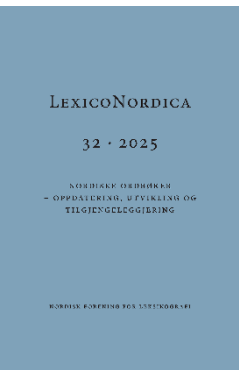


LexicoNordica

Titel:	Gamle historiske sprogarkiver og deres værdi i nutiden	
Forfatter:	Ásta Svavarsdóttir & Helga Hilmisdóttir	
Kilde:	LexicoNordica 32, 2025, s. 155-175	
URL:	https://tidsskrift.dk/lexn/issue/archive	

© 2025 LexicoNordica og forfatterne

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Gamle historiske sprogarkiver og deres værdi i nutiden

Ásta Svavarsdóttir & Helga Hilmisdóttir

Thus far, archives of the Lexicographic Institute in Reykjavík, compiled in the 20th century, have been the only lexicographical resources available for older Icelandic. With new historical corpora in the making the situation has changed. In this article, we compare the two types of resources from different perspectives and conclude that archives and corpora have different strengths and weaknesses and thus supplement each other. Together, both the archives and corpora can continue to serve the needs of lexicographers and linguists in the future.

1. Indledning

Orðabók Háskólans (OH) blev stiftet ved Islands universitet tidligt i 1940'erne med det formål at forberede, redigere og udgive en videnskabelig, historisk ordbog over islandsk. Fokus blev sat på perioden fra midten af 1500-tallet, da de første islandske bøger blev trykt, til nutiden. De første årtier var arbejdet koncentreret omkring indsamling af materiale til ordbogen, og det resulterede i traditionelle seddelsamlinger af forskellig art og størrelse, den største af dem det såkaldte *Ritmálssafn* (ROH), et skriftsprogsarkiv med citater især fra trykte værker (se Svavarsdóttir 2024, Pind et al. 1993).

Digitalisering blev indført ganske tidligt på OH (nu en del af Árni Magnússon-Instituttet for Islandske Studier (ÁMI)), dvs. i 1980'erne. Digitaliseringen blev, i begyndelsen, først og fremmest betragtet som en måde at udføre en effektiv og systematisk redigering af materialet i OH's arkiver. Tidligt viste det sig at den almindelige digitalisering af trykkeriernes arbejds gange gav mulighed

for at få fat i en ny type materiale, dvs. sammenhængende tekster i digital form, og dermed opbygge et nyt grundlag for leksikografisk arbejde (se fx Pind et al. 1993:5). Fra trykkerierne kom der dog først og fremmest nye, samtidige tekster selv om der indimellem også fulgte enkelte genudgivelser af ældre værker. Man så derfor arkiverne fremadrettet som det vigtigste grundlag for en historisk beskrivelse af ordforrådet, mens opbygning af digitale, søgbare tekstsamlinger hovedsagelig ville erstatte traditionel excerpering når det gjaldt nyere sprog.

Nu er det blevet realistisk at digitalisere gamle tekster med scanning og OCR-behandling og dermed at opbygge store historiske sprogkorporer med tekster fra forskellige perioder og genrer. Og så er det på tide at gøre sig overvejelser om de gamle arkivers status i nutid og fremtid i forhold til nye digitale sprogresurser (jf. diskussion om forholdet mellem seddelsamlinger og korpusser i Hovmark 2012, 2023 og Trap-Jensen 2014).

Formålet med denne artikel er at sammenligne de gamle sprogarkiver med nye eller planlagte digitale sprogkorporer og diskutere deres fordele og ulemper vis-a-vis hinanden. Hovedspørgsmålene er de følgende: 1) Hvilken slags informationer findes der i de forskellige typer sprogresurser, og hvordan er forholdet mellem dem? 2) Gør udviklingen af digitale korpusser de gamle leksikalske arkiver forældede og overflødige, eller vil de gamle arkiver beholde deres værdi for fx leksikografer og sprogforskere? 3) Hvordan kan eller skal de gamle samlinger videreudvikles, fx med hensyn til (yderligere) digitalisering og analyse?

Dispositionen er som følger: I afsnit 2 beskrives ÁMI's vigtigste sprogresurser med hensyn til en udførlig, historisk beskrivelse af islandsk ordforråd, både de traditionelle sprogarkiver og nyere digitale sprogkorporer. Afsnit 3 omfatter en sammenligning af resurserne, som viser at hver og én har sine fordele og ulemper i forskellige sammenhænge og således supplerer hinanden. Afsnit 4 indeholder en kortfattet sammenfatning og konklusioner.

2. Leksikografiske sprogresurser for islandsk

2.1. Sprogarkiver

2.1.1. Seddelsamlinger

De gamle sprogarkiver fra OH indbefatter forskellige store og små samlinger. Vi fokuserer her på hovedsamlingerne, *Ritmálssafn Orðabókar Háskólans* (ROH), et skriftsprogsarkiv, og *Talmálssafn Orðabókar Háskólans* (TOH), som er et talesprogsarkiv. Begge er i udgangspunktet traditionelle seddelsamlinger, men er senere blevet digitaliseret i større eller mindre grad. Samlingerne er ellers vidt forskellige, med hensyn til deres størrelse, indhold, struktur, tidsrum osv. (se Tabel 1).

	ROH	TOH
Kilder	Trykte kilder og håndskrifter	Breve fra og samtaler med informanter
Type af information	Belæg på ord fra skriftsproget	Kommentarer om ord og ordbrug i daglig tale
Opslagsord	ca. 700.000	ca. 56.000
Omfang og indhold	ca. 2,2 mio. sedler med opslagsord, ordklasse, citat og kildehenvisning	ca. 200.000 sedler med diverse oplysninger (fx betydning, bøjning, udtale, syntaks mm.)
Tidsrum	16.-20. århundrede	20. århundrede
Formidling	Digitaliseret; søgbart på åbent website	Delvis digitaliseret; ikke alment tilgængeligt

Tabel 1: Oversigt over ROH og TOH.

ROH, skriftsprogsarkivet, er langt det største af samlingerne. Ved excerperingen var sigtet at dække så mange og varierede tekster som muligt fra midten af det 16. til slutningen af det 20. århundrede. Alle sedlerne har samme udformning og parallelle oplysninger: (1) opslagsordet samt ordklasse, fx *nýmæli*, n. 'nyhed, inno-

vation' (n = substantiv, neutrum); (2) det exciperede citat med den pågældende ordform understreget, fx *Eg veit engin nýmæli herra!* 'Jeg kender ingen nyheder, herre'; (3) kildehenvisning, dvs. værkets faste akronym, eventuelt bind eller årgang, og sidetal, fx *StThLear 14* (en oversættelse af Kong Lear, side 14).

TOH, talesprogsarkivet, blev bygget op som resultat af et radioprogram som OH's medarbejdere styrede fra 1950'erne (jf. Ingólfsson 1988:68) til 2005. Ved siden af folkelig oplysning om sproget blev udsendelserne brugt til at komme i kontakt med lyttere som viste sig at være villige til at være sproginformanter. Dette gav leksikograferne mulighed for at få svar på spørgsmål og indsamle information, fx om bestemte ord og ordforbindelser, eller om ord inden for bestemte domæner. Svarene kom med breve eller i telefonsamtaler, og de pågældende oplysninger blev registreret på sedler med henvisning til informanten, hans eller hendes hjem- og fødeegn og alder.

I TOH består oplysningerne på en arkivseddel af fire dele: (1) opslagsord med ordklasse og køn af substantiver; (2) informantens kommentar om ordet; (3) det geografiske område som informanten stammer fra; (4) informantens navn og adresse samt brevets eller samtalens dato. Kommentarerne er vidt forskellige, afhængig af bl.a. spørgsmålet som blev stillet og informantens indstilling (jf. Hovmark 2023:166-168 om seddelsamlingen for *Ømålsordbogen*). De kan være alt fra et nej, at informanten ikke kender til ordet, til en udførlig kommentar om et ords eller en ordforbindelses semantiske, morfologiske og/eller syntaktiske egenskaber, eventuelt med brugseksempler og endda etnologisk information. Som et eksempel på en kommentar kan nævnes en informants udsagn om ordet *della* '(en type) bagværk'. I dansk oversættelse lyder kommentaren: "Dellur [plur.] (I udtalt som i Ella) = ástarpungar [dvs. samme slags boller], langt hyppigere at kalde dette for *dellur* i min egn og jeg vil fortsat gøre det." Her er der altså sagt noget om både ordets udtale, betydning og brug. En anden informants udsagn, om ordet

baukur, er derimod kun en meget kortfattet forklaring af dets betydning (dansk oversættelse): “= en beholder af metal.”

2.1.2. Digitalisering af arkivmaterialet

Digitalisering af materialet i ROH begyndte allerede i 1980'erne (jf. Svavarsdóttir et al. 1993, Prospectus 1993; se Svensén 2004:54 om digitalisering af leksikografiske seddelsamlinger alment). Det første skridt var en systematisk registrering og grov analyse af opslagsordene. Ordregistret gav en mere fleksibel adgang til materialet uafhængig af den alfabetiske organisering, som havde været den eneste tilgang til seddelsamlingen. Desuden fik man et vigtigt overblik over antallet af opslagsord og hvordan de fordelte sig på forskellige ordklasser (fx omtrent 590.000 substantiver og 17.000 verber). Således gav registret et vigtigt indblik i ordforrådets sammensætning ifølge arkivmaterialet. Samtidig blev der også oprettet et digitalt kilderegister over de skrifter som var blevet excerperede.

Næste skridt i arkivets digitalisering var indtastning af udvalgte citater så man systematisk kunne redigere materialet, bl.a. for at finde ud af hvor nyttige og fyldestgørende de foreliggende belæg viste sig at være i forbindelse med redigering af en historisk ordbog for islandsk. Mellem 1983 og 1993 blev der gjort forskellige forsøg i den retning (se Jónsson 1988), bl.a. blev der udgivet et prøvehæfte med over 100 redigerede verber (Svavarsdóttir et al. 1993; Prospectus 1993). Redigeringsarbejdet viste at selv om de gamle arkiver vanligvis gav et godt overblik over ordenes semantiske, morfologiske og syntaktiske karakter og udvikling, virkede de indimellem mangelfulde. Den digitale tekstsamling *Íslenskt textasafn* (ÍT, se afsnit 2.2.1), som var under udarbejdelse i 1990'erne, kunne bruges af redaktørerne til at bekræfte og eventuelt supplere den beskrivelse som tegnede sig på basis af skriftsprogsarkivet. Forsøg med redigering resulterede således i vigtig erfaringsopsamling og bedre forståelse af hvordan forskellige kilder kan spille sammen. Af adskillige årsager fortsatte redigeringen af det historiske ma-

teriale dog ikke på dette tidspunkt, bl.a. fordi OH fik midlertidig økonomisk støtte til at digitalisere det meste af citatsamlingen.

Selv om ROH-arkivet er stort, er strukturen forholdsvis enkel, og det kræver derfor kun en forholdsvis simpel database for at hver oplysningstype kan blive optegnet i eget felt. Opslagsord og kildeakronymer får et fast id-nummer som kobler citaterne til den relevante enhed i henholdsvis ordregistret (lemmalisten) og kilderegistret.

Skriftsprogsarkivet, ROH, havde længe været den vigtigste kilde for historiske undersøgelser af islandske ord og ordforråd, men seddelsamlingen var først og fremmest tilgængelig for forskere og universitetsstuderende. Ved systematisk digitalisering af arkivmaterialet, inklusive de fleste citater, i en database og tilkomsten af webben blev det til gengæld muligt at gøre det tilgængeligt og åbent for en langt større brugerkreds. Dermed kunne hvem som helst konsultere det direkte med hensyn til forskellige spørgsmål og interesser angående islandske ord, deres grammatiske kendetegn og anvendelse gennem de godt 450 år som kildematerialet dækker. I den forstand kan man sige at den digitale version af ROH allerede fungerer som en vigtig historisk sprogresurse.

Brugergørsefladen og den underliggende database er efterhånden blevet renoveret og udvidet, bl.a. ved en mere detaljeret analyse af citaterne og mere fleksible søgemuligheder, for bedre at kunne tilfredsstille brugernes forskellige behov. Omkring år 2000 blev databasen til ROH gjort alment tilgængelig og søgbar på webben, og for nylig blev både databasen og brugergørsefladen renoveret, bl.a. med mere fleksible søgemuligheder (Svavarsdóttir 2024:195ff.).

En del af materialet i TOH er allerede blevet digitaliseret, dvs. lemmalisten, informantregistret, brevsamlingen og en mindre del af kommentarerne (Svavarsdóttir & Kvaran 1994), selv om det kun er tilgængeligt internt. På grund af kommentarernes diversitet og materialets mangfoldighed er det kompliceret at oprette en vel-

struktureret database som rummer hele TOH-arkivet og knytter de forskellige dele sammen på en formålstjenlig og gennemskuelig måde. Der arbejdes nu på udvikling og opbygning af en sådan database, og i fortsættelsen er det planlagt at indføre kommentarerne og andet materiale tilknyttet samlingen. Således vil hele arkivet efterhånden blive let tilgængeligt, i det mindste for leksikografer og sprogforskere, og det diskuteres også hvordan udvalgte dele af materialet kunne præsenteres på en interessant måde for almenheden.

2.2. Sprogkorporusser

På OH, og senere ÁMI, har man opbygget et antal digitale tekstsamlinger og sprogkorporusser til brug for islandsk leksikografi, sprogforskning og sprogteknologi.

2.2.1. Foreliggende korpusser

Den ældste digitale tekstsamling er *Íslenskt textasafn* (ÍT), som nu består af 65 millioner løbende ord fra gamle, men dog mest nyere tekster af forskellig slags. Den er søgbar på webben, men kun som ren tekst. Teksterne i ÍT er helt uden annotering, og eksempel materialet fra et sådant korpus bliver hurtigt uoverskueligt, især når det drejer sig om undersøgelsen af forholdsvis almindelige ord (se fx Trap-Jensen 2014:49). Den begrænsning gælder derimod ikke teksterne i nyere sprogkorporusser, hvor teksterne er grammatisk taggede og lemmatiserede, og hvor man derfor har flere og mere fleksible søgemuligheder, fx efter opslagsord, ordformer eller orddele og/eller efter forskellige grammatiske kendetegn.

Det største islandske korpus er *Risamálheildin* (RMH), et gigakorpus som består af 2,8 milliarder tokens (se Steingrímsson et al. 2018), for det meste fra forholdsvis nye tekster. Ved opbygningen af korpusset blev der fokuseret på mængden, dvs. at få fat i så mange tekster der allerede forelå digitalt, som muligt. Trap-Jensen

(2014:53) har påpeget at det netop er kendetegnende for de store korpusser “at de ikke er særlig godt balancerede.”

Et mindre korpus, men mere balanceret med hensyn til materialets mangfoldighed, er det såkaldte *Mörkuð íslensk málheild* (MÍM) som havde det britiske BNC som det vigtigste forbillede. MÍM består af 25 millioner tokens i forskellige tekster fra 2000-2010 (Helgadóttir et al. 2012) som både er grammatisk taggedede og lemmatiserede. Disse tekstsamlinger og korpusser har været en vigtig kilde ved udarbejdelsen af forskellige sprogresurser, bl.a. ordbøger over samtidssproget, den islandske bøjningsbeskrivelse *Beygingarlýsing íslensks nútímamáls* (BÍN) og *Íslenskt orðanet*, dvs. et islandsk ordnet. Derimod repræsenterer kun en brøkdel af korpusmaterialet ældre sprogbrug så disse resurser er til begrænset nytte ved en historisk beskrivelse af sproget og ordforrådets udvikling over længere tid.

2.2.2. Korpusser under udvikling

I dag er der tre nye sprogkorpusser bestående af ældre tekster på tegnebordet på ÁMI. Sammenlagt vil de dække samme tidsrum som skriftsprogsarkivet, dvs. fra ca. 1550 til 2000, og delvis også de samme kilder. Tabel 2 viser en oversigt over de korpusser der er under udvikling.

Korpus	Indehold	Omfang	Status
Historisk korpus	Håndskrifter, dokumenter og bøger fra 1550-1800	Endnu uklart	Forberedes; materialeindsamling
1800-tals korpus	Aviser og tidsskrifter	270 mio. tokens	Færdig i foreløbig udgave på webben
1900-tals korpus	Bøger	153 mio. tokens	Under udarbejdelse; begrænset adgang

Tabel 2: Nye korpusser der indeholder tekster fra 1550-2000.

Det historiske korpus der er planlagt, vil dække perioden fra midten af 1500-tallet og frem til 1800, dvs. de omtrent 260 år som svarer til den ældre del af ROH. Der foreligger kun få allerede digitaliserede tekster fra denne periode, så de må enten indtastes eller skannes og gøres læsbare med OCR. Digitalisering af gamle tekster er ikke en nem opgave, fx stammer mange tekster fra håndskrifter eller bøger trykt med fraktur, og typisk er der stor variation i skriftformer både inden for samme tekst og mellem tekster. Korpusset bliver på to niveauer, det ene med originalteksten (med latinske tegn) og det andet med en ortografisk standardiseret version for at lette automatisk analyse og søgning. Tekstbehandlingen er således kompliceret og tidskrævende, og det er endnu uklart hvor stort korpusset vil blive, og hvornår det bliver færdigt.

1800-talskorpusset består af tekster fra aviser og tidsskrifter, som på forhånd forelå digitalt (oprindeligt fra databasen *Tímarit.is*), og fejl i OCR-behandlingen var allerede til dels blevet korrigeret manuelt eller halvautomatisk. Korpusset består af 270 millioner tokens og vil indbefatte både en bogstavret version og en standardiseret version af teksterne. Det foreligger i skrivende stund i en foreløbig udgave.

Det sidstnævnte korpus, 1900-talskorpusset, består af tekster fra trykte bøger, både skøn- og faglitterære. Korpusset dækker det 20. århundrede, især før it-teknologiens indtog i 1980'erne. På grund af ophavsret kan det kun anvendes til forskningsformål.

3. Sammenligning af sprogarkiver og digitale sprogkorpusser

3.1. Resursernes forskellige natur og nytte

I afsnit 2 blev de vigtigste leksikografiske og sproglige resurser for islandsk præsenteret. Her vil vi fokusere nærmere på deres

indhold med hensyn til hvordan de kan spille sammen og opveje hinandens begrænsninger for at nå frem til en så detaljeret og pålidelig beskrivelse af ordenes grammatiske, semantiske og pragmatiske egenskaber som mulig.

3.1.1. Arkiver og korpusser som ordbogsressurser

ROH består af citater som er blevet exciperet manuelt, dvs. at mennesker har læst kilderne og udvalgt ord og udtryk som de har anset som interessante og informative. Ved excipering er der dog en fare for tilfældighed ved valg af citater, bl.a. fordi man har tendens til lettere at få øje på det usædvanlige end det almindelige, og således kan frekvente udtryksmåder blive overset. Dette kan lede til en ubalance i belægsamlingen for bestemte ord og reducere dens repræsentativitet (jf. Hovmark 2012:297-299), bl.a. ved at “ge en oproportionerligt stor tyngd åt det ovanliga och perifera språkbruket” (Svensén 2004:57). Digitale tekstsamlinger og korpusser består derimod af sammenhængende tekster, ofte hele værker, og derfor kan man i princippet finde alle eksempler af et ord som forekommer i teksterne i deres forskellige former og sammenhænge (ordforbindelser, konstruktioner osv.). Korpusserne kan således være nyttige til at supplere belæggene som findes i arkiverne (jf. Pind et al. 1993:14-15; Hovmark 2012), ikke mindst når det gælder meget almindelige og frekvente ord som gerne er underrepræsenterede i seddelsamlingerne.

Hvis et korpus skal være hovedgrundlag for en leksikografisk beskrivelse, må den dog opfylde en række krav til dets sammensætning, især teksternes variation med hensyn til alder, type, emne, stil osv. Desuden vil det være et krav at teksterne er godt analyseret, og at søgesystemet er både meget effektivt og fleksibelt. Det er også vigtigt at leksikograferne har grundigt kendskab til søgesystemets muligheder, og at de til deres rådighed har et godt redigeringsystem, helst med direkte kobling til korpusset.

3.1.2. Emneområde og tekstgenrer

Tidsmæssigt indeholder de eksisterende korpusser næsten udelukkende nyere tekster, mest fra 2000-tallet, selv om der findes enkelte ældre tekster i ÍT og i mindre grad i RMH. Men nu forberedes der nye historiske korpusser som tilsammen vil dække samme periode som ROH (se afsnit 2.2). Materialeudvalget i de fleste korpusser har hidtil beroet på at teksterne allerede fandtes digitalt, og det præger deres sammensætning med hensyn til emneområde og stil, selv hvis man har stræbt efter en så varieret sammensætning som muligt. MÍM-korpusset kommer nærmest en bred og balanceret dækning med hensyn til emne og teksttype, mens RMH hovedsagelig består af ikke-skønlitterære tekster, hvoraf omtrent halvdelen er nyheder (fra aviser, radio og tv), og andre store tekstkategorier er sociale medier, administration (især fra *Alþingi* og domstole) og lokale eller faglige tidsskrifter. Andelen af bøger, bl.a. skønlitteratur og videnskabelige tekster, er derimod lille.

Problemer ved indsamling af materiale til skriftsprogsarkivet, ROH, og de planlagte historiske korpusser er anderledes. Her er udvalget af mulige tekster begrænset af hvad der blev skrevet, og hvad der er bevaret fra forskellige perioder. I alt blev der excerperet fra omtrent 2.000 titler til ROH, og disse kilder inddeles i mere end 30 kategorier med hensyn til teksttype og emneområde. Teksternes antal og mangfoldighed varierer imidlertid meget efter periode. Således excerperedes størstedelen af de eksisterende værker fra 1600-tallet, selv om der kun er 32 og de fleste er religiøse. Derimod er der blevet excerperet over 1.000 titler fra 1900-tallet, omfattende næsten alle kategorier (Svavarsdóttir et al. 1993:3), men som alligevel kun er et udvalg af det som publiceredes i perioden.

Det begrænsede antal af eksisterende tekster fra gammel tid betyder at tekstmaterialet i de nye historiske korpusser uundgåeligt kommer til at overlappe det som blev excerperet for ROH, især fra perioden før ca. 1800. Desuden bliver 1800- og 1900-talskorpusserne begrænset til bestemte teksttyper. Det førstnævnte vil

udelukkende bestå af aviser og tidsskrifter, mange af dem allerede excerperede for ROH, mens det sidstnævnte bygger på skanning af trykte bøger. Mange bøger i 1900-talskorpusset tilhører de samme kategorier som de excerperede tekster til ROH, mens andre repræsenterer teksttyper som ikke fylder meget i ROH, fx populærlitteratur, børnebøger og oversættelser. Til trods for overlappning i kildegrundlaget bliver de historiske korpusser et vigtigt supplement til arkivmaterialet.

3.1.3. Talesprog

Uanset resursetype er talesproget stærkt underrepræsenteret. Det meste af tiden har der ikke været metoder til at bevare talesprog, og der findes ingen direkte historiske kilder om det før 1900-tallet. Talesprogsarkivet, TOH, retter sig mod daglig tale, især dagligdags ord, ordforbindelser og ordbrug som sjældent forekommer i skriftlige kilder, og er således et vigtigt tillæg til andre resurser. Det består dog af informanternes kommentarer om ord og deres vurdering af hvad de betyder og hvordan de bruges, ikke af autentisk forekommende talesprog. De fleste informanter, i alt over 3.000 kvinder og mænd fra forskellige dele af landet, var født og opvokset i 1800- og 1900-tallet. TOH kan således anses som historisk, selv om tidsrummet er begrænset.

Hvad korpusserne angår, er det meget arbejds- og tidskrævende at indsamle og forberede store mængder talesprogsmateriale på en sådan måde at det kan lægges ind i et korpus. En lille samling af transskriberet spontant islandsk talesprog, svarende til ca. 50 timers optagelser, er dog inkluderet og søgbart både i MÍM og ÍT. Desuden findes der et korpus bestående af interview med unge mennesker, i alt omtrent 20 timer, som giver tilgang både til lyd og transskriptioner (Hilmisdóttir 2024:108-110). Selv om dette materiale ikke rækker langt i leksikografisk sammenhæng, kan det i nogle tilfælde supplere skriftsprogsilderne ved at fange udtryk som først og fremmest forekommer i tale (se Svavarsdóttir 2003).

3.2. Sociolingvistiske faktorer

TOH er i mange tilfælde en god kilde til sproglig variation som kan kobles til sociologiske faktorer. Ved hvert udsagn er der oplysninger om informanten, dvs. navn, geografisk herkomst, bopæl, køn osv., og arkivmaterialet kan således give oplysninger om sproglig variation som er knyttet til sociolingvistiske faktorer, ikke mindst dialektal udbredelse af ord, ordformer og ordbrug.

3.2.1. Dialektal ordbrug som kun findes i talesprog

Som et eksempel på at TOH er hovedkilde til et dialektalt ord, kan man tage ordet *della* i betydningen '(en type) stegt bagværk'. I arkivet er der 12 sedler hvor ordet er nævnt som et alternativ til *ástarþungur* med samme betydning. Ifølge den etymologiske ordbog svarer *della* muligvis til sidstedelen af *frikadella* 'kødbolle', et låneord fra italiensk gennem dansk (*frikadelle* < *frittadella*) (Magnússon 1989; jf. også Árnason 2011:151-152). Denne tolkning støttes af udtalen som ifølge kommentarer i TOH er langt [l:], typisk for bl.a. låneord, men ikke det hyppigere [tl] (som fx i homografen *della* 'nonsens, dille'). Alle belæg i talesprogsarkivet er knyttet til det nordlige Island, hvilket tyder på at ordet har haft en begrænset geografisk udbredelse. Desuden viser nogle af kommentarerne informanternes opfattelse af og holdninger til ordet, fx siger én at *della* blev anset som finere end det mere udbredte *ástarþungur* (antagelig med hensyn til sidstedelen *-þungur* 'þung'), og at det derfor mest blev brugt af "hisþursdömmum", dvs. fine damer. En anden informant hævder at *della* var det almindelige ord, mens *ástarþungur* mest blev brugt ironisk.

Udsagnene i TOH kan således give værdifuld indsigt i sprogbruges opfattelse af ord og ordanvendelse, men de kan også være i modstrid med hinanden, så det er ofte svært at drage generelle konklusioner ud fra de få kommentarer som findes. Alligevel indeholder TOH vigtige og mangfoldige oplysninger, med hensyn

til *della* og mange andre ord, som næppe findes i skriftsproglige kilder.

3.2.2. Morfologisk variation i tale og skrift

Der er også eksempler på at ROH og TOH supplerer og støtter hinanden. Det viser fx oplysninger om homonymerne *skúr* ‘nedbør’ og *skúr* ‘(lille, enkel) bygning’. De fleste sprogbrugere anvender det første som femininum (def. *skúrin*, pl. *skúrir*) og det sidste som maskulinum (def. *skúrin*, pl. *skúrar*), men hos andre falder de to sammen morfologisk så de enten begge er i maskulinum eller begge i femininum (jf. Blöndal 1920-1924). Dette leder til variation i brugen, som ifølge Blöndal er regional: maskulinum af *skúr* ‘nedbør’ findes i nord- og østlandet, og femininum af *skúr* ‘bygning’ knytter sig til et område i Vestfjordene. Andre steder er homonymerne adskilte morfologisk så vel som semantisk.

Tabel 3 viser fordelingen mellem varianter ifølge sprogarkiverne (eksempler hvor kønnet ikke er utvetydigt, medregnes ikke).

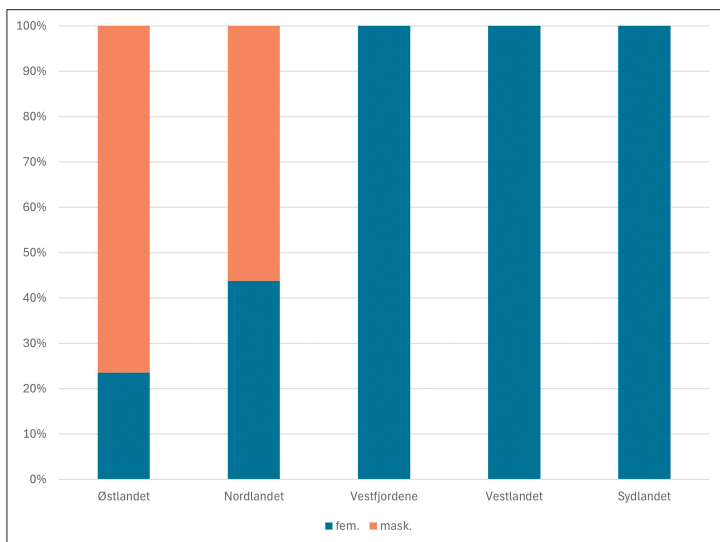
	<i>skúr</i> ‘nedbør’		<i>skúr</i> ‘bygning’	
	fem.	mask.	fem.	mask.
ROH	38 (70 %)	16 (30 %)	4 (19 %)	17 (81 %)
TOH	23 (51 %)	22 (49 %)	5 (45 %)	6 (55 %)

Tabel 3: Køn af homonymer i skrift- og talesprogarkiverne.

Det fremgår af Tabel 3 at materialet fra de to arkiver stort set stemmer overens med Blöndals beskrivelse fra 1920'erne. I begge homonymer er der variation i køn således at femininum dominerer i *skúr* ‘nedbør’, men maskulinum i *skúr* ‘bygning’. Jævnere distribution i TOH kan skyldes indsamlingsmetoden, dvs. at den sjældnere variant har været i fokus i leksikografernes forespørgsel. Det ville være tidskrævende at komme frem til sådanne oplysninger.

ger i store korpusser, fx i RMH's 13.000 og 17.000 belæg om *skúr* i femininum respektive maskulinum, fordi homonymer af samme ordklasse ikke kan adskilles i søgningen.

Oplysninger om geografisk udbredelse af varianter ville man først lede efter i TOH. Figur 1 viser hvordan brugen af maskulinum og femininum af *skúr* 'nedbør' fordeler sig efter region ifølge informanterne (hvis de har meddelt at begge køn var brugt i deres hjemegn, var begge registreret).



Figur 1: Distribution af femininum og maskulinum af ordet *skúr* 'nedbør' ifølge talesprogsarkivet (oplysninger fra 45 informanter).

Figur 1 tyder på at i syd- og vestlandet samt Vestfjordene bruges *skúr* 'nedbør' udelukkende som femininum (*skúrin*). I nord- og østlandet er maskulinum (*skúrinn*) derimod dominerende, men femininum er også kendt og brugt. Nogle informanter hævdede at maskulinum havde været hyppigere før i tiden, og det tyder på at udbredelsen af maskulinum kan være på retur.

Der er for begrænsede oplysninger om *skúr* 'bygning' i TOH til

at vise en landsdækkende distribution af varianterne. Der findes fem udsagn om brugen af femininum, og interessant nok kommer fire af dem fra samme område som Blöndal (1920-1924) knyttede varianten til, dvs. de nordlige Vestfjorde. I tillæg er der fire belæg i ROH på femininum, alle fra kilder som stammer fra Vestfjorderne. Således supplerer og støtter arkiverne hinanden og de meget begrænsede oplysninger om denne variants udbredelse i Blöndal (1920-1924).

3.3. Skriftsprogsarkivet som sprogteknologisk resurse

Den digitale version af ROH kan anses som et korpus med sine ca. 2 millioner citater indlagt i en systematisk opbygget database hvor forskellige oplysningstyper står i adskilte felter. Hvert belæg er knyttet til et opslagsord (i standardiseret nutidsform), og det kan således siges at den pågældende ordform i citatet er blevet lemmatiseret. Opslagsordene er desuden markeret med hensyn til ordklasse (plus genus for substantiver) og morfologisk struktur (simpleks, afledning med præfiks og/eller suffiks, sammensætning). I hvert belæg er ordformen afgrænset og opført bogstavret i et særfelt i databasen, og i et andet felt indføres den tilsvarende standardiserede nutidsform. I tillæg er hvert citat, og dermed den tilhørende skrift- og bøjningsform af opslagsordet, tidsfæstet. Til sidst er belægget knyttet til sin kilde.

Den analyse og kategorisering af historisk sprogmateriale som på denne måde er systematisk registreret i databasen og hovedsageligt bygger på manuel analyse, kan forventes at være nyttig ved opbygning af historiske sprogkorporer som dækker samme tidsrum som ROH. Det kan navnlig gælde ved rettelser i teksterne, deres standardisering, samt grammatisk tagging og lemmatisering af ordformer. På den måde kan det gamle skriftsprogsarkiv med de yderligere analyser som er blevet tilført i den digitale version, blive et vigtigt sprogteknologisk hjælpemiddel ved opbygning af

nye, historiske korpusser, især ved en (halv)automatisk analyse af teksterne. Samtidig vil sådan et nyt korpus blive et vigtigt supplement til ROH som kilde til en historisk, leksikografisk beskrivelse af islandsk.

4. Konklusioner

Udgangspunktet i denne artikel var at diskutere og vurdere den leksikografiske værdi af store historiske sprogarkiver, bygget op på basis af excerpering af tekster, vis-a-vis moderne digitale sprogkorpusser med sammenhængende tekster. Foruden tekstsamlinger og korpusser med nutidssprog, som længe har eksisteret, arbejdes der nu på oprettelse af historiske korpusser som dækker helt eller delvis samme periode som arkiverne. Dette giver anledning til at spørge: Vil de nye sprogresurser afløse de gamle, eller kan de fungere side om side og støtte hinanden? Det har vi drøftet i artiklen med hensyn til forskellige islandske sprogresurser og deres betydning for leksikografi og sprogforskning.

I en artikel om brugen af seddelsamlinger og korpusser ved redigering af *Ømålsordbogen* skriver Hovmark (2012:298): “det er fristende og nærliggende at anse seddelsamlinger for at være et ringe(re) kildegrundlag end tekstkorpora, bl.a. fordi korpora rummer mulighed for kvantitative undersøgelser og giver et bedre indblik i fx syntaktiske forhold og mere upåfaldende ord og betydninger der ikke altid er blevet indstreget når man har excerperet til en seddelsamling.” De to slags resurser viser imidlertid sprogets egenskaber på forskellig vis og supplerer hinanden: Korpusserne indeholder mange eksempler på almindelige ord og sprogbrug, mens seddelsamlingerne gerne fremhæver det som er sjældent forekommende (Svensén 2004:57). Hovmark (2012:305) konkluderer at begge resurser har selvstændig værdi for leksikografen, og at spørgsmålet først og fremmest drejer sig om kildekritik og

repræsentativitet. Leksikografen må huske at begge datatyper er “et udsnit af virkeligheden som skal vurderes kritisk som kildegrundlag”. Vi er kommet til lignende konklusion, dvs. at både de traditionelle arkiver og de digitale korpusser udgør værdifulde ressourcer for islandsk leksikografi og sprogforskning.

ROH holder sin værdi som historisk oversigt over ordforrådet og er en vigtig kilde til det islandske sprogs udvikling fra det 16.-20. århundrede. Den digitale version er nu i forholdsvis fin stand med nyligt renoveret database og grænseflade, tilføjelse af citater fra seddelsamlingen, rettelser og standardisering. Yderligere analyse og kategorisering af citaterne med hensyn til strukturelle og/eller semantiske egenskaber ville, med hensyn til fortsat anvendelse af arkivet, være ønskværdig. I det arbejde må der tages hensyn både til de forskellige brugeres behov, og til at materialet kan være til nytte ved opbygning af historiske sprogkorpusser, fx ved tilpasning af sprogteknologiske værktøjer til automatisk analyse af ældre tekster.

TOH indeholder forskellige oplysninger som ikke findes i andre kilder, og derfor er det en uvurderlig sprogresurce. Arbejde på digitalisering af materialet er nu i gang, og en skræddersyet database er under udvikling. Desuden diskuteres det hvordan udvalgte oplysninger fra samlingen kan præsenteres på en interessant måde, fx ved hjælp af kort som viser udbredelsen af sprogvarianter.

Således vil arkiverne efter vores vurdering fortsat udgøre vigtige kilder til vores viden om islandske ord, ordbrug og ordforråd, og samtidig er tilkomsten af nye historiske sprogkorpusser både nyttig og ønskværdig. Digitale sprogkorpusser har allerede bevist deres værdi ved undersøgelse og leksikografisk analyse af nutidsproget, og brugen af de forholdsvis få ældre tekster i eksisterende digitale tekstsamlinger har vist at de er et vigtigt supplement til arkivmaterialet.

Litteratur

Ordbøger, korpusser og digitale resurser

BÍN = *Beygingarlýsing íslensks nútímamáls*. Kristín Bjarnadóttir (red.). Árni Magnússon-instituttet for islandske studier. <<https://bin.arnastofnun.is>> (marts 2025).

Blöndal, Sigfús (1920-1924): *Íslensk-dansk Ordbog*. Reykjavík. <<https://blondal.arnastofnun.is>> (marts 2025).

BNC = *British National Corpus*. University of Oxford. <<http://www.natcorp.ox.ac.uk>> (marts 2025).

Íslenskt orðanet. Jón Hilmar Jónsson (red.). Árni Magnússon-instituttet for islandske studier. <<https://ordanet.arnastofnun.is>> (marts 2025).

ÍT = *Íslenskt textasafn*. Árni Magnússon-instituttet for islandske studier. <<https://corpus.arnastofnun.is>> (marts 2025).

Magnússon, Ásgeir Blöndal (1989): *Íslensk orðsifjabók*. Reykjavík: Orðabók Háskólans.

MÍM = *Mörkuð íslensk málheild*. Árni Magnússon-instituttet for islandske studier. <<https://malheildir.arnastofnun.is>> (marts 2025).

RMH = *Risamálheildin* (2024). Árni Magnússon-instituttet for islandske studier. <<https://malheildir.arnastofnun.is>> (marts 2025).

ROH = *Ritmálssafn Orðabókar Háskólans*. Árni Magnússon-instituttet for islandske studier. Digital version: <ritmalssafn.arnastofnun.is> (marts 2025).

Tímarit.is. Landsbókasafn Íslands Háskólabókasafn. <<https://tímarit.is>> (april 2025).

TOH = *Talmálssafn Orðabókar Háskólans*. Árni Magnússon-instituttet for islandske studier.

Ømålsordbogen. *En sproglig-saglig ordbog over dialekterne på Sjælland, Lolland-Falster, Fyn og omgivende øer 1- (1992-)*. Køben-

havn: Center for Dialektforskning, Københavns Universitet, og Universitets-Jubilæets danske Samfund.

Anden litteratur

- Árnason, Mördur (2011): Ásgeir, Orðsifjabókin og endurútgáfa Íslenskrar orðabókar 1983. I: *Orð og tunga* 13, 145-171.
- Helgadóttir, Sigrún, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir & Hrafn Loftsson (2012): The Tagged Icelandic Corpus (MÍM). I: *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages – SaLTMiL 8 – AfLaT 2012*. Istanbul. 67-72.
- Hilmisdóttir, Helga (2024): Talspråkskorpuser som resurs för isländska ordböcker. I: *LexicoNordica* 31, 103-125.
- Hovmark, Henrik (2012): Data og repræsentativitet i ordbogsarbejdet. I: Birgit Eaker et al. (red.): *Nordiska studier i lexikografi* 11. Lund: Nordiska föreningen för lexikografi. 296-308.
- Hovmark, Henrik (2023): Seddelsamlinger – historisk arkivmateriale eller levende resurse? I: Louise Holmer et al. (red.): *Nordiska studier i lexikografi* 16. Lund och Göteborg: Nordiska föreningen för lexikografi. 165-176.
- Ingólfsson, Gunnlaugur (1988): Söfnun Orðabókar Háskólans úr mæltu máli. I: *Orð og tunga* 1, 65-72.
- Jónsson, Jón Hilmar (1988): Sagnorðagreining Orðabókar Háskólans. I: *Orð og tunga* 1, 123-174.
- Pind, Jörgen, Kristín Bjarnadóttir, Jón Hilmar Jónsson, Guðrún Kvaran, Friðrik Magnússon & Ásta Svavarsdóttir (1993): Using a Computer Corpus to Supplement a Citation Collection for a Historical Dictionary. I: *International Journal of Lexicography* 6/1, 1-18.
- Prospectus (1993) = *A Prospectus to a Dictionary of Icelandic Verbs. English Version of the Introduction*. Reykjavík: Institute of Lexicography.

- Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson & Jón Guðnason (2018): Risamálheild: A Very Large Icelandic Text Corpus. I: *Proceedings of LREC 2018*. Miyazaki. 4361-4366.
- Svavarsdóttir, Ásta (2003): Ordbogen og den daglige tale. Om den islandske talesprogsbank (ISTAL) og dens betydning i ord-bogsredaktion. I: Zakaris Svabo Hansen & Anfinnur Johansen (red.): *Nordiske Studier i Leksikografi* 6. Tórshavn: Nordisk for-ening for leksikografi. 43-48.
- Svavarsdóttir, Ásta (2024): Ritmálssafn Orðabókar Háskólans. Samtíð og saga. I: *Orð og tunga* 26, 185-210.
- Svavarsdóttir, Ásta & Guðrún Kvaran (1994): The computeriza-tion of the spoken language archive. I: Karl Hyldgaard-Jensen & Viggo Hjørnager Pedersen (red.): *Symposium on Lexicogra-phy* VI. Tübingen: Max Niemeyer Verlag. 127-139.
- Svavarsdóttir, Ásta, Guðrún Kvaran, Jón Hilmar Jónsson & Kristín Bjarnadóttir (1993): *Sýnihefti sagnorðabókar*. Reykjavík: Orðabók Háskólans.
- Svensén, Bo (2004): *Handbok i lexikografi. Ordböcker och ord-boksarbete i teori och praktik*. Stockholm: Nordsteds.
- Trap-Jensen, Lars (2014): Leksikografisk tradition og fornyelse: tre revolutioner på 100 år? I: Ruth Vatvedt Fjeld & Marit Hov-denak (red.): *Nordiske studier i leksikografi* 12. Oslo: Novus forlag. 42-68.

Ásta Svavarsdóttir
forskningsdocent (emerita)
Árni Magnússon-instituttet for
islandske studier
Edda, Arngrímögötu 5
IS-107 Reykjavík
asta.svavarsdottir@arnastofnun.is

Helga Hilmisdóttir
forskningsdocent, ph.d.
Árni Magnússon-instituttet for
islandske studier
Edda, Arngrímögötu 5
IS-107 Reykjavík
helga.hilmisdottir@arnastofnun.is