

Holms studier i *Swensk Ordabok* resulterade utöver doktorsavhandlingen i ett flertal artiklar och föredrag om innehållet i den (bl.a. om ”fula ord”, polysemi, mat och dryck, sjukdomar) och kulminerade 2009, 23 år efter disputationen, i en mönstergill textedition av Swedbergs *Swensk Ordabok* med Uppsala-handskriften som textkodex och med tillägg och rättelser ur övriga handskrifter. Uppsala-handskriften är en avskrift utförd 1724–1725 av sonen Jesper Swedenborg[2] och försedd med egenhändiga tillägg av Swedberg senior. Den vackra utgåvan, som inleds med en introduktion som bl.a. innehåller en presentation av Jesper Swedbergs liv och karriär, redigeringsprinciper och en värdefull 40-sidig förteckning över de källor och referenser som anges i ordboken, får betraktas som en kulturgärning av stora mått.

Texutgivare (utöver *Swensk Ordabok*)

Hiob Ludolfs *Dictionarium Sueco-Germanicum*

År 2017 utgav Lars Holm tillsammans med lärarkollegan Gunnar Graumann, romanist och latinist, med inledning och kommentarer en svensk-tysk ordbok, tidigare inte utgiven – *Dictionarium Sueco-Germanicum*. Ordboken innehåller drygt 2100 svenska uppslagsord omfattande hela alfabetet med översättning till tyska,

2 Släkten Swedberg adlades 1719.

ofta också till latin, holländska, franska och flera andra språk. Ordboken sammanställdes 1649–1651 (med senare tillägg) för eget bruk av den tyske mångspråkkunnige orientalisten och diplomaten Hiob Ludolf (1624–1704), som omkring mitten av 1600-talet vistades som informator hos ämbetsmannen Schering Rosenhane på godset Torp i Södermanland. *Dictionariet* föreligger i två handskrifter, båda tillhöriga Uppsala universitetsbibliotek. Utgåvan av Ludolfs *Dictionarium* har gjorts efter en CD-kopia framställd på Uppsala universitetsbibliotek.

Ludolfs *Dictionarium* redovisar främst innehållsord, inte formord. Definitionerna anges dels med ekvivalenter, synonymer, på olika målspråk (tyska och latin vanligast) eller med en deskriptiv omskrivning på tyska eller latin. Vissa uppslagsord saknar översättning. Av stort intresse är ett tjugotal uppslagsord som svällt ut till ”essäer” om ämnen som folktro, svenskt djurliv, fiske, seder och bruk, nöjen och festligheter. De encyklopediska artiklarna är skrivna på svenska, tyska eller latin, ofta i blandning. Ordboksartiklar som omfattar minst tre rader har översatts till nutida svenska, som står i hakparentes efter respektive artikel. Vissa artiklar innehåller lågspråkliga ord, språkprov och uttalsuppgifter, vilket vittnar om att Ludolf var långt före sin tid som lexikograf. Värt att notera är att ett trettiotal uppslagsord visar sig vara äldsta belägg.

Efter ordbokstexten följer två tematiska avsnitt. I det första kommenterar Gunnar Graumann med utgångspunkt i uppslagsorden dagligt liv i 1600-talets Sverige. I det andra avsnittet redogör Lars Holm ingående för Ludolfs språkliga repertoar. Även Ludolfs källor och referenser anges.

Den gedigna utgåvan är särskilt välkommen, eftersom ordboken trots sitt ringa omfång är ett värdefullt tillskott i 1600-talets magra ordboksflora i Sverige.

35 brev kring en resa – korrespondens mellan Jacob Jonas Björnståhl och medlemmar av familjen Rudbeck, 1768–1778.

Holm gav 2019 ut *35 brev kring en resa* som är en kommenterad brevutgåva av hög klass. Alla breven utom fyra tillhör Lunds universitetsbibliotek. Holm har bokstavstroget skrivit av breven efter original och/eller fotokopior. Efter varje brev följer språkliga och sakliga förklaringar och kommentarer. Särskilt värdefullt är att personerna som resenärerna träffat i möjligaste mån identifieras. I kommentarerna tar Holm upp innehållet i breven ur olika aspekter. Den språkligt och kulturhistoriskt värdefulla boken är därtill en njutbar läsupplevelse.

Ordbokssamlare

Lars Holm var en hängiven ordbokssamlare. Han donerade sin omfattande ordbokssamling innehållande drygt hundra titlar till institutionen för nordiska språk vid Göteborgs universitet.

I artiklarna ”Om en Calepinus 1616 med svenska som tolfte språk” (2002)[3] och ”En svensk Calepinus från 1600-talet” (2003)[4] berättar han om sitt samlarintresse och om sitt exemplar av Calepini *Dictionarivm Vndecim Lingvarum* tryckt 1616 i Basel. Den italienske munken och lexikografen Ambrogio Calepino (ca 1440–1510) – mest känd under den latinska namnformen *Ambrosius Calepinus* – gav ut sitt *Dictionarivm* 1502. Ordboken blev mycket populär och spreds över hela Europa i en mängd upplagor, eftersom den inte bara var en ordbok utan också ett uppslagsverk. I upplagorna kan det latinska uppslagsordet ha mellan två och elva ekvivalenter. Lars Holms Baselupplaga är unik. Den innehåller nämligen inte bara de elva gängse ekvivalenterna utan dessutom – som tolfte språk – svenska ekvivalenter tillskrivna i marginaler-

3 I: *Alla ord är lika roliga.* Festskrift till Lars Svensson […] 2002.
4 I: *Språk och stil* 13. Ny följd 2003.

na och mellan uppslagsorden. Skrivaren är anonym. Handstilen är vårdad och dateras av Holm till sista tredjedelen eller fjärdedelen av 1600-talet. Den svenska lexikografen var en skicklig latinist men kunde också främmande levande språk. Han korrigerade och kompletterade flera tryckta ordboksartiklar.

Det svenska ordförrådet är trots luckor rikt företrädd i Lars Holms Calepinus-upplaga. Ca 75 % av uppslagsorden har översatts till svenska efter *Lincopensen* 1640[5], Wolimhaus 1649[6] och *Hamburgensen* 1700[7] som förebilder eller källor.

Ämnesmetodiskt författarskap

Utöver sin lexikografiska forskning och lärartjänat ägnade Lars Holm sig också åt ämnesmetodiskt författarskap. Tillsammans med Kent Larsson gav han 1976 ut en främst för gymnasieskolan avsedd elementär språklära, *Svenska meningar*, som bygger på Paul Diedrichsens positionsmodell. Boken kom ut 1980 i en andra upplaga.

Björnståhlforskare

Som flitig resenär blev Lars Holm tidigt fascinerad av orientalisten, polyglotten och reseskildraren Jacob Jonas Björnståhls resa i Europa 1767–1779, som han skrev flera skrifter om. Här skall bara nämnas *Brev till Jacob Jonas Björnståhl om ett besök i Konstantinopel/Istanbul* (1984) och *35 brev kring en resa* – hans sista bok (se **Textutgivare**).

5 *Lincopensen* 1640 = [Grubb, Nicolaus &] Gothus, Jonas Petri *Dictionarium Latino-Sveco-Germanicum* [...] Lincopiæ.
6 Wolimhaus 1649 = Wolimhaus, Johannes, *Syllabus* […]. Holmiæ.
7 *Hamburgensen = Novum dictionarium Latino-Sveco Germanicum, Sveco-latinum et Germanico-Latinum* [...] Hamburg & Stockholm.

Personlighet

Till det yttre var Lars Holm en stark personlighet, en trivsam sällskapsmänniska med god berättarkonst, mycket allmänbildad och humoristisk, rakryggad och bestämd. Han var med i sällskapet ”lunchgubbarna”, en grupp pensionerade lärarkolleger, som under gemytligt samkväm brukade äta en gemensam måltid om tisdagarna. Och om torsdagarna sågs han – så länge hälsan tillät det – ofta i SAOB-redaktörernas pensionärsgäng på restaurang Finn Inn.

Lars var en stor naturvän. I yngre år var han aktiv fågelskådare och ringmärkare. Han gillade vandring, bad, bär- och svampplockning. Familjen tillbringade därför under många år somrarna i sin sommarstuga i Krogshult i norra Skåne med närhet till skog och sjö.

I många av sina resor har han bokstavligen följt Björnståhl i spåren och använt hans resebrev som resehandbok. Som resenär fick han nytta av sina goda språkkunskaper i tyska, engelska och franska. Han talade även hjälpligt italienska och nygrekiska.

Lars var för mig inte bara en lexikografikollega utan en sann vän med värmande vänskap. De sista åren bodde han på ett äldreboende. Han blev alltid glad vid besök. Han grämde sig dock över ett uppslag han inte kunde ge sig i kast med, nämligen Swedbergs stora fågelintresse!

Lars Svensson
fil. dr, docent
Dammfrivägen 48 A
SE-217 63 Malmö
laurentiusgerd@gmail.com

MEDDELELSER

Nyt fra bestyrelsen for Nordisk Forening for Leksikografi

Thomas Widmann

Nordisk Forening for Leksikografi (NFL) arrangerer symposier hvert år og konferencer hvert andet år.

Symposierne samler et mindre antal specialister fra de forskellige nordiske lande, som holder foredrag om og diskuterer et bestemt tema; foredragene danner basis for udarbejdelsen af artikler der udgives i tidsskriftet *LexicoNordica*. Symposiernes format giver rig mulighed for fordybelse i leksikografiske emner, for udveksling af erfaringer og for dannelse af netværk.

Det 31. *LexicoNordica*-symposium 2024 fandt sted på Voksenåsen i Oslo d. 15.-17. februar 2024. Temaet var *Hvilket datamateriale bygger nordiske ordbøger på? Skævheder, udfordringer og løsninger*, og bidragene herfra er publiceret i den tematiske del af dette bind af *LexicoNordica*.

Konferencerne henvender sig til et bredere publikum, og målet er at samle forskere og andre leksikografiinteresserede fra hele Norden. Hver konference har et overordnet tema, men der er generelt plads til alle leksikografiske foredrag og posterpræsentationer. Konferencerne afholdes på skift i landene i Norden, og de arrangeres af et eller flere fagmiljøer der. Skriftlige versioner af konferencebidragene udgives i *Nordiske Studier i Leksikografi*. Konferencerapporten fra konferencen i Bergen udkommer i slutningen af 2024 eller i begyndelsen af 2025.

Planlægningen af næste års symposium er i fuld gang. Det har temaet *Nordiske ordbøger – opdatering, udvikling og tilgængeliggørelse* og vil finde sted på Voksenåsen i Oslo d. 23.-25. januar 2025.

Den 18. konference om leksikografi i Norden er også ved at fal-

de på plads. Den vil blive afholdt i Vejle d. 21.-23. maj 2025 (ikke i Odense som oprindeligt annonceret). Flere oplysninger kan findes på konferencens hjemmeside: <nfl-nsn2025.dsn.dk/>.

Referater fra NFL's generalforsamlinger og information om bl.a. kommende symposier, konferencer og publikationer kan findes på NFL's hjemmeside <nordisk-leksikografi.com> og på NFL's Facebook-side (søg på "Nordiska Föreningen för Lexikografi – NFL"). Medlemmer af NFL modtager også et nyhedsbrev med aktuel information. Samtlige numre af *LexicoNordica* og *Nordiske Studier i Leksikografi* publiceres digitalt på <tidsskrift.dk> i samarbejde med Dansk Sprognævn.

De administrative opgaver i NFL varetages af bestyrelsesmedlemmerne og af *LexicoNordicas* redaktion, som består af to hovedredaktører og fem landsredaktører. Betaling for medlemskab foregår digitalt, således at hvert medlem betaler direkte via NFL's hjemmeside, og de registreres i et fælles medlemsregister. Man kan kontakte bestyrelsen direkte via e-mail: nordisk.lexikografi@gmail.com. Den næste generalforsamling vil finde sted på konferencen i Vejle i 2025.

NFL har grundlæggende en sund økonomi, men der er fortsat behov for ekstern økonomisk støtte for at kunne fortsætte med at arrangere resursekrævende initiativer som symposier og konferencer. Herudover bidrager værtsinstitutionerne med arbejdstid og andre resurser.

NFL takker for alle eksterne økonomiske bidrag til konferencer, symposier og publicering af konferencerapporterne – ikke mindst Nordplus Nordens Sprog. Uden denne støtte ville disse aktiviteter ikke være mulige.

Mange personer bidrager også med deres arbejdsindsats, ikke mindst i forbindelse med udgivelse af *LexicoNordica* og konferencerapporterne – også en stor tak til dem.

NFL's bestyrelse 2023-25

Formand: Thomas Widmann
Næstformand: Margunn Rauset
Kasserer: Pär Nilsson
Sekretær: Kristin M. Magnussen
Medlem: Caroline Sandström
Suppleanter: Einar Freyr Sigurðsson og Carina Nilstun

Thomas Widmann
seniorkonsulent
Dansk Sprognævn
Adelgade 119 B
5400 Bogense
tw@dsn.dk

REDAKTIONSANVISNINGER

1. LexicoNordica udkommer hvert år i november. Tidsskriftet indeholder leksikografiske bidrag som er skrevet på et af følgende nordiske sprog: dansk, finsk, færøsk, islandsk, norsk (bokmål eller nynorsk) og svensk. Bidrag på engelsk kan også optages hvis særlige forhold taler for det.

2. **Bidrag** sendes til det medlem af redaktionskomitéen som repræsenterer bidragyderens land:

   - Danmark: Liisa Deth Theilgaard, Helgesensgade 1, 2. tv., DK-2100 København Ø. <liisa.theilgaard@gmail.com>.
   - Finland: Caroline Sandström, Institutet för de inhemska språken, Hagnäskajen 6, FI-00530 Helsingfors. <caroline.sandstrom@sprakinstitutet.fi>.
   - Island: Helga Hilmisdóttir, Árni Magnússon-instituttet for islandske studier, Arngrímsgötu 5, IS-107 Reykjavík. <helga.hilmisdottir@arnastofnun.is>.
   - Norge: Kjetil Gundersen, Erika Nissens gate 7, NO-0480 Oslo. <kjetil.gundersen@sprakradet.no>.
   - Sverige: Louise Holmer, Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet, Box 200, SE-40530 Göteborg. <louise.holmer@svenska.gu.se>.

   Fristen for aflevering af bidrag er **den 1. april** hvis artiklen skal kunne trykkes i det nummer af tidsskriftet som udkommer samme år. Bidraget indleveres digitalt i både tekstbehandlingsformat (.docx) og i PDF-format. Dette gælder også evt. reviderede versioner.

3. **Illustrationer** der skal medtages i artiklen, indsættes i manuskriptet og vedlægges som separate billedfiler, helst i JPG-format og minimum 300 ppi. Tabeller udført i Word indsættes i manuskriptet og kræver ikke særskilt billedfil. Der refereres

eksplicit til figurer og tabeller undervejs i teksten, fx "jf. figur 1", "se tabel 3" e.l., men IKKE "se følgende figur/tabel". De vedlagte billedfiler nummereres tydeligt og i overensstemmelse med den rækkefølge og den angivelse som bruges i manuskriptet. Dette gælder også evt. reviderede versioner af artiklen, hvor antal og rækkefølge af illustrationer/billedfiler (og tabeller) kan være ændret.

4. Bidraget skal forfattes i LexicoNordicas **stilark** (.docx), der kan rekvireres ved henvendelse til redaktionen. Stilarket er på forhånd opsat med korrekte marginer og forhåndsdefinerede typografier. Når man har modtaget stilarket, tages der en kopi af stilarket, som herefter omdøbes efter følgende model: "[forfatter(e)]_LN30". Artiklen udfærdiges herefter i det omdøbte stilark, og der vælges kun foruddefinerede typografier fra stilarket. Ved evt. problemer eller tvivlsspørgsmål rettes henvendelse til redaktionen.

5. **Manuskriptet** indledes med titel på artiklen og forfatterens navn. For tematiske og ikke-tematiske bidrag følger et **abstract** på engelsk på op til 10 linjer og dernæst selve artiklen, som opdeles i afsnit. Bidrag på engelsk tilføjes et abstract på dansk, norsk eller svensk. **Afsnit** nummereres efter følgende model: **1.**; **1.1.**; **1.1.1.** (højst tre niveauer; ved henvisninger i teksten udelades slutpunktum, fx "jf. afsnit 2.1"; der henvises eksplicit til afsnit i teksten, fx "jf. afsnit 2.4", "se videre afsnit 4", men IKKE "se ovenstående/følgende afsnit"). Bidraget afsluttes med angivelse af forfatterens navn, titel samt post- og e-mailadresse. Bidrag kan normalt have et omfang på højst 20 sider inkl. litteraturliste. Jf. i øvrigt stilarket mht. typografi, blanke linjer o.l.

6. **Citater**: Kortere citater (op til 3 linjer) bringes som en del af teksten med dobbelte anførselstegn omkring, mens længere ci-

tater eller fremhævelser af større vigtighed gives i et afsnit for sig selv **uden** anførselstegn (vælg typografien LN-citat i stilarket).

7. Vi anbefaler en meget tilbageholdende brug af **fodnoter**. Evt. nødvendige noter gennemnummereres i teksten med højtstillet angivelse uden parentes. Der anvendes fodnoter, ikke slutnoter.

8. **Litteraturhenvisninger** foretages i teksten efter følgende model:

   som det fremgår af Herbst (2009)
   som det fremgår af Borin & Forsberg (2011:18)
   (se Herbst 2009:158ff.)
   (jf. Borin & Forsberg 2011:49–52)

   For kilder med tre eller færre forfattere anføres efternavnene på alle forfattere i henvisningen, fx "Gudiksen 2009", "Gudiksen & Hovmark 2020", "Gudiksen, Hovmark & Monka 2015". For kilder med fire eller flere forfattere anføres kun det første efternavn efterfulgt af "et al.", fx "Gudiksen et al. 1999". Forfatternavnene adskilles af komma, undtagen de to sidste navne, som adskilles af "&" (tilsvarende hvis der kun er to forfatternavne). I litteraturlisten skrives alle forfatternavne imidlertid ud, også hvis der er flere end tre forfattere.

   I den løbende tekst angives IKKE hele internetadresser, men et forfatternavn eller en angivelse af titlen på internetbidraget, som herefter bruges i litteraturlisten. I litteraturlisten angives internetadresser uden "http(s)://" eller evt. "www." og uden understregning, men omgivet af < >, og måned og årstal for sidste tidspunkt for opslag på adressen anføres i parentes, fx "<ordnet.dk/ddo> (april 2021)". Hvis der er oprettet en unik

reference til en digital kilde (doi:), anføres denne i litteraturlisten. Internetadresser, som har vundet indpas som titler af proprial karakter (fx "svenska.se"), kan undtagelsesvis anføres som sådan i den løbende tekst.

9. **Særlige angivelser**: Vær meget tilbageholdende med brug af **fede** typer; **sprogeksempler** markeres med kursiv, fx: "ordet *ungkarl* har synonymet *alenemand*"; **betydninger** af sproglige enheder angives ved hjælp af enkelte anførselstegn, fx: 'en ugift mand'; dobbelte anførselstegn bruges ved **citater** eller **forbehold**, fx: De er vokset op i de "glade" tressere. Tegnsætningsreglerne, bl.a. for brug af komma, tankestreg, bindestreg (i betydningen 'fra ... til'), er forskellige i de nordiske lande, og forfatterne skal følge reglerne for det sprog som bruges i artiklen. Titler på ordbøger o.l. sættes i kursiv, fx "*Den Danske Ordbog* er ..." og gives evt. en introduktion første gang titlen nævnes. Hyppigt anvendte titler kan evt. erstattes af en forkortelse, der indsættes i parentes første gang titlen nævnes, og som herefter anvendes, fx "*Den Danske Ordbog* (DDO) er ...".

10. **Litteraturangivelser**
    I litteraturlisten anføres forfatternavne efter følgende model:

    Gudiksen, Asgerd ([årstal])
    Gudiksen, Asgerd & Henrik Hovmark ([årstal])
    Gudiksen, Asgerd, Henrik Hovmark & Malene Monka ([årstal])

    Ved mere end ét bidrag fra samme forfatter anføres bidragene i kronologisk rækkefølge. Alle bidrag hvor en person er eneforfatter anføres før bidrag hvor samme person er førsteforfatter sammen med andre forfattere, fx "Nielsen 2020, Nielsen 1999, Nielsen & Krogh 2010".

I tilfælde af en længere litteraturliste kan den inddeles i to dele i lighed med nedenstående eksempel. Hvad angår angivelser som *red.*, *eds.*, *Hrsg.*, anbefales det så vidt muligt at bruge originalsproget. Det vigtigste er dog konsekvens inden for samme liste.

I tvivlstilfælde rettes henvendelse til redaktionen.

Litteratur

Ordbøger, korpusser og digitale resurser

ALD (1948) = A.S. Hornby, E.V. Gatenby & H. Wakefield: *A Learner's Dictionary of Current English*. London: Oxford University Press.

BÍN = *Beygingarlýsing íslensks nútímamáls*. Kristín Bjarnadóttir (red.). Árni Magnússon-instituttet for islandske studier. <bin.arnastofnun.is> (marts 2021).

COBUILD (1987) = *Collins* COBUILD *English Language Dictionary*. Editor in Chief: John Sinclair, Managing Editor: Patrick Hanks. London/Glasgow: Collins.

DDO = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ddo> (april 2021).

Italiensk-Dansk Ordbog (1999). Knud Andersen & Giovanni Mafera. København: Gyldendal.

Jarvad, Pia (1999): *Nye Ord. Ordbog over nye ord i dansk 1955-1998*. København: Gyldendal.

LBK = Leksikografisk bokmålskorpus. Tekstlaboratoriet, Institutt for lingvistiske og nordiske studier, Universitetet i Oslo. <tekstlab.uio.no/glossa2/bokmal> (august 2020).

Norstedts stora engelska ordbok (2000). Stockholm: Norstedts.

Oxford-Hachette French Dictionary (1994). Oxford: Oxford University Press.

Risamálheildin (2017-2018). Stofnun Árna Magnússonar í íslenskum fræðum. <malheildir.arnastofnun.is> (februar 2020).

Språkbanken Text. <spraakbanken.gu.se/> (marts 2021).

Svenska.se = Svenska Akademiens ordboksportal. <svenska.se/> (april 2021).

Anden litteratur

Delvin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. I: *Proceedings of the 2019 Conference of NAACL: Human Language Technologies*, Volume 1. Minneapolis, Minnesota: Association for Computational Linguistics. 4171–4186.

Faarlund, Jan Terje, Kjell Ivar Vannebo & Svein Lie (1997): *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Haiman, John (1980): Dictionaries and Encyclopedias. I: *Lingua* 50, 329-357.

Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir (2011): ISLEX – en flersproget nordisk ordbog. I: Birgit Eaker, Lennart Larsson & Anki Mattisson (red.): *Nordiska studier i lexikografi* 11. Lund: Nordisk förening för lexikografi. 353–366.

Lakoff, George & Mark Johnson (1980): *Metaphors we live by.* Chicago/London: The University of Chicago Press.

Mugdan, Joachim (1985a): Grammatik im Wörterbuch: Wortbildung. I: Herbert Ernst Wiegand (Hrsg.): *Studien zur neuhochdeutschen Lexikographie IV*. Hildesheim/Zürich/New York: Olms. 237-308.

Nikula, Kristina (2012): Samspelet mellan text och bild i enspråkigt svenska ordböcker. I: *LexicoNordica* 19 (dette bind).

Nordenstorm, Leif (2017): Tro och tradition enligt Svenska Akademins Ordlista. I: *Svensk kyrkotidning* 11/2017. <svenskkyrkotidning.se/recension/tro-och-tradition-enligt-svenska-akademins-ordlista/> (april 2021).

Nørstebø Moshagen, Sjur, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen & Per Langgard (2022): *Report on the Nordic Minority Languages*. doi:10.1163/9789004298507.

NRG = *Norsk referansegrammatikk*, se Faarlund et al. (1997).

11. LexicoNordica udkommer både som trykt tidsskrift og i en **digital udgave** på open access-platformen Tidsskrift.dk. Ved indsendelse af et bidrag til redaktionen erklærer forfatterne sig derfor indforstået med både en trykt udgave og en digital udgave på open access-platformen Tidsskrift.dk.