LexicoNordica

LexicoNordica

29 · 2022

NORDISK LEXIKOGRAFI
– NU OCH I FRAMTIDEN

NORDISK FÖRENING FÖR LEXIKOGRAFI

LexicoNordica 29 · 2022
Nordisk lexikografi – nu och i framtiden

Omslag och sättning: Laurids Kristian Fahl
Tryckt hos: Tarm Bogtryk a-s, Danmark

LexicoNordica trycks med ekonomiskt stöd från
Nordplus Nordens Språk

ISSN 0805-2735
ISSN 1891-2206 (online)

Innehåll

Nordisk leksikografi – nu og i fremtiden

Henrik Hovmark & Anna Helga Hannesdóttir

Nordisk Forening for Leksikografi (NFL) har hermed fornøjelsen af at præsentere 29. bind af *LexicoNordica*. Hovedparten af årets nummer består vanen tro af tematiske bidrag – ni i alt – der på forskellig vis belyser emnet ved det 29. symposium: Nordisk leksikografi – nu og i fremtiden. Symposiet blev afholdt 10.-12. februar 2022 på Höllviksnäs kursgård i Skåne, Sverige. Derudover indeholder nummeret to anmeldelser, en orientering fra bestyrelsen for Nordisk Forening for Leksikografi samt til slut redaktionelle anvisninger.

Det er efterhånden snart 40 år siden at digitaliseringen og sprogteknologiske metoder, værktøjer og tankegange begyndte at ændre leksikografiens praksis og produkter radikalt. Det leksikografiske miljø er vant til at tænke på sprogteknologien som en naturlig medspiller. Ordbøger redigeres i databaser og udgives digitalt på internettet og som mobilapplikationer. Ikke desto mindre har udviklingen inden for de allerseneste år nået et punkt hvor den traditionelle leksikografi og dens selvforståelse for alvor er blevet udfordret. Det leksikografiske produkt er ikke længere nødvendigvis den velkendte ordbog, men kan fx være forskellige strukturerede vidensudtræk fra den oprindelige database – og man kan føle sig kaldet til at tale om leksikografiske resurser i stedet for ordbøger. Og en bruger er ikke længere kun et menneske i en bestemt kommunikationssituation, men kan være et program der henter data fra ordbogsprojektets vidensdatabase og tilgængeliggør eller anvender dem i en ny kontekst, menneskelig eller digital.

Udviklingen kan let opfattes på den måde at sprogteknologien efterhånden har opløst leksikografien som fag og praksis. Det 29. LexicoNordica-symposium havde netop til formål at undersøge

og nuancere denne opfattelse. Vi skal i de tematiske bidrag i dette nummer af tidsskriftet se eksempler på hvordan sprogteknologien er i stand til at udføre flere og flere af leksikografens kerneopgaver. Men vi skal også se hvordan leksikografiske metoder og kompetencer er nødvendige for at kunne udvikle eller forbedre sprogteknologiske applikationer. Det vil desuden fremgå at udviklingen i sidste ende ikke er bestemt af sprogteknologien isoleret set, men nærmere af mere omfattende digitaliseringsprocesser i det omgivende samfund og af de brugskontekster som ordbøger og andre leksikografiske resurser indgår i her. Brugerperspektivet står således mere centralt end nogensinde i leksikografien, og artiklerne i det følgende giver en lang række eksempler på i hvor høj grad leksikografens arbejde og rolle i vore dage består i overvejelser og arbejdsprocesser relateret til videreformidlingen af det leksikografiske indhold: Hvordan kan forskellige (sproglige) data i et bestemt værk eller en database bedst tilgængeliggøres for relevante brugergrupper?

Brugerspørgsmålet dukker naturligt op i forbindelse med revisioner af eksisterende ordbøger. Pär Nilsson & Bodil Rosqvist giver i deres artikel fx en beskrivelse af de mange formidlingsmæssige tiltag der må gøres når et stort leksikografisk værk der er udkommet i trykt form over en lang tidsperiode (*Svenska Akademiens ordbok*, SAOB), skal revideres og udkomme bredt på internettet, men samtidig stadig tilgodese en specialiseret brugergruppe. Og Margunn Rauset viser i detaljer hvordan publicering på internettet til stadighed må følge med tiden og tilpasses nye generationer – arbejdet med en ny hjemmeside for de reviderede udgaver af *Bokmålsordboka* og *Nynorskordboka* har affødt målrettede brugerundersøgelser som også har afsløret forskellige dilemmaer.

Yderligere to bidrag beskæftiger sig med de rammer og muligheder som publicering på en hjemmeside bringer med sig: Ordbøger begynder at indgå i større vidensnetværk. Emma Sköldberg beskriver hvordan den ændrede kontekst for publiceringen af an-

den, reviderede udgave af *Svensk ordbok* – fulddigital publicering side om side med *Svenska Akademiens ordlista* (SAOL) og SAOB på ordbogssitet *svenska.se* – har haft væsentlig indflydelse på det leksikografiske revisionsarbejde og på overvejelser omkring *Svensk ordboks* rolle og profil. Tarrin Wills viser hvordan *Ordbog over det norrøne prosasprogs* hjemmeside er begyndt at linke til en lang række andre digitale resurser (korpusser, ordbøger, tekstudgaver). De nye digitale og sprogteknologiske muligheder udnyttes som et internt redaktionsredskab, men bruges nu også til at komplettere den endnu ufærdige ordbog på hjemmesiden. I og med at de mange eksterne resurser også bliver tilgængelige via ordbogens hjemmeside, bliver siden også en bredere vidensresurse.

Men som nævnt finder sprogteknologiske applikationer i kombination med leksikografiske kompetencer også anvendelse i nye sammenhænge. Magnus Breder Birkenes, Lars G. Johnsen & Andre Kåsen beskriver udviklingen af en applikation som kan analysere og håndtere ikke kun synkrone, men også diakrone stavevarianter, og som dermed kan forbedre korpussøgninger og analyser af formmæssig variation og udvikling. Katarzyna Dominczak, Lene Antonsen og Trond Trosterud undersøger i deres artikel hvordan en sprogteknologisk applikation kan forbedre en ordbog mellem to strukturelt set meget forskellige sprog: nordsamisk og norsk. Applikationen foretager en baggrundsanalyse af det grammatiske og semantiske indhold i en søgning på det ene sprog, med henblik på at kunne vise mere præcise oversættelsesforslag på målsproget. Applikationen udvikles på basis af brugerinddragelse i form af spørgeskema og logfiler af brugen af ordbogen. Endelig giver Henrik Køhler Simonsen et eksempel på brug af leksikografisk og sprogteknologisk kompetence i en helt ny kontekst, nemlig som redskab i udformningen og brugen af en specialiseret chatbot rettet mod sundhedspersonale.

Temasektionen indeholder også to artikler der viser hvordan leksikografisk og sprogteknologisk kompetence væves sammen,

og i hvor høj grad begge kompetencer hver især kan bidrage til udviklingen af leksikografiske værktøjer og resurser. Sanni Nimb, Bolette S. Pedersen, Nathalie Carmen Hau Sørensen, Ida Flörke, Sussi Olsen & Thomas Troelsgård gør rede for udviklingen af den semantiske del af det nye Centrale OrdRegister for dansk (COR). Projektet trækker på en lang række eksisterende leksikografiske resurser, og leksikografisk kompetence er en nødvendig del af arbejdet med at finde frem til kernebetydningerne i det centrale ordforråd. Samtidig udnytter projektet sprogteknologiske redskaber og kunstig intelligens i de semantiske analyser og udvælgelsen af betydninger. Endelig viser Þórdís Úlfarsdóttir & Steinþór Steingrímsson hvordan man med hjælp fra sprogteknologiske metoder, bl.a. brug af pivotsprog og oversættelsesmaskiner, automatisk kan danne store dele af indholdet i en helt ny islandsk-engelsk ordbog. Men også her er anvendelsen af leksikografisk og lingvistisk kompetence nødvendig, både i kvalitetssikringen af de sprogteknologiske metoder og i den endelige redigeringsproces.

Herefter følger to anmeldelser: Lars Trap-Jensen anmelder anden og reviderede udgave af *Svensk ordbok*, og Bo A. Wendt anmelder en ny svensk etymologisk ordbok, *Ordklok*, der har fokus på slægtskabet med andre ord i samtiden.

Årets nummer afsluttes med en rapport fra Nordisk Forening for Leksikografi ved formand for bestyrelsen, Hanne Lauvstad.

Redaktionen af dette nummer består af de to hovedredaktører Anna Helga Hannesdóttir og Henrik Hovmark (nytiltrådt), samt landsredaktørerne Anna Braasch (Danmark), Kjetil Gundersen (Norge), Lennart Larsson (Sverige), Harry Lönnroth (Finland) og Ásta Svavarsdóttir (Island).

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

2023: Oversættelse og leksikografi i Norden
2024: Nordiske ordbøger og valget af datamateriale

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. Også tak til bestyrelsen for Nordisk Forening for Leksikografi, især formanden Hanne Lauvstad og kassereren Pär Nilsson, for godt samarbejde og både stor og uvurderlig hjælp i forbindelse med ansøgninger og den praktiske gennemførelse af symposiet på Höllviksnäs. 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. Endelig skal vi takke 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

Anna Helga Hannesdóttir
professor em.
Institutionen för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
Box 200
SE-405 30 Göteborg
anna.hannesdottir@svenska.gu.se

TEMATISKA BIDRAG

Om å bygge en leksikalsk ressurs for diakron skriftspråksvariasjon

Magnus Breder Birkenes, Lars G. Johnsen & Andre Kåsen

The present article sketches the process of constructing a pairing of modern word forms with their historical counterparts. We describe a particular pipeline for inducing such a lexical mapping, which results in a digital lexicographic resource. This resource can be used to amend existing digital dictionaries and build historical dictionaries, and it may form an essential part in applications and fields that work with textual data from a wide timespan.

1. Innledning

Moderne språkteknologi og digital tekstanalyse antar som regel at språket den opererer på, er standardisert og har få innslag av variasjon. For eksempel forutsettes det ofte at en gitt stamme- eller bøyningsform skrives på bare én måte innenfor et leksem. Dette er imidlertid ikke tilfellet. I norsk kan for eksempel en grammatisk funksjon med en viss morfologisk koding, for eksempel bestemt form entall, i noen tilfeller uttrykkes med ulike bøyingsaffikser og dermed anta ulike skriftformer. For eksempel kan bestemt form entall av *sol* i bokmål skrives enten *solen* eller *sola*.

Her er det altså snakk om synkron variasjon – som er ganske omfattende i norsk sammenheng (både i bokmål og i nynorsk). Men formvariasjon skaper også problemer i diakrone sammenhenger, for eksempel når man ønsker å se på historisk språk eller sammenligne eldre og nyere tekster, ettersom ordenes skrivemåte kan ha endret seg. For å kunne håndtere de forskjellige og noen ganger mange variantene rent praktisk er man imidlertid også nødt til å etablere én form som alle varianter kan kobles til, slik at for eksempel formerne *qvinde* og *kvinne* automatisk analyse-

res som varianter av samme leksem. I det følgende omtales denne samleformen som «grunnform».

Problemet med formvariasjon skyldes at koblingen mellom formene ikke er eksplisitt kodet i eksisterende ordbøker, men likevel underforstått. Mens det synkrone aspektet ved slik variasjon er godt ivaretatt i moderne ordbøker, både for menneskelig og maskinell tolkning, er det diakrone ikke tatt hensyn til i samme grad. Beskrivelse av det diakrone aspektet er ofte begrenset til etymologiske forhold eller henvisninger til eldre belegg (implisitt informasjon), hvor ordet gjerne har en helt annen skrivemåte. Ved å slå opp på leksemet *kvinne* i NAOB vil man i belegglisten finne formene *qvinde* (for eksempel hos Henrik Wergeland), *kvinde* (for eksempel hos Henrik Ibsen) og *kvinne* (for eksempel hos Dag Solstad) i både entall og flertall. Det er imidlertid ikke angitt eksplisitt at disse formene skal eksemplifisere *kvinne*. Den kunnskapen forutsettes det at leseren har selv.

Fra et leksikografisk ståsted tilhører historiske stavingsvarianter som *kvinne* og *kvinde* samme leksem. Nærmere bestemt er *kvinne* og *kvinde* varianter av den grammatiske formen ubestemt form entall, og *kvinne* representerer den moderne skrivemåten. Når eldre stavingsvarianter som *kvinde* innlemmes i tekstgrunnlaget til en ordbok, kan vi tenke oss at leksemet utvides med historiske former. I digital tekstanalyse er det likevel fullt mulig å tenke seg at det opprettes et nytt leksem med utgangspunkt i formen *kvinde* (bøyningsparadigmet er også grafemisk forskjellig fra det moderne).

Det at nye former introduseres gjennom modernisering og språkreformer, omtales i den datalingvistiske litteraturen som *lexical replacement*, altså leksikalsk utskiftning eller erstatning. Det vil si at en ny form erstatter den gamle. Formene fra ulike tidsepoker refereres til som historiske kognater. Den moderne formen er en liten endring av den historiske. I denne artikkelen tar vi bare for oss kognater og utskiftning med utgangspunkt i språkets gra-

femiske representasjon og holder talespråket utenfor. Skriftspråket kan endre seg uten at talen gjør det, og omvendt, og førstnevnte har i Norge endret seg mye gjennom rettskrivingsvedtak som har kommet på løpende bånd fra 1860-tallet og frem til i dag. Hvorvidt disse endringene reflekterer fonologiske endringer, skal vi ikke forfølge videre.

I denne artikkelen skal vi beskrive en metode for å knytte grafemiske ordformer fra ulike tidsperioder (som altså også kan kalles historiske kognater) sammen uten å se på leksemtilhørighet, og deretter skal vi presentere en måte å organisere formene videre i leksemer på, det vil si historisk fulle leksemer. Nærmere bestemt går metoden ut på at de historiske ordformene, enten de er grunnformer eller andre morfologiske varianter, blir koblet til en av de tilsvarende variantene i moderne norsk, som så kobles til et leksem.

Med denne metoden planlegger vi å lage en historisk ordliste. Ordlisten vil bli en digital ressurs som blant annet kan fungere som en utvidelse av eksisterende ordbøker ved at ordbøkene inneholder de historiske skrivemåtene. For å konstruere ordlisten benytter vi oss av det digitaliserte materialet ved Nasjonalbibliotekets samling samt metadata for gruppering og organisering av tekster.

I dette arbeidet vil vi av praktiske hensyn begrense oss til bokmål, som har en veldokumentert utviklingsbane fra dansk og frem til sin moderne form. Bokmål har dessuten et tilstrekkelig datagrunnlag. Nynorsk har en til dels uavhengig utvikling samt et betydelig magrere datagrunnlag.

2. Problemet

Fra et datamaskinelt perspektiv kan vi ta utgangspunkt i hva som møter algoritmene i digital tekstanalyse. For datamaskinen vil alt

som skrives likt, være likt, og alt som skrives forskjellig, være forskjellig. For å gruppere sammen distinkte forekomster etter andre kriterier enn at de er like på overflaten, kreves en representasjon i form av for eksempel en digital ordbok. Ved hjelp av en digital ordbok kan algoritmer gruppere ulike grunnformer og/eller bøyde former under det samme leksemet, for eksempel fastslå at forekomstene av *spise* og *spiste* kan grupperes under leksemet *spise*.

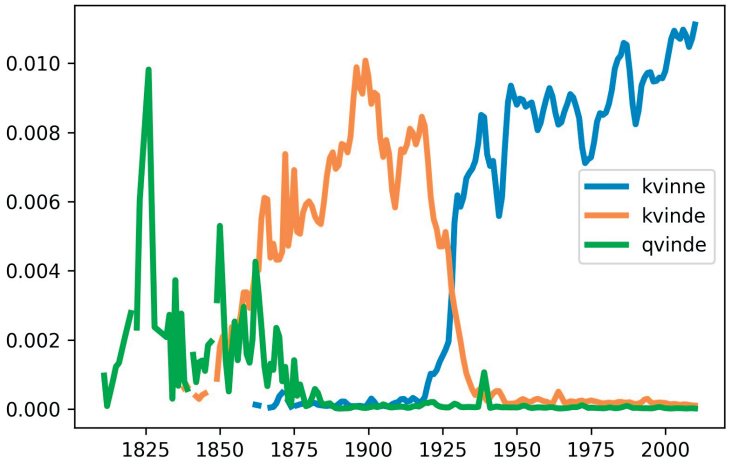
Vi føyer til at homografer ikke blir behandlet her. En og samme ordform kan tilhøre forskjellige leksemer, som *legger*, som kan være presens av verbet *legge* eller flertall av substantivet *legg*. Den grafemiske skrivemåten gir ingen hint om hvilket leksem som er i sving. For en slik disambiguering må man se på konteksten ordet står i. I det historiske tilfellet vil skriftkonvensjoner begrense omfanget av homografi, da substantiver på 1800-tallet typisk ble skrevet med stor forbokstav. Man hadde altså en kontrast mellom *Lægger* (substantiv) og *lægger* (verb).

I arbeidet vi presenterer her, er det forskjellige skrivemåter av en ordform som skal grupperes, og disse må derfor kodes med leksemtilhørighet og eventuelt grammatiske funksjoner. Men det må påpekes at denne kodingen i første omgang vil vise seg ved at for eksempel *legger* ses på som en utvikling av *lægger*. Grammatiske egenskaper blir tilordnet den historiske formen basert på en beskrivelse av den moderne. Se under for en beskrivelse av hvordan det kan gjøres.

2.1. Noen eksempler på historisk formvariasjon

Det er den nakne ordformen slik den står i teksten, som er utgangspunktet for arbeidet vårt med å utvikle en historisk ordliste. For eksempel ville en naiv analyse av 1800- og 1900-tallets tekster se på *kvinde* og *kvinne* som to forskjellige ord. I dag er det ingen digitale ordbøker for norsk som kobler de to formene sammen. Begge ordformene refererer til det samme og burde derfor kunne

relateres til hverandre. Vi skal senere beskrive en løsning på dette problemet, men først skal vi se litt nærmere på den historiske utviklingen i bruk av kjente skrivemåter av ordet *kvinne* ved hjelp av såkalte n-gram-trendlinjer[1] (Birkenes et al. 2015), som vist i figur 1.



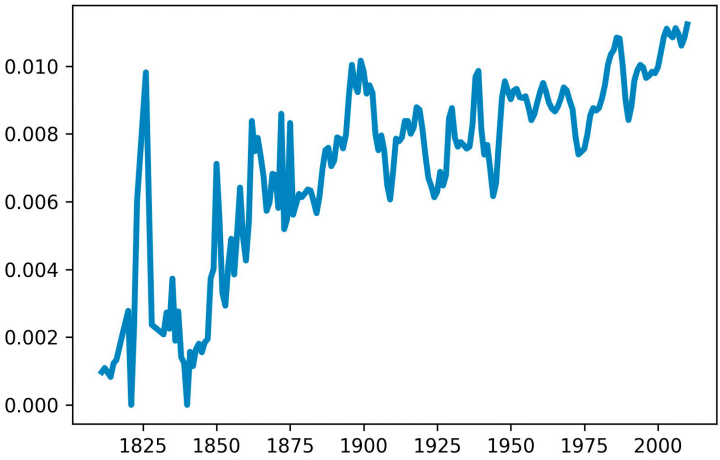
Figur 1: Trendlinjer for ulike skrivemåter av ordet *kvinne*.

Her har vi i tillegg med varianten *qvinde*. Grafen viser den relative frekvensen av de ulike stavingsvariantene av *kvinne* fra ca. 1800 og frem til år 2000.

I en praktisk søkekontekst der noen ønsker å se utviklingen i den relative frekvensen av ordet *kvinne* i norske bøker, kunne man tenke seg at de historiske skriftvariantene (de historiske kognatene) står for det samme, slik at trendlinjen for *kvinne* kan illustreres ved å slå sammen trendlinjene for de ulike skrivemåtene, som vist i figur 2.

En annen søkesituasjon kan være at noen er ute etter forskjellige kontekster for ordet *kvinne* i et historisk korpus og ønsker å

1 Trendlinjen viser den relative (prosentvise) frekvensen av en ordform i et korpus, det vil si hvor mange ganger ordformen forekommer sammenlignet med det totale antallet ordformer i korpuset.



Figur 2: Sammenslåtte trendlinjer fra figur 1.

få opp alle dokumenter som inneholder ordet, med tilhørende eksempler (konkordanser).

Digital tekstanalyse gir ofte en liste med ordformer som resultat, gjerne sammen med kvantitativ informasjon. Resultatet av slike lister gir grunnlag for aggregering, for eksempel at ordformene *kvinne* og *kvinnen* skal summeres opp som et mål på forekomsten av entall *kvinne*. Selv når ordformene kommer fra forskjellige perioder med forskjellige skrivemåter, skulle de kunne la seg gruppere.[2] For eksempel vil ekstraksjon av kollokasjoner gi en vektet liste av ordformer som statistisk er nær knyttet til hverandre. Om kollokatene stammer fra en tekstsamling som spenner over et tidsrom med språkendringer, vil det oppstå slike kognat-koblinger som ikke fanges inn med dagens digitale ressurser. Om vi tar ordet *kaffe*, for eksempel, og prøver å hente ut de viktigste foranstilte ordene (kollokatene) fra perioden 1900 til 1930, finner vi blant topp

2 Merk at det her er snakk om en annen type gruppering enn den som gis ved lemmatisering eller semantisk likhet.

ti kognatsett som *kop*/*kopp* og *sekker*/*sække*/*sækker*.[3] Her er det tenkelig at man i et gitt forskningsprosjekt ville foretrekke at hele ordlisten moderniseres slik at kollokaten *kopp* innbefatter begge formene *kop* og *kopp*.

2.2. Skjematisk fremstilling av prosessen

Målet vårt er å lage en ressurs som inneholder koblinger mellom den moderne formen av et ord og ordets historiske kognater. Vi viser hvordan koblinger lages mellom enkeltord, det vil si slik de opptrer grafemisk, og hvordan de koblingene på et senere stadium kan benyttes til å utvide leksemer. Skjematisk ser det slik ut for grunnformen *kvinne* og flertallsformen *kvinnene*:

kvinne → {kvinde, qvinde}
kvinnene → {kvinderne, qvinderne}

Slike koblinger legger grunnlaget for å utvide spørringer (såkalt *query expansion*) i søkesammenheng, i tillegg til at de danner grunnlaget for å gruppere former som nevnt ovenfor. Nedenfor ser vi også på hvordan koblingene kan benyttes til å konstruere historiske leksemer.

I en søkesituasjon der man leter etter bøker som tar for seg begrepet *kvinne*, vil man gjerne få treff på alle bøkene som inneholder en eller annen form av ordet *kvinne*. Det vil si *kvinne*, *kvinnen*/*kvinna*, *kvinner* og *kvinnene* i tillegg til alle de historiske kognatene, som for eksempel *qvinde*, *qvinden*, *qvinder* og *qvinderne*, og de tilsvarende formene basert på grunnformen *kvinde*. I en leksikografisk sammenheng vil en slik utvidelse også kunne benyttes til å finne historiske belegg for varianter av ord/leksemer. Det vil si at

3 Analysen ble laget med kollokasjonsappen fra DH-LAB (Nasjonalbibliotekets laboratorium for digital humaniora, se litteraturlisten). Perioden 1900–1930 inneholder de viktige 1907- og 1917-reformene i norsk normering.

man kan oppgi formen *kvinne* som søkeuttrykk og få treff på alle formene av ordet, også de historiske. Hvor mye en søking skal utvides, vil være opp til den som søker, for eksempel om det bare skal lages varianter for en grammatisk form som bestemt form entall.

Ressursen vil derfor kunne anvendes til generelt begrepssøk av brukere uten kjennskap til leksemkategorier og historiske former og også til mer spissede leksikografiske søk av brukere som har oversikt over formene og grammatikken til ordet de er ute etter.

3. Data og metode

I det følgende skal vi presentere datagrunnlaget og metodene vi bruker for å lage en historisk ordliste på basis av Nasjonalbibliotekets samlinger. Vi skal benytte teknikker fra datautvinning og statistisk maskinoversettelse. For å ha et tilstrekkelig datagrunnlag trenger vi et korpus bestående av bøker som foreligger både i en historisk og i en moderne variant. Med slike par av bøker på plass kan vi starte prosessen med å finne ord som korresponderer med hverandre, og benytte den korrespondansen til å konstruere par av kognater.

3.1. Datagrunnlag

Vi skal benytte oss av Nasjonalbibliotekets tekstsamlinger, slik de er eksponert gjennom bibliotekets DH-LAB med tilhørende API (*application programming interface*, norsk: *programmeringsgrensesnitt*), i tillegg til en såkalt spesialbibliografi, nemlig *Nasjonalt autoritetsregister for verk* (Verksregisteret). DH-LAB tilbyr en datamaskinell inngang til bibliotekets samlinger som lar en studere kvantitative aspekter ved samlingene på en programmatisk måte. DH-LAB består av flere komponenter som kan være av interesse for forskere med ulik grad av datateknisk kyndighet.

Formålet med Verksregisteret er å gruppere de forskjellige utgavene av et verk slik at alle utgavene som faller inn under det, skal kunne gjenfinnes og brukes. Verksregisteret ble utviklet i prosjektet SHARE-VDE, hvor flere amerikanske universitetsbiblioteker samt nasjonalbibliotekene i Finland og Norge deltar.

Ved hjelp av Verksregisteret finner vi altså alle utgaver av et verk, men for å kunne lage en historisk ordliste trenger vi i tillegg informasjon om utgavens språkform. En ny utgave av et verk er ikke nødvendigvis moderne i språkformen. Det finnes for eksempel både faksimiler og diplomatiske utgaver som gjengir den opprinnelige formen. Derfor må vi dele inn utgavene fra Verksregisteret i bolker basert både på utgivelsestidspunkt og på språkform. Dette skal vi se nærmere på i neste kapittel.

3.2. Parallelltekst

En viktig teknikk i det automatiserte arbeidet med å lage en historisk ordliste er såkalt *alignering*, det vil si å parallellstille tekster eller deler av tekster. I datalingvistikk er det vanlig å bruke termen *bitext* om én og samme tekst som er oversatt til ett eller flere språk (Tiedemann 2011), eller om flere versjoner av samme grunnlagstekst. Termen *parallel text* (norsk: *parallelltekst*) er også vanlig i samme betydning. På samme måte som om vi skulle alignert tekster på to ulike språk, velger vi i denne sammenhengen å alignere varianter av den samme teksten fra ulike språkstadier. Her er vi til dels begrenset av Verksregisteret og hva det inneholder. Det betyr at tilfanget av verker vi ser på, i prinsippet kan være mindre enn det faktiske antallet verker, men vi legger til grunn av varianter innad i et verk er komplett for alle praktiske formål.

For å identifisere tekster som kan representere ulike språkstadier, ser vi på frekvensen av de ulike formene som et høyfrekvent ord opptrer i. Høyfrekvente ord er med i de fleste tekster, og skrivemåten avslører tidsepoken. Hvis for eksempel ordformen *paa*,

og ikke *på*, går igjen i en tekst, gir det et hint om at teksten tilhører tiden før rettskrivingsreformen fra 1917. Et sett med slike høyfrekvente ord ble benyttet til å avgjøre om teksten er historisk eller moderne.[4] Fra et tilfeldig utvalg av fem hundre tekster fra de to periodene fant vi følgende (illustrert i tabell 1) blant de ordene som 1) har høy frekvens og 2) har en tilstrekkelig skillende effekt. Det siste vil si at formene har vesentlig forskjellig frekvensfordeling i de to periodene.

Ordform	**Historiske tekster**	**Moderne tekster**	**Kognat**
lese	0	1637	*læse*
fat	2712	0	*fatt*, men også *fat* substantiv
vide	5146	0	*vite*, men også *vid* adjektiv
paa	216 235	10 071	*på*
inn	0	27521	*ind*
ind	22 370	0	*inn*
på	9143	244 801	*paa*
lægge	2527	0	*legge*

Tabell 1: Et lite utvalg av ordformer med frekvenser. Frekvensene er for de utvalgte ordformene i henholdsvis historiske og moderne tekster, sammen med en indikasjon av ordformenes historiske kognater.

Ved å sammenligne frekvensen av moderne og historiske former i en variant av et verk med frekvensen av de samme formene i tekstutvalget kan vi avgjøre hvorvidt varianten tilhører et tidligere språkstadium eller ikke. Hver tekst får en vektet score[5] basert på hvor mange «moderne» ordformer den har, og hvor mange som tilhører den «historiske» perioden.

4 Her er det gjort et utvalg av historiske tekster i perioden 1700–1917, mens de moderne er hentet fra perioden 1920–2010.

5 For eksempel kan man bare telle opp forventet antall av de forskjellige ordformene.

En mulig utfordring med metoden er eventuelle forskjeller i måter å modernisere enkelte forfatterskap på. Man kunne se for seg at det er ulike tradisjoner for modernisering av for eksempel Henrik Ibsen og Bjørnstjerne Bjørnson. Dette problemet skal vi ikke forfølge videre her.

En annen utfordring er anakronismer: Selv i moderniserte utgaver vil visse former holde seg lenger, for eksempel riksmålsformer av hyppige ord, som *nu* og *efter*, eller pronominaladverb som *dertil* og *hvorpå*, som må kunne betraktes som gammeldagse i moderne bokmål. For å omgå effekten av enkeltord tar vi med en forholdsvis stor liste[6] med skillende ord i vurderingen av om en tekst er moderne eller historisk.

3.3. Setningsalignering

Når dokumentene er alignert, må setningene og ordene innad i dokumentene aligneres. Dette er to separate automatiserte prosesser der setningene sammenstilles først, så ordene.

En gjengs og mye brukt setningsalignerer er beskrevet i Varga et al. (2005): *hunalign*. Denne tar utgangspunkt i en maskinell ord-for-ord-oversettelse mellom en kildetekst og en måltekst. I vårt arbeid benytter vi nyere modeller for alignering, som sBERT (Reimers & Gurevych 2020). Disse modellene gjør ikke noen oversettelse *per se*, men omdanner tekst til maskinlesbare trekkstrukturer, og sammenligner de ulike trekkene for å finne de setningene som ligner mest på hverandre.

Setningsaligneringen utføres så utgave for utgave ved at utgavene i den historiske delen av korpuset sammenlignes med utgavene i den moderne. Det genereres altså kombinasjoner av gammel og ny tekst hvor setningene fra hvert par sidestilles. Ligger sannsynligheten for at setningene i setningsparet faktisk er like, under et visst nivå, forkastes setningsparet. Ikke alle setninger er paral-

6 I testinger har vi brukt en liste med 1500 ord.

lelle (det kan være tilføyelser eller andre endringer, eller det kan være forskjeller i tekstkvalitet forbundet med digitaliseringen av teksten, for eksempel automatisk bokstavgjenkjenning).

3.4. Ordalignering

Når setningene er alignert, kan vi så alignere på ordnivå, og det er slik vi får den tidligere illustrerte koblingen fra moderne til historiske ordformer:

kvinne → qvinde

For denne typen alignering har vi benyttet verktøyet SimAlign, som til dels bygger på samme teknologi som sBERT. Språkmodellen vi har brukt i SimAlign, er den som beskrives i Kummervold et al. (2021), som er trent på et stort materiale fra Nasjonalbibliotekets digitale samlinger. Giza++, beskrevet i Och & Ney (2003), er et alternativ til SimAlign, men Sabet et al. (2020) viser at sistnevnte er mer treffsikker.

4. Historisk ordliste som leksikografisk ressurs

Resultatet av aligneringen er en ordliste som kan brukes til forskjellige formål. Ett er å gi en beskrivelse og en kategorisering av historiske former, et annet er å oversette eldre tekster automatisk.

Om man i en moderne ordbok prøver å finne informasjon om ordformene *sygdom* eller *kvinder*, får man ikke noen treff,[7] men man får treff på *sykdom* og *kvinner*. Når historiske tekster er digitalt tilgjengelige, er det ønskelig å ha en fyldig leksikografisk beskrivelse av dem. I kapittel 4.1 ser vi på hvordan kognatparene kan brukes til å konstruere fulle historiske leksemer. I kapittel 4.2 tar vi

7 De aktuelle ordene er sjekket med NAOB og *Bokmålsordboka*.

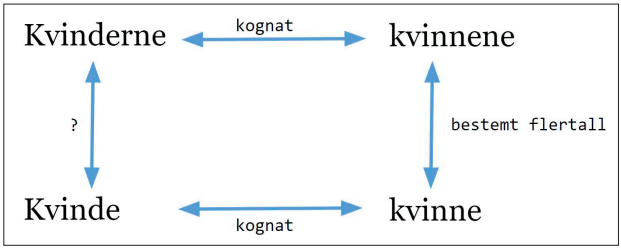
for oss muligheter for å modernisere tekster ut fra det perspektivet at modernisering er en form for oversettelse. Kapittel 4.3 tar for seg et problem som oppstår i forbindelse med automatisk modernisering: tvetydighet ved funksjonell splitting.

4.1. Fra ordformer til leksemer

Når koblingen mellom historiske og moderne ordformer er etablert, kan vi prøve å overføre informasjon fra den moderne beskrivelsen til den historiske. Så langt har vi kun sett på kobling mellom ordformer, som at *kvinne* korresponderer med *kvinde*, og *kvinner* med *kvinder*. Men vi kan også stille spørsmålet om det er mulig å koble relasjonene mellom ordformene sammen til et helt leksem, slik at alle de historiske formene blir gruppert sammen med de respektive moderne formene til ett leksem.

I figur 3 er det illustrert to relasjonstyper, én som viser at en ordform er kognat av en annen (heretter kognatrelasjonen), og én som angir ordformenes plass i leksemet (heretter leksemrelasjonen). Figuren viser med andre ord at *kvinne* og *kvinnene* er kognater av henholdsvis *Kvinde* og *Kvinderne* og samtidig at *kvinnene* er bestemt flertall av *kvinne*. Den ukjente relasjonen – som ikke er direkte kodet i tilgjengelige ordbøker – er den mellom *Kvinde* og *Kvinderne*.

Leksemrelasjonen og kognatrelasjonen antas å være kommutative. Det betyr at om vi starter i hjørnet som inneholder ordformen *Kvinde*, og så følger pilene for kognaten og deretter pilene for leksemrelasjonen bestemt flertall, vil vi få samme resultat som om vi først velger leksemrelasjonen bestemt flertall og så kognaten. Vi får *kvinnene* som resultat i begge tilfeller.



Figur 3: Kommutativt diagram over relasjonene *kognat* og *flertallsform*.

I praksis betyr det at informasjon om de historisk relaterte kognatene kan berikes med informasjon fra beskrivelser av det moderne språket. Resultatet er at de historiske kognatparene kan danne et utgangspunkt for å konstruere en digital historisk ordbok.

4.2. Automatisk modernisering

Både hver for seg og i kombinasjon kan ordaligneringen og setningsaligneringen komme til nytte i automatisk modernisering av tekst. Resultatet av ordaligneringen er en ordliste (som beskrevet over) som kobler to ordformer sammen. Formålet med modernisering kan være å tilgjengeliggjøre eldre tekster for de som ikke er kjent med historiske ordformer, eller det kan være forskjellige formål innen tekstanalyse – for eksempel å koble tekstene til moderne ordbøker.

En automatisk modernisering kan med de beskrevne ressursene ta to former: 1) Den konstruerte ordlisten kan fungere som grunnlag for substitusjon, det vil si at de historiske formene byttes ut med de moderne. 2) Ved hjelp av det setningsalignerte materialet (det vil si listen med par av setninger) kan vi trene maskinoversettere (for eksempel slike som er beskrevet i Raffel et al. 2020). Alternativ 2 vil kunne gi mer idiomatisk moderne norsk enn det som oppnås med alternativ 1.

4.3. Funksjonell splitting

Alignering vil gi flere koblinger mellom ord enn de som forbinder kognater. Disse kan oppstå for eksempel som følge av OCR-feil[8] (feil ved automatisk tekstgjenkjenning), som i *kjcerlighed*/*kjærlighet*, ved at ordet er oversatt, som i *kvinde*/*woman*, eller ved at ordet er byttet ut med en semantisk ekvivalent, som i *glad*/*lykkelig*.

Særlig utfordrende i den sammenhengen er tilfeller der et ord får en oppsplitting av funksjoner. Den historiske formen beholder et bruksområde i det moderne språkstadiet, samtidig som en nyere form erstatter den historiske innenfor andre områder. To eksempler er eldre *at* vs. nyere *at*, *aa* og *å*, og eldre *der* vs. nyere *der* og *det*.

Oppsplitting i funksjon er en utfordring i prosessen med å lage en modernisert tekst på grunnlag av en eldre. Om vi benytter alternativ 1 i forrige kapittel for automatisk modernisering av tekst (der eldre ordformer blir byttet ut med korresponderende moderne former), vil *at* byttes ut med *å* betingelsesløst, også der det ikke er riktig å gjøre det. I setningene *han prøvede at gaa* og *han troede at han var syg* vil utskifting av *at* med *å* i den første gi riktig resultat, mens i den siste vil *å* for *at* resultere i en ugrammatisk (og meningsløs) setning. For de andre ordene kan man bare velge en moderne form uten å se på kontekst. Samme situasjon gjelder for *der* vs. *det*, for eksempel i *der staar en kat i haven* vs. *han stod der i haven*. I den første setningen går det fint å bytte ut *der* med *det*, men i den siste setningen vil det resultere i en ugrammatisk setning.

5. Oppsummering

Vi har beskrevet en metode for å automatisk utvide eksisterende ordbøker til å omfatte historisk materiale, både på ordformnivå

8 OCR = Optical Character Recognition (optisk tegngjenkjenning).

og på leksemnivå. Vårt fokus har vært på leksikalske ressurser for digital prosessering av tekst, og vi har sett både på hvordan slike ressurser kan benyttes i beskrivelsen av språket selv (det rent leksikografiske), og på hvordan slike ressurser kan forbedre søk i litteratur og ellers gjøre det mulig å sammenligne tekster fra ulike tidsperioder i digital tekstanalyse.

Litteratur

Digitale ressurser

Bokmålsordboka. Språkrådet og Universitetet i Bergen. <ordbokene.no> (juli 2022).

DH-LAB = Nasjonalbibliotekets laboratorium for digital humaniora. <nb.no/dh-lab> (juli 2022).

NAOB = *Det Norske Akademis ordbok*. <naob.no> (juli 2022).

Verksregisteret = Nasjonalt autoritetsregister for verk. <bibliotekutvikling.no/kunnskapsorganisering/kunnskapsorganisering/nasjonalt-autoritetsregister-for-verk/> (juli 2022).

Annen litteratur

Birkenes, Magnus Breder, Lars G. Johnsen, Arne M. Lindstad & Johanne Ostad (2015): From digital library to n-grams: NB N-gram. I: *Proceedings of the 20th Nordic Conference of Computational Linguistics*. Linköping: Linköping University Electronic Press, 293–295.

Kummervold, Per E., Javier De la Rosa, Freddy Wetjen & Svein Arne Brygfjeld (2021): Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. I: *Proceedings of the 23rd Nordic Conference on Computational*

Linguistics (NoDaLiDa), 20–29. <aclanthology.org/2021.nodalida-main.3> (september 2022).

Och, Franz Josef & Hermann Ney (2003): A Systematic Comparison of Various Statistical Alignment Models. I: *Computational Linguistics* 29(1), 19–51.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li & Peter J. Liu (2020): Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. I: *Journal of Machine Learning Research* 21, 1–67.

Reimers, Nils & Irina Gurevych (2020): Making monolingual sentence embeddings multilingual using knowledge distillation. I: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. <arxiv.org/abs/2004.09813> (september 2022).

Sabet, Masoud Jalili, Philipp Dufter, François Yvon & Hinrich Schütze (2020): SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. I: *Findings of the Association for Computational Linguistics: EMNLP* 2020, 1627–1643. <dx.doi.org/10.18653/v1/2020.findings-emnlp.147> (september 2022).

Tiedemann, Jörg (2011): Bitext alignment. I: *Synthesis Lectures on Human Language Technologies* 4(2), 1–165.

Varga, Daniel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh & Viktor Trón (2005): Parallel corpora for medium density languages. I: *Proceedings of the RANLP* 2005, 590–596.

Magnus Breder Birkenes
Forskningsbibliotekar
dr.phil.
Nasjonalbiblioteket
Henrik Ibsens gt. 110
NO-0255 Oslo
magnus.birkenes@nb.no

Lars G. Bagøien Johnsen
Forskningsbibliotekar
dr.art.
Nasjonalbiblioteket
Henrik Ibsens gt. 110
NO-0255 Oslo
lars.johnsen@nb.no

Andre Kåsen
Forskningsbibliotekar
M.Sc.
Nasjonalbiblioteket
Henrik Ibsens gt. 110
NO-0255 Oslo
andre.kasen@nb.no

Fra partikkelverb og preposisjoner til verbavledninger og kasus. Brukerstudie av ei nordsamisk-norsk-nordsamisk ordbok

Katarzyna Dominczak, Lene Antonsen & Trond Trosterud

The article discusses challenges facing the bidirectional North Saami-Norwegian e-dictionary *Neahttadigisánit* when used as a production dictionary, based upon a set of writing and translation tasks presented to second term students and logging of their dictionary use. The dictionary's ability to both analyse and generate Norwegian and Saami forms is of great help to the students. The dictionary still has too strong a focus upon Saami, and a more thorough analysis of the morphological and the lexical components of the Norwegian part of the dictionary is needed.

1. Introduksjon

Artikkelen analyserer hvilke svakheter og muligheter *Neahttadigisánit nordsamisk-norsk-nordsamisk ordbok* (NNNNO) har som digital produksjonsordbok for norskspråklige studenter som skal lære seg nordsamisk, et språk som typologisk sett er svært ulikt norsk. 33 studenter på nordsamisk innføringskurs II ved UiT Norges arktiske universitet deltok i en brukerstudie bestående av to produksjonsoppgaver: ei skriveoppgave og ei oppgave med oversetting fra norsk til nordsamisk. Fokuset i oppgavene var transitivitet, verb som krever adverbial i bestemt kasus, og partikkelverb. I tillegg besvarte studentene ei refleksjonsoppgave om bruken av ordboka.

Ordbokbruk har etter hvert blitt et etablert forskingsfelt, spesielt etter at e-leksikografien ga tilgang til å forske på ordboklogger. Litteraturen er framfor alt konsentrert om enspråklige ordbøker for majoritetsspråk, som LexicoNordicas temanummer i 2008 (se

Bergenholtz & Malmgren 2008) og Robert Lews oversiktsartikkel (Lew 2011). Fokuset i denne artikkelen er språkinnlæreres bruk av ei tospråklig ordbok til tekstproduksjon på et minoritetsspråk som typologisk sett står langt unna morsmålet deres. Utfordringene deres er i liten grad drøfta i litteraturen om ordbokbruk, og vi refererer dermed ikke til den her.

Kapitlene 2 og 3 presenterer de mest relevante typologiske forskjellene mellom nordsamisk og norsk og deretter NNNNO, ordboka i brukerstudien. Kapitlene 4 og 5 presenterer innsamlinga og analyse av data fra brukerstudien. Kapittel 6 analyserer studentsvara i refleksjonsoppgava. I konklusjonen viser vi at ordboka i større grad enn i dag bør være eksplisitt i å presentere de grammatiske egenskapene til de norske oppslagsorda når det gjelder transitivitet og valg av kasus. I utforminga av lemmaartiklene er det avgjørende at informasjonen om valens i nordsamisk er entydig.

2. Typologisk sett ulike språksystemer

Nordsamisk har rik morfologi. Verba har 45 finitte bøyningsformer – som følge av at nordsamisk har 3 personer x 3 tall i 4 modi og dessuten tempusbøyning i moduset indikativ. Morfologien er komplisert, med morfofonologiske vekslinger som diftongforenkling i første stavelse, stadieveksling i konsonantsentrum og vokalveksling i trykklett stavelse. Nordsamisk har seks kasus. Nomen i illativ (‘til’) og lokativ (‘i, på, fra’) danner adverbial alene, der man på norsk bruker preposisjon. Også adposisjoner har illativisk eller lokativisk betydning, f.eks. *ala* (‘i retning på’) og *alde* (‘på’).

På norsk kan mange verb være både transitive og intransitive, som *rulle* og *velte*. I andre tilfeller er det to ulike verb med homonymi i infinitiv. For eksempel har *brenne* formene *brente* (transitiv) og *brant* (intransitiv) i preteritum. På nordsamisk blir transitiv og intransitiv handling alltid uttrykt med ulike verb.

Samiske språk har et rikt system av verbavledninger, som kan deles inn i to hovedtyper: grammatiske avledninger, som endrer verbets valens, og aspektuelle avledninger. Istedenfor norsk perifrastisk passiv har samisk ulike verbavledninger med hver sine konnotasjoner. I nordsamisk uttrykker suffikset (*o*)(*j*)*uvvot* (som i *borrojuvvot* 'bli spist') bare at det finnes en underforstått agens i ytringa. Suffikset *hallat* (som i *vuojahallat* 'bli overkjørt') uttrykker i tillegg at patiens har opplevelsen av å være uheldig. Det er også verbavledninger som uttrykker refleksivitet, som *čiehkat* > *čiehkádit* ('gjemme' > 'gjemme seg') og resiprokitet, som *deaivat* > *deaivvadit* ('treffe' > 'treffe hverandre'). Mange aspektuelle verbavledninger tilsvarer verb pluss adverb på norsk, som *vázzilit* ('gå av gårde'). Men norske partikkelverb kan også ha overført betydning, som *komme an på* og *komme seg*, som ikke kan oversettes med den nordsamiske ekvivalenten av verbet (*boahtit*) og heller ikke med ei verbavledning av dette.

3. Ordboka NNNNO og NDS-plattforma

Dagens ordbøker mellom norsk og nordsamisk har opphav i to tradisjoner. Den ene er ordbøkene til Kåven et al. (1995) og Kåven (2000) mellom nordsamisk og norsk, som henter den nordsamiske lemmalista si fra Pekka Sammallahti (1989). Den andre tradisjonen tar utgangspunkt i Nils Jernslettens nordsamisk-norsk *Álgosátnegirji* (1991). Forfatteren ga denne ordboka til den språkteknologiske forskingsgruppa Giellatekno i Tromsø, og den blei dermed utgangspunkt for nettordboka NNNNO, som analysene i denne artikkelen baseres på.

Ordboka er likevel sterkt utvida: Mens *Álgosátnegirji* inneholder 4332 nordsamiske lemma, inneholder NNNNO 40 500 nordsamiske lemma (10 000 av dem er egennavn, hvorav halvparten er norskspråklige navn som er med for å kunne gi bøyningspa-

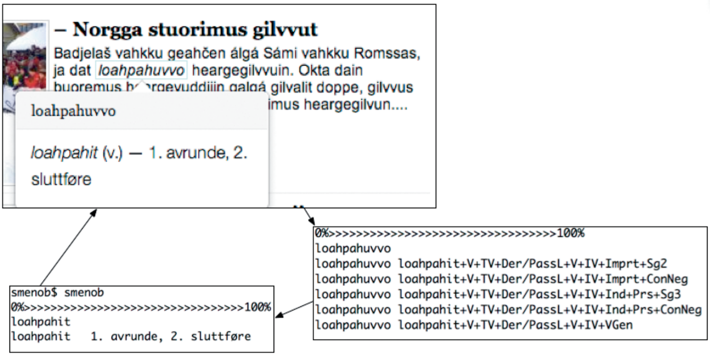
radigme). Den norsk-nordsamiske delen av NNNNO blei snudd fra nordsamisk-norsk, men snuoperasjonen skjedde relativt tidlig heller enn på ei ferdig ordbok, og storparten av lemmaartiklene er laga for to atskilte ordbøker. Nye norsk-nordsamiske ordartikler har blitt lagt til basert på norske eller tospråklige frekvenslister fra både dagligspråk og ulike domener. Blant annet blei det lagt til ordpar fra offentlig forvaltning i regi av et prosjekt finansiert av det daværende Forbruker- og administrasjonsdepartementet (jf. Gerstenberger, Eskonsipo & Eira 2013). Dagens norsk-nordsamiske ordbok inneholder 25 000 lemma.

I den første digitale versjonen av ordboka, *Vuosttaš Digisánit*, blei ordbokmaterialet utvida ved at bøyningsformer blei generert til en database, både for treff i ordboka og for bøyningsparadigmer. Ordboka var ikke online, men fungerte i de fleste programma på datamaskina (dette er nærmere presentert i Johnson, Antonsen & Trosterud (2013)).

Nå hører NNNNO til NDS-plattforma[1], ei online ordbokplattform som opprinnelig blei laga for ordbøker mellom nordsamisk og norsk og sørsamisk og norsk, og som nå er utvida til 31 språkpar. Den tekniske løsninga i NDS-plattforma innebærer at innputt først blir sendt til den morfologiske analysatoren Giella-sme[2] for nordsamisk og til Giella-nob for norsk. Analysatoren finner lemma, som igjen sendes for treff i ordbokfila, som illustrert i figur 1.

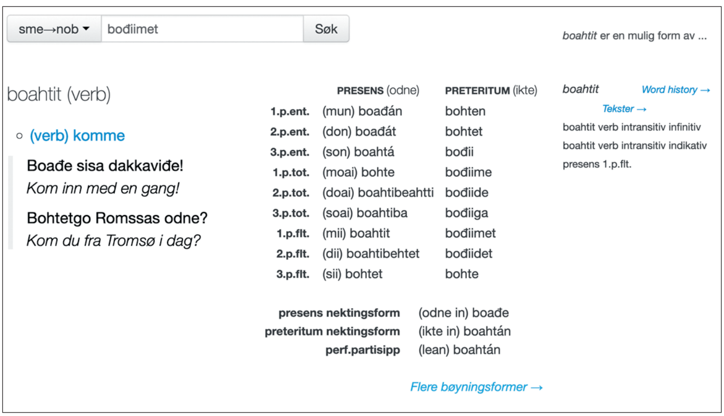
Den morfologiske analysen gjør ordboka tolerant overfor bøyde former og også overfor vanlige skrivefeil, som er lagt til analysatoren. Brukeren kan også bruke ordboka via en lesehjelpapplikasjon som legger seg i bokmerkelinja på nettleseren. Når brukeren klikker på et ord i teksten, sendes ordet til analyse, og oversettelsen vises i en dialogboks, men uten ekstra informasjon.

1 NDS er en forkortelse for *Neahttadigisánit*. Det er ordbøker for mange språk i plattforma, og derfor kaller vi den nå bare NDS.
2 Giella-sme er nærmere beskrevet i Antonsen & Trosterud (2017).



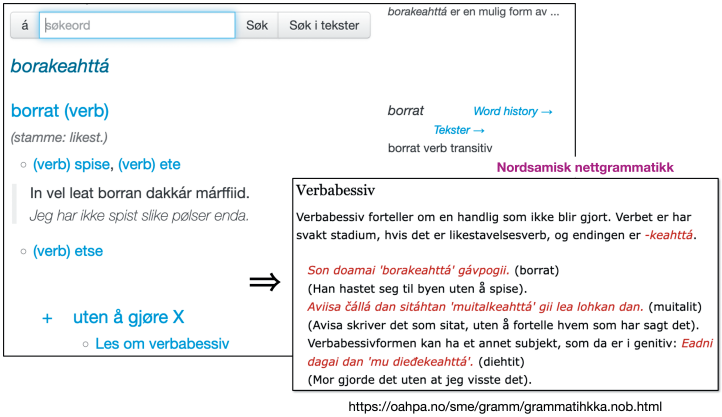
Figur 1: I NDS-plattforma blir ordbokoppslaget sendt til en morfologisk analysator før lemmaet blir slått opp i ordboka.

Brukeren kan gå fra norsk lemmaartikkel til nordsamisk lemmaartikkel, og motsatt veg, ved å klikke på lemmaet som foreslås som oversettelse. Ved at man klikker på det nordsamiske lemmaet i nordsamisk-norsk ordbok, blir bøyningsparadigmet generert ved hjelp av Giella-sme og et ferdig oppsett for hvilke former som skal genereres, se figur 2. Ordboka er i flittig bruk. I perioden mars 2019–februar 2020 var det over 1,4 millioner oppslag i NNNNO (Antonsen & Trosterud 2020:11). Om bruken av ordbøkene på NDS-plattforma, se også Trosterud (2019) og Eskonsipo (2020).



Figur 2: Artikkel for verbet *boahtit*, med generert bøyningsparadigme. Brukeren har søkt med ordforma *bođiimet*.

Den morfologiske analysen av innputt gir ikke bare lemma, men også tagger for morfologi. For de mest produktive avledningene og for noen infinitte verbformer genererer plattforma en forklaring til lemmaartikkelen. Et eksempel er *borakeahttá* (et av orda i oppgava som studentene fikk), som gir analysen borrat+V+TV+VAbess, hvor både lemmaet *borrat* ('å spise') og taggen *VAbess* får treff i ordbokleksikonet. *VAbess* angir at ordet er verbabessiv, og dette presenteres for brukeren med formuleringa «uten å gjøre X» og lenke til en nettgrammatikk, jf. figur 3 (alle henvisninger til nettgrammatikk gjelder Antonsen & Baal (2011–2022)). (For samspillet mellom Giellasme og NDS-plattforma, se Antonsen et al. 2009 og Antonsen 2018.)



Figur 3: Oppslag for verbabessiv *borakeahttá*, med generert informasjon «uten å gjøre X» og lenke til nettgrammatikken (Antonsen & Baal 2011–2022).

NDS-plattforma genererer også lenker merka «Søk i tekster» fra nordsamisk lemma eller ordform til tekstkorpuset SIKOR, som hadde ca. 30 millioner ord da denne brukerstudien blei gjennomført. Treff på norske ord blei lenka til et atskillig mindre tospråklig korpus på 3,4 millioner ord.

I tillegg til ordbøkene på NDS-plattforma fins det også andre e-ordbøker for nordsamisk. Forlaget Davvi Girji har publisert lemma og oversettelser (men ikke forklaringer og eksempelsetninger) fra papirordbøkene sine som et elektronisk ordbokpar nordsamisk-norsk-nordsamisk (*Davvi Girji digital ordbok*). Mikael Svonnis *Nordsamisk-svensk, Svensk-nordsamisk ordbok* (Svonni 2013) er tilgjengelig som en mobiltelefonapp. Pekka Sammallahtis nordsamisk-finske ordbok (Sammallahti 1989) er tilgjengelig via termportalen Sátni.org.

4. Analyse av tekster og logger

4.1. Utforming av brukerstudien

Brukerstudien besto av to skriftlige oppgaver med produksjon av tekst og var utforma basert på erfaringer fra samiskundervisning. I den første oppgava (skriveoppgava) skulle deltakerne skrive en kort tekst med tolv oppgitte nordsamiske verb som hadde ulik mengde informasjon i ordboka. Ti verb var i infinitiv, ett var en verbabessiv, og ett var en verbgenitiv. Den andre oppgava gikk ut på å oversette en tekst fra norsk til nordsamisk. Begge oppgavene inneholdt transitive og intransitive verb, partikkelverb, refleksive verb og verb som krever adverbial i bestemte kasus. I tillegg inneholdt begge oppgavene en refleksjonsdel hvor studentene kommenterte bruken av ordboka (se NDS-studie for detaljer). Deltakerne var 33 studenter på videregående innføringskurs i nordsamisk som fremmedspråk. 8 av disse gjennomførte oppgavene under tilsyn i et klasserom, og deres bruk av ordboka blei logga. De 25 andre deltakerne besvarte oppgavene uten tilsyn og uten logging av ordbokbruken. Med bakgrunn i begge produksjonsoppgavene ser vi på hvordan de ulike lemmaartiklene som studentene slo opp, påvirka studentenes leksikalske og grammatiske valg. Istedenfor å ta for

oss oppgavene hver for seg velger vi å se dem under ett og fokusere på bruken av ordboka.

4.2. Bruk av ordbokas grammatiske analyse

For en del infinitte verbtyper og ordavledninger genererer NDS-plattforma en generell informasjon som inneholder lenke til en nettgrammatikk, som vist i figur 3. I skriveoppgava fikk studentene oppgitt to infinitte verbformer som skulle brukes i setninger. Det ene var en verbabessiv, *borakeahttá*, av verbet *borrat* ('å spise'). Når studentene søkte på denne forma, kjente ordboka igjen grunnverbet og genererte lemmaartikkelen for denne sammen med ei forklaring basert på taggen for verbabessiv: «uten å gjøre X». Fra denne forklaringa går det ei lenke til nettgrammatikken, som inneholder utdypende forklaring og eksempelsetninger. Ei av eksempelsetningene inneholder *borakeahttá*. To tredeler av studentene brukte denne forma riktig.

Den andre infinitte verbforma var verbgenitiven *vácci* ('til fots'). Også for denne var ordbokas forklaring basert på lemmaartikkelen for *vázzit* ('å gå') pluss tagg (VGen) i analysen som genererte forklaringa «måten å gjøre noe på» og lenke til nettgrammatikken. I grammatikken forklares verbgenitiv, men uten eksempelsetning med *vácci*. En tredel av studentene oppfatta forma som den finitte verbforma *váccii* ('han/hun gikk'). Problemet her er tosidig: *Vácci* mangler suffiks og er derfor mindre gjenkjennbar som infinitt verbform, og forma representerer også en vanlig skrivefeil av den finitte forma. I lemmaartikkelen kommer informasjonen om verbgenitiv under presentasjonen av lemmaet, og man må dermed lese hele lemmaartikkelen for å forstå at dette ikke er et finitt verb.

4.3. Informasjon om kasus i eksempelsetninger

Mange av lemmaartiklene har eksempelsetninger for å vise valg av kasus for et mulig adverbial. På nordsamisk markerer man med illativ at en handling betegner retning mot noe, mens lokativ betegner en tilstand. Materialet viser at det ikke er uproblematisk å forstå bruken av illativ, og studentenes tekster viser at det er sammenheng mellom på den ene sida forekomsten av ei eksempelsetning og kvaliteten på den og på den andre sida studentenes valg av kasus.

I skriveoppgava var det mulig for studentene å la være å legge til adverbial. For verbet *čáhkat* ('å få plass') var det ingen eksempelsetning i lemmaartikkelen, og bare 5 av 28[3] studenter (fra nå av 5/28) la til adverbial i illativ ('å få plass i X'), som er riktig kasus. 9/28 studenter valgte lokativ, som er feil kasus, og 8/28 studenter brukte ikke adverbial i setninga. For verbet *čiehkat* ('å gjemme noe') var det ei eksempelsetning med postposisjonen *sisa* ('inn i'). Postposisjonen har illativisk betydning, men uten den prototypiske kasusendinga *-i*, som markerer illativ for substantiv. Dette gjør det vanskeligere å generalisere til at verbet krever illativ. 10/28 studenter unngikk å bruke adverbial, 3/28 studenter valgte illativ, og 7/28 studenter valgte lokativ, som var feil kasus. Hele 11/28 studenter brukte verbet som et refleksivt verb, noe som er feil. I lemmaartikkelen er verbet forklart med «transitiv: å gjemme noe», og det refleksive verbet *čiehkádit* står lenger nede i artikkelen med forklaringa 'intransitiv: å gjemme seg'.

I oversettelsesoppgava kunne ikke studentene unngå å legge til adverbial. I teksten de skulle oversette, var verba *å henge*, *å sette* og *å sette seg*, som alle krever adverbial i illativ. Alle lemmaartiklene for de tilsvarende nordsamiske verba hadde eksempelsetninger med adverbial i illativ med suffikset *-i*, og tallet for studenter som valgte illativ, var henholdsvis 14/24, 17/23 og 26/32.

3 Antall studenter som har brukt hvert verb, varierer.

4.4. Informasjon om transitivitet og valg av kasus

Studentenes besvarelser viser at de har problemer med å skille mellom transitive og intransitive verb. Transitive verb krever objekt i akkusativ. Oversettelsesoppgava krevde at studentene skulle skille mellom transitiv og intransitiv oversettelse av verba *å tørke* og *å henge*. Den norske teksten inneholdt formuleringa «tørke bort skiten», og for denne valgte alle studentene riktig nordsamisk verb. For den intransitive betydninga av *å tørke* var det ikke like bra resultat. I ordboka er verbet *goikat* merka «intransitiv», og bruken er illustrert med eksempelsetninga «klærne tørker». Likevel valgte 9/31 studenter det transitive verbet *goikadit*. Dette verbet er forklart med 'transitiv: gjøre tørr, både i lufta og med apparat', og eksempelsetninga omhandler klær.

For verbet å henge gir ordboka to betydninger med hvert sitt sett av ekvivalente verb i nordsamisk. Den ene er forklart med «intransitiv: være festet» og illustrert med en eksempelsetning. Den andre er forklart med «transitiv: få til å henge». 8/33 studenter valgte likevel intransitivt verb istedenfor riktig transitivt verb, og 7/33 studenter valgte transitivt verb istedenfor riktig intransitivt verb.

Forskjellen i resultata mellom de ulike instansene av *å tørke* og *å henge* viser at *å tørke bort noe* er lett å forstå som en konstruksjon med transitivt verb. I andre tilfeller er det imidlertid vanskelig å skille mellom transitive og intransitive betydninger av *å tørke* og *å henge*, til tross for forklaringer og eksempelsetninger i lemmaartiklene.

Studentene skulle også oversette det intransitive *å høres* (*gullot*) og transitivt *å høre* (*gullat*) til nordsamisk. I ordboka er begge verba presentert med eksempelsetninger med argumenter, men for *gullot* er nominativ illustrert med det ubestemte pronomenet *mihkkege* ('(ikke) noe'), som ikke følger bøyningsmønsteret for substantiv og dermed ikke har eksplisitt kasusmarkering. En

tredel av studentene valgte feil kasus for argumentet for *gullot*. At studentene trenger eksplisitt informasjon om kasusbøyning, vises også i hvordan de bruker det illativiske adverbet *olggos* ('ut'). Dette adverbet er brukt 40 ganger, og i 14 av tilfellene brukte studenten det feil, som et lokativisk adverb ('ute'), mens det riktige ville vært *olgun*. *Olggos* har *-s* som utlyd, noe studentene nok assosierer med lokativsuffikset (*-s*) fra nomenparadigmet.

Studentenes oversettelse av «På snora hang ei hvit skjorte» viser også problemet med å forstå forskjellen mellom nominativ og akkusativ. I denne setninga kommer subjektet etter verbet, og i oversettelsene valgte 14/31 studenter akkusativ, som er kasus for objekt, istedenfor nominativ, som er riktig kasus for subjekt. Årsaken er nok at objektet ofte kommer etter verbalet på nordsamisk. Slike feil viser at forståelsen av transitivitet og akkusativ versus nominativ ikke er internalisert hos studentene, og at forklaringer og eksempelsetninger i ordboka ikke er nok. Dette stemmer overens med resultatene i Dominczak (2020), som viste at språkinnlærere hadde ujevn progresjon i forståelse av transitivitet og akkusativ versus nominativ.

Studentene skulle også bruke verbet *deaivvadit*. Ordboka gir to betydninger og bruksmåter av verbet: 'å møtes' med bare subjekt, i tillegg til 'å møte', med informasjonen «intransitiv, krever komitativ». 14/30 studenter brukte verbet uten adverbial og 9/30 med adverbial i komitativ. 7/30 studenter brukte verbet som et transitivt verb med objekt i akkusativ, noe som er feil. Det transitive verbet er *deaivat*.

4.5. Oversetting av partikkelverb i norsk til nordsamisk

I skriveoppgava skulle studentene bruke verbet *ohcat*, og ifølge ordboka betyr verbet 'å lete' eller 'å søke'. Noen studenter brukte verbet flere ganger i skriveoppgava, slik at det til sammen var 53 tilfeller av *ohcat* i tekstene. I 33 tilfeller blei verbet brukt i betydninga

'lete etter', og i 6 av disse la studentene til et adverb etter verbet, *ohcat bearrái* (5 tilfeller), *ohcat maŋŋel* (1 tilfelle), noe som blir feil på nordsamisk.[4] Ordbokartikkelen for *ohcat* gir ikke forklaring på bruken og heller ingen eksempler. Tilsvarende er det i artikkelen for verbet *å lete* (som det er lenket til i *ohcat*-artikkelen). Derimot var det to eksempelsetninger i artikkelen for verbet *å søke*, og de fire studentene som brukte verbet med betydninga 'å søke om noe', valgte riktig kasus (illativ) og unngikk å bruke adverb.

Dette viser at norske partikkelverb kan være vanskelige å oversette når de ikke er forklart i ordboka. Det finnes 565 partikkelverb med egne oppslag, som *å kle på seg*. De aller fleste av disse er opprinnelig oversettelser fra den nordsamisk-norske ordboka, som har blitt til lemma i prosessen med å snu ordboka til norsk-nordsamisk.

4.6. Problemer med oppslag fra norsk

I oppgave 2 skulle studentene oversette «det ble en stygg flekk igjen på skjorta». Dette var vanskelig for studentene. Bare 6/33 studenter brukte riktig verb, *báhcit*, og av disse var det bare to som brukte argument i riktig kasus, illativ. Her var det i ordboka lagt til misvisende informasjon om *báhcit*. Betydninga var forklart som «bli (med vilje)», noe som er feil fordi verbet også kan brukes med et inanimat (dvs. viljeløst) subjekt. Dette var kanskje grunnen til at 12/33 studenter oversatte med *šaddat* (som betyr '*å bli*' i betydninga at noe vokser, eller at noe endrer seg til noe), mens andre skrev setninga om. 16/33 studenter la til adverbet *fas* ('om igjen'), noe som ikke passer i denne sammenhengen. På norsk brukes verbet å bli også i translativ betydning og i perifrastisk passivkonstruksjon, mens man på nordsamisk i tilsvarende tilfeller bruker avleda verb.

4 Adverbet og postposisjonen *bearrái* blir brukt primært sammen med verb og kan oversettes med 'etter' eller 'med', som i *geahčča bearrái* 'se etter, holde øye med'. Adverbet og adposisjonen *maŋŋel* betyr (temporal) 'etter', som i *dálvvi maŋŋel* 'etter vinteren'.

Dette bør komme fram i ordboka, med lenker til nettgrammatikken.

Ei oppgave var også å oversette «satt ute rundt et bord». Når man søker på *rundt* i ordboka, kommer adjektivet *rund* før preposisjonen *rundt*. To av studentene brukte dermed adjektivet *jorbbas* ('rund'). Her er det to problemer: Ordboka burde ha vist treff der søkeordet er lik lemma, dvs. preposisjonen, først. Det andre problemet er at ordboka er tolerant og gir treff på bøyningsforma *rundt* av *rund* og dermed gjør det mulig for studenten å velge feil lemma.

I tillegg til bøyningsformer som *rundt* gir den norske analysatoren også dynamiske sammensetninger, som fører til at brukeren får førsteledd og sisteledd oversatt hver for seg. Dette er en fordel i tilfeller der det sammensatte ordet ikke er et eget lemma i ordboka, og det kan også gi bedre innsikt i det nordsamiske ordet sjøl der ordboka inneholder ei leksikalsk sammensetning. Men dynamisk sammensetning med substantiv med bare to eller tre bokstaver som førsteledd kan gi feil analyse, som når *forsoning* blir sammensetning av substantivene *fôr* og *soning* og *donasjon* blir *do* + *nasjon*. Hvis orda *forsoning* og *donasjon* finnes i ordboka, blir de vist øverst, men de ukorrekte sammensetningsanalysene burde ikke vært vist i det hele tatt. Problemet kan løses ved at slike korte substantiv hindres i å delta i dynamisk sammensetning. Da bør man analysere et større tekstkorpus for å finne reelle sammensetninger som *forstoff* og *dolokk*, som da kan leksikaliseres i ordboka og analysatoren.

5. Logging av studentenes ordbokbruk

Åtte av studentene satt i et klasserom der bruken deres av ordboka blei logga, og vi så på ordbokoppslagene for disse studentene samla.

I gjennomsnitt slo hver student opp 139 ganger i ordboka i løpet av tida testen varte,[5] og 32 % av oppslagene var fra norsk til nordsamisk (de tilsvarende medianverdiene var 122 og 30,5 %). Andelen av oppslagene som var fra norsk, var noe lavere enn for gjennomsnittlig bruk av NDS-plattforma. I 2021 var 38 % av oppslaga i ordbokplattforma fra norsk til nordsamisk. Trass i at verb var en sentral del av oppgavene, var andelen norske ord blant verboppslaga lavere enn for oppslaga generelt. Bare 20 % av verba som blei slått opp i løpet av testen, var norske. Disse talla er i seg sjøl interessante: Sjøl om den største oppgava i testen var å oversette fra norsk til nordsamisk, brukte studentene likevel den nordsamiske ordboka mer enn dobbelt så ofte som den norske. Dette kan tyde på at den formelle hjelpa ordboka kan gi, er viktigere for studentene enn den semantiske: Heller enn å søke etter hva som tilsvarer et norsk ord, ser de på hvordan det nordsamiske ordet skrives, bøyes og brukes.

Problemet med å slå opp bøyde former av partikkelverb (jf. 4.6) viste seg også i ordbokloggen. 39 av 350 oppslag (11 %) i den norsk-nordsamiske delen av NNNNO var flerordsuttrykk, og av dem hadde 31 infinitiv og 3 bøyde former (*listet seg*, *tar på seg*, *tok på seg*). Dette er en høyere andel enn i loggen for all bruk av NNNNO i 2021, der 7 % av de norsk-nordsamiske ordboksøka var flerordsuttrykk. Tilsvarende utgjorde søk med bøyde verbformer 8 % av flerordssøka i testen, mens tallet for hele 2021-loggen var 3 %[6] (av totalt 90 308 flerordssøk). Problemet med søk på partikkelverb i bøyd form er at sjøl om NNNNO omdirigerer søk på både *går*, *gikk* og *gått* til lemmaet **gå**, er det samme ikke mulig for oppslagsformene *gikk på ski*, *gikk seg vill*, *gikk rundt* og *gikk ned*,

5 Som «oppslag» registrerer loggen både vanlig søk på ordform og en eventuell oppfølgende forespørsel om å få se bøyningsparadigmet til det samiske ordet. I tillegg er oversettelsene i lemmaartiklene klikkbare, slik at ordbokbrukerne etter å ha slått opp på et samisk ord kan slå opp på ordbokartikkelen for den norske oversettelsen og omvendt.

6 Tallet ser bort fra ordformer som *for* og *så*, som kan være bøyde verbformer i tillegg til preposisjoner og adverb. Det viste seg imidlertid at i alle tilfeller unntatt ett var orda brukt som adverb eller preposisjon.

og slike oppslag gir ikke tilslag. Feilen gjelder få oppslag, men den er systematisk.

49 av studentenes oppslag i norsk-nordsamisk ordbok returnerte ingen treff. 13 av disse var nordsamiske ord, 7 var resultat av bruk av nordsamisk tastatur, der bokstaven *ŧ* er plassert på posisjonen til bokstaven *y*, som ikke er i bruk i nordsamisk (bl.a. blei *flŧ* tastet for *fly*), og 3 var skrivefeil (bl.a. *forsvnat*). 12 ord var flerordsuttrykk (bl.a. *om våren*, *det var en gang*, *slippe ut*), og 11 ord mangla (bl.a. *bagasjerom*, *flyvende*, *gjetning*).

6. Analyse av spørreskjema

32 av 33 studenter besvarte to refleksjonsoppgaver, som begge inneholdt tre åpne spørsmål: (I) «Hvordan har du brukt nettordboka mens du arbeidet med oppgaven?», (II) «Hva har vært spesielt nyttig for deg?» og (III) «Var det noe som manglet?». 11 av de 32 studentene svarte på samisk, og 21 på norsk.

Studentene rapporterte om tre hovedbruksområder: oversetting (alle), bøyningsparadigmer (29/32) og eksempler på bruk på samisk, med semantisk og syntaktisk informasjon (11/32). 13 fortalte at de søkte på bøyd ordform, og 10 at de brukte ordboka til å sjekke rettskrivinga for samiske ord de kjente fra før, men var usikre på hvordan staves (f.eks. om det skulle være *a* eller *á*). Fem studenter fortalte at de brukte lenker til nettgrammatikken og SIKOR.

Samtlige som oppga at det var nyttig å se eksempelsetninger i ordbokartikkelen, etterlyste flere eksempler og forklaringer. En svakhet som blei nevnt, var at noen ord ikke fantes i ordboka, f.eks. norske flerordsuttrykk (bl.a. *ta på seg*), avledede substantiv (bl.a. *brøl*) og adjektiv (bl.a. *flygende*). For hvert lemma er det ei lenke til et bøyningsparadigme bestående av de mest sentrale ordformene, og derfra er det ei ny lenke til det fullstendige paradigmet («Flere

bøyningsformer») på Giellatekno si nettside. Noen kommenterte at det var vanskelig å forstå de grammatiske taggene der, f.eks. *V Cond Prs Du2* (verb kondisjonalis presens 2. person totall).

Studentene var fornøyde med eksempelsetningene, bøyningsparadigmene, muligheten til å søke opp noen flerordsuttrykk på norsk og ordbokas fleksibilitet til å godta ord uten samiske diakritiske tegn når brukeren ikke huska den eksakte skrivemåten. I tillegg blei grammatikkforklaringene i ordbokartikkelen og lenkene til nettgrammatikken sett på som nyttige.

7. Diskusjon og konklusjon

I likhet med andre norsk-nordsamiske ordbøker har NNNNO hovedfokus på samisk ordforråd og morfologi, uten å ta norsk språkstruktur nok på alvor. Analysen viser at svakheter i ordboka blir reflektert i ordbokbrukernes språkbruk.

Ved å ta høyde for vanlige skrivefeil og bøyde former kommer ei tolerant ordbok brukeren i møte, men toleransen gir også mange flere treff for brukeren å velge mellom, og analysen viser at brukeren derfor noen ganger velger feil oversettelse. Analysen viser også at brukeren forventer at ordboka er tolerant når den ikke er det, dvs. når norske partikkelverb ikke blir gjenkjent i preteritum eller perfektum.

Morfologisk analyse av søkeordet gir i mange tilfeller dynamisk sammensetning med substantiv. Hvis førsteleddet er på to eller tre bokstaver, er analysen ofte feil. Feilanalyser som *for + soning* og *do + nasjon* kan unngås ved at slike substantiv hindres i å delta i dynamiske sammensetninger.

Det er som nevnt bare 565 partikkelverb som oppslagsord i den norsk-samiske ordboka, og som regel er de opprinnelig oversettelser av nordsamiske lemma (jf. 4.5 slutten). Partikkelverb må tas på alvor, og bør enten bli lemma både i analysatoren og i ordboka,

eller de bør legges til i verbartikkelen. Dette bør gjøres systematisk ved hjelp av norske ordbøker og korpussøk. Verb som *bli* og *være* må i artikkelen behandles både som hovedverb og som hjelpeverb i perifrastiske verbformer. Lenker til nettgrammatikken vil være nyttig for å formidle informasjon om egenskapene til slike verb.

Opplysninger om reksjon bør være eksplisitte. Analysen viser at eksempelsetninger ikke alltid er nok, spesielt hvis argumentet ikke er et substantiv. Tydelig informasjon i form av eksempelsetninger med substantiv hjelper brukeren med å velge riktig kasus for adverbialer. Men sjøl om artiklene inneholder eksplisitt informasjon om transitivitet i tillegg til eksempelsetning, er både valget mellom transitivt og intransitivt verb og mellom nominativ og akkusativ i subjekt og objekt vanskelig å forstå for ordbokbrukerne, og det må læres utafor ordbokas rammer. Informasjon om noen av de infinitte verbformene gis via ei generell forklaring og lenke til nettgrammatikken. Spesielt for verbgenitiv, som ikke har et eksplisitt suffiks, er ikke dette nok. Denne verbforma er ikke spesielt produktiv. Det er bare ca. 80 verb som brukes som verbgenitiv (Antonsen 2018:65), og man kunne gjøre dem tydeligere for brukerne ved å leksikalisere verbgenitiver både i analysatoren og i ordboka.

Litteratur

Ordbøker

Davvi Girji digital ordbok. Kárášjohka: Davvi Girji. <533.davvi.no/> (juli 2022).

Jernsletten, Nils (1991): *Álgosátnegirji sámi-dáru sátnegirji = Samisk-norsk ordbok*. Kárášjohka: Davvi Girji.

Kåven, Brita, Johan Jernsletten, Ingrid Nordal, John Henrik Eira & Aage Solbakk (1995): *Sámi-dáru sátnegirji = Samisk-norsk ordbok*. Kárášjohka: Davvi Girji.

Kåven, Brita E. (red.) (2000): *Stor norsk-samisk ordbok = Dáru-sámi sátnegirji*. Kárášjohka: Davvi Girji.

NNNNO = Lene Antonsen, Trond Trosterud & Berit Merete Nystad Eskonsipo (2013–2022): *Neahttadigisánit Davvisámi-dáru-davvisámi sátnegirji = Neahttadigisánit Nordsamisk-norsk-nordsamisk ordbok*. Tromsø: UiT Norges arktiske universitet. <sanit.oahpa.no> (juli 2022).

Sammallahti, Pekka (1989): *Sámi-suoma sátnegirji = Saamelais-suomalainen sanakirja*. Ohcejohka: Jorgaleaddji.

Sátni.org. Ordbokplattform for samisk termwiki og flere samiske ordbøker. Tromsø: UiT Norges arktiske universitet. <satni.org> (juli 2022).

Svonni, Mikael (2013): *Davvisámegiela-ruotagiela, ruotagiela-davvisámegiela sátnegirji = Nordsamisk-svensk, svensk-nordsamisk ordbok*. Deatnu: ČálliidLágádus.

Annen litteratur

Antonsen, Lene (2018): *Sámegielaid modelleren – huksen ja heiveheapmi duohta giellamáilbmái* [Abstract: *Modeling Saami languages. Construction and adaptation to real-world linguistic issues*]. Ph.d.-avhandling. Tromsø: UiT Norges arktiske universitet.

Antonsen, Lene & Trond Trosterud (2017): Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. I: *Norsk lingvistisk tidsskrift* 35:2, 153–185.

Antonsen, Lene & Trond Trosterud (2020): Med et tastetrykk. Bruk av digitale ressurser for samiske språk. I: *Samiske tall forteller* 13, 53–67. Kautokeino: Sámi allaskuvla.

Antonsen, Lene, Ciprian Gerstenberger, Sjur Moshagen & Trond Trosterud (2009): Ei intelligent elektronisk ordbok for samisk. I: *LexicoNordica* 16, 271–283.

Antonsen, Lene & Berit Anne Bals Baal (2011–2022): *Nordsamisk nettgrammatikk*. UiT Norges arktiske universitet. <oahpa.no/sme/gramm/grammatihkka.nob.html> (juli 2022).

Bergenholtz, Henning & Sven-Göran Malmgren (2008): Ordbogsbrug i Norden. I: *LexicoNordica* 15, 1–4.

Dominczak, Katarzyna Zofia (2020): *Nominatiiva vai akkusatiiva? Davvisámegiela studeanttaid kásusoahppan*. Masteroppgave i samisk språkvitenskap. UiT Norges arktiske universitet. <munin.uit.no/handle/10037/19353> (oktober 2022).

Eskonsipo, Berit Merete Nystad (2020): *Sátnegirjegeavaheami čalmmustahttin neahttasátnegirjji loggafiilla analysa bokte*. Masteroppgave i samisk språkvitenskap. UiT Norges arktiske universitet. <munin.uit.no/handle/10037/18505> (oktober 2022).

Gerstenberger, Ciprian, Berit Merete Eskonsipo & Márjá Eira (2013): *Digging for domain-specific terms in North Saami*. Presentasjon på konferansen Oovtâst, Inari. <giellatekno.uit.no/publications/fad_inari2013.pdf> (juli 2022).

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

Lew, Robert (2011): Studies in dictionary use: Recent developments. I: *International Journal of Lexicography* 24, 1–4.

NDS-studie. <giellatekno.uit.no/research/ndsstudie.html> (september 2022).

SIKOR = *Samisk Internasjonalt KORpus*. UiT Norges arktiske universitets og det norske Sametingets samiske tekstsamling, versjon 01.12.2021. <gtweb.uit.no/korp/> (juli 2022).

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

Katarzyna Dominczak
ph.d.-student
UiT Norges arktiske universitet
NO-9037 Tromsø
katarzyna.z.dominczak@uit.no

Lene Antonsen
førsteamanuensis
UiT Norges arktiske universitet
NO-9037 Tromsø
lene.antonsen@uit.no

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

Nya tider, nya möjligheter – inför en reviderad version av SAOB i en helt ny tid

Pär Nilsson & Bodil Rosqvist

The *Swedish Academy Dictionary* (SAOB) will be completed next year (i.e. 2023). Since the project started already in the 1890s a large part of the dictionary will be outdated as soon as it is complete – as a natural consequence of the passing of time. It has therefore been decided that there will be a revision of the SAOB after its completion, with the goal to transform the current digital version into an updated, functional and user friendly online edition. In this article, we discuss user input of different kinds and outline some changes that could be made to the structure and content of the dictionary.

1. Inledning

Ordbok över svenska språket utgiven av Svenska Akademien (SAOB) är en historisk och samtidsspråklig ordbok som beskriver skriven svenska från 1521 fram till nutid. På grund av verkets långa utgivningstid har ”nutid” i realiteten kommit att avse olika årtal för olika delar av alfabetet: A–ANLÖPNING (band 1) publicerades exempelvis år 1898 och VRETT–ÅVÄXT (band 38) år 2021. År 2023 förväntas hela ordboken vara färdigställd från A till Ö. Men färdigställandet innebär inte slutet på ordboksprojektet, utan snarare startskottet till ett helt nytt arbete med nya förutsättningar. I denna artikel diskuteras ordboksredaktionens preliminära planer inför en kommande uppdatering av SAOB.

En digitaliserad version av SAOB föreligger redan idag på nätet, dels på den egna hemsidan saob.se, dels – tillsammans med SAOL14 (2015) och SO (2021) – på ordboksportalen svenska.se. Denna första version kommer även fortsättningsvis att vara åtkomlig. Tanken är att den framtida versionen ska bli en mer ren-

odlat digital ordbok. En självklar målsättning är att denna ordbok ska vara användbar och relevant för en så bred målgrupp som möjligt.

2. Från ordboksband i bokhyllan till databas på nätet

Fram till år 1997, när SAOB blev tillgänglig på nätet i en första version (den s.k. OSA-versionen), fick flertalet av dess användare bege sig till ett bibliotek eller en universitetsinstitution för att kunna slå upp i ordboken. Användaren hade således i regel en viss kännedom om det verk han eller hon sökte information i, och var förmodligen också villig ”att egna en stund till att göra sig hemmastadd i dess beteckningssätt”, så som det första förordet uppmanade till (saob.se/om/).

När SAOB nu kan nås på ett betydligt enklare sätt, via någon av Svenska Akademiens fritt tillgängliga webbsidor, är antalet användare betydligt större, men med all sannolikhet har kunskapen om SAOB inte ökat i motsvarande grad. Numera hamnar många i någon av SAOB:s artiklar genom en sökning på Google eller annan sökmotor, och känner kanske inte till något om det verk man har klickat sig vidare till. För att SAOB ska vara relevant och ändamålsenlig idag och i framtiden behöver den således vara möjlig att förstå för den tillfällige besökaren, samtidigt som den ska erbjuda den återkommande och målmedvetna användaren förväntad komplexitet och vetenskaplig noggrannhet.

Ett framtida SAOB-projekt innebär därmed inte bara att den befintliga ordboken behöver kompletteras med ord, belägg, betydelser och användningar som har tillkommit under verkets utgivningstid – utan också att innehållet behöver presenteras på ett tydligare och mer överskådligt sätt och med förbättrade sökmöjligheter. Målet är en användarvänlig och funktionell digital ord-

bok som uppdateras kontinuerligt. Det är därför mycket viktigt för redaktionen att ta del av (potentiella) användares synpunkter och önskemål rörande ordbokens utformning.

Ordböcker har naturligtvis alltid utformats med tanke på hur de ska användas. En av Samuel Johnsons ”guiding principles” som ofta citerats är att ”the value of a work must be estimated by its use” (Johnson 1747 enl. Atkins & Rundell 2008:5). Idag står emellertid användarnas behov ännu tydligare i centrum när ordböckers utformning diskuteras, både i lexikografiska handböcker (se t.ex. Svensén 2004:26 och Atkins & Rundell 2008:4–5) och inom lexikografisk teori och forskning (se t.ex. Tarp 2008:44–50 och Abel & Meyer 2013:179–180).

3. Återkoppling från användare

Målsättningen är alltså att den framtida versionen av SAOB ska vara användbar och begriplig för fler användare än sin föregångare, samtidigt som kärnanvändarna, exempelvis språkforskare och filologer, fortsatt ska kunna finna den information de förväntar sig. Den nya teknikens möjligheter bör tas till vara och erbjuda nya ingångar till ordboksinnehållet för alla SAOB:s besökare.

För att komma i kontakt med så många potentiella användare som möjligt har redaktionen under de senaste åren deltagit som föredragshållare i en rad seminarier där olika perspektiv på revideringen av SAOB har diskuterats. Vidare har redaktionen alltid fått viss återkoppling från användare via telefonsamtal och brev, och numera även via ett kontaktformulär på hemsidan. Vi har på detta sätt fått påpekanden om felaktigheter som behöver rättas och idéer om hur ordboksinnehållet skulle kunna presenteras tydligare.

I avsnitt 3.1 diskuteras några önskemål som kommit redaktionen till del genom kontakt med användare, s.k. *explicit feedback*.

Därefter ser vi (i avsnitt 3.2) på användarstatistik från Google Analytics, s.k. *implicit feedback* (Abel & Meyer 2013:180).

3.1. Revideringsseminarier och annan användarkontakt

Sedan hösten 2017 har företrädare för redaktionen deltagit som föredragshållare i ordinarie forskarseminarier för svenska (eller nordiska språk) vid olika universitet, närmare bestämt i Lund, Göteborg, Växjö, Stockholm, Uppsala och Umeå. Där har olika aspekter av SAOB:s innehåll presenterats och diskuterats, med en uppmaning i seminarieinbjudan om att deltagarna ska passa på att komma med synpunkter ”eftersom det för första gången på 140 år är möjligt att tänka i nya banor vad gäller artiklarnas utformning och innehåll”. Teman som har tagits upp är bl.a. förkortningar, bruklighetsangivelser, grammatisk terminologi, beskrivningen av fasta uttryck och den viktiga frågan om vilket material en uppdaterad historisk ordbok bör grunda sig på, liksom – inte minst – hur ordboksartiklarna kan göras överskådligare.

Ett särskilt seminarium har också anordnats i redaktionens lokaler, i samband med konferensen Svenska språkets historia i november 2021. Där diskuterades SAOB:s innehåll ur olika synvinklar av flera olika föredragshållare. Bl.a. gjorde företrädare för Språkrådet i Stockholm och Institutet för de inhemska språken i Helsingfors en gemensam genomgång av vilka förändringar de önskar se i en framtida uppdaterad SAOB. Här framkom t.ex. att ordbokens många förkortningar upplevs försvåra läsningen, att det ibland kan vara svårt att skilja olika informationskategorier från varandra och att definitionerna är kompakta och svårlästa. En tydlig separering och rubricering av olika informationskategorier önskas, liksom sammanfattande översikter och dessutom en tydlig användarguide. Fler språkliga exempel efterlystes också, särskilt i fråga om sammansättningar som ofta är mycket kortfattat beskrivna. Vidare efterfrågades en ”uppdateringslogg”. Användaren bör

kunna se när en artikel är publicerad resp. reviderad. I ett annat seminariebidrag fastslogs att SAOB:s uttalsuppgifter är utformade enligt ett ålderdomligt och egenartat system. Rekommendationen blev att uttalsuppgifterna i SAOB helt enkelt kan utgå, eftersom en revision av dem skulle vara alltför arbetskrävande för det mervärde som eventuellt skulle kunna uppnås.

Liknande synpunkter har också kommit redaktionen till del genom telefonsamtal, brev och meddelanden via kontaktformulär. Flera enskilda felaktigheter har också rapporterats in under årens lopp, som t.ex. att den engelska pluralform av ordet *hobby* som vanligen används i svenska är *hobbies*, inte *hobbys*. Vi har också fått förslag på ord som saknas i SAOB och bör läggas till, t.ex. *försvarsvilja*, *isterband* och *grattis*. Det är först nu när den första upplagan av SAOB snart är färdigställd som dessa synpunkter kan leda till konkreta förändringar i SAOB:s innehåll.

En användarundersökning av Svenska Akademiens ordböcker på nätet har gjorts av Bäckerud, Nilsson & Sköldberg (2020). Bland annat analyserade de frågor och synpunkter från användare av sidorna saob.se och svenska.se som kommit in under år 2018. Det visade sig att cirka 30 % av meddelandena som var riktade till SAOB utgjordes av kommentarer eller frågor om ett specifikt ord. Cirka 15 % var frågor kring hur ordbokens information är strukturerad och hur beskrivningsspråket ska tolkas; användarna ville t.ex. ofta ha hjälp med att uttyda förkortningar eller uttalsuppgifter. Den största andelen meddelanden som skickats, cirka 50 %, utgjordes av påpekanden om konkreta felaktigheter i artiklarna (Bäckerud, Nilsson & Sköldberg 2020:97–98). Här ingick även klagomål på (ålderdomliga) definitioner som representerar en annan tid och som uppfattas som stötande och olämpliga idag. Påpekanden från användarna hjälper alltså redaktionen att identifiera vilka ändringar som bör prioriteras i den reviderade digitala utgåvan av SAOB.

Redaktionen har redan fattat preliminära beslut om en del förändringar. Bl.a. har ett enhetligt och tydligt operationaliserat system för bruklighetsangivelser utarbetats (se vidare Larsson 2020).

Olika förslag om hur informationskategorier på bästa sätt kan delas upp och rubriceras för att öka översiktligheten har också tagits fram. Några av dessa kommer att diskuteras vidare i avsnitt 4 nedan (se även Rosqvist & Wendt 2018).

3.2. Användarstatistik från saob.se och svenska.se

Den statistik som kommer att diskuteras i detta avsnitt har tagits fram av SAOB:s IT-ansvarige, Erik Bäckerud, med hjälp av verktyget Google analytics. Siffrorna avser år 2020 och/eller 2021, och ibland mer specifikt vecka 50 under något av dessa år. Statistiken kan därmed jämföras med siffror från 2018 och vecka 50 det året, som analyserats av Bäckerud, Nilsson & Sköldberg (2020). Den enhet som diskuteras härvid är *besök*. Ett besök påbörjas när en webbsida visas för en användare och avslutas när användaren har varit inaktiv i 30 minuter. Om samma användare öppnar en ny sida när 30 minuter har passerat räknas detta som ett nytt besök.

Under hela år 2021 var antalet besök på ordboksportalen svenska.se 6,4 miljoner. Detta kan jämföras med 4,4 miljoner besök under år 2020 och 2,5 miljoner under 2018 (Bäckerud, Nilsson & Sköldberg 2020:93). Antalet besök på Svenska Akademiens tre ordböckers gemensamma portal fortsätter således att stiga.

Motsvarande siffror för webbsidan saob.se visar en motsatt trend. Antalet besök ökade visserligen från 1,8 miljoner 2018 (Bäckerud, Nilsson & Sköldberg 2020:93) till 2,7 miljoner 2020, men under 2021 har siffran istället sjunkit något, till 2,3 miljoner. Förklaringen är troligtvis att fler användare väljer den gemensamma ordboksportalen istället, och/eller att sökmotorer som Google har börjat länka till denna i ökande omfattning. Vecka 50 år 2020

kom 29 % av läsarna till ordboksportalen svenska.se via en sökning på en sökmotor, och samma vecka år 2021 hela 59 %.

Fortfarande kommer en ännu större andel av besöken som görs på SAOB:s egen webbsida, saob.se, via sökmotorer. Här var siffrorna i det närmaste oförändrade under de aktuella veckorna och åren, nämligen cirka 88 % vecka 50 år 2020 och 85 % samma vecka år 2021. Ett typiskt scenario är nog, som nämnts ovan, att man googlar på ett okänt ord och får träff som leder till SAOB. Det är endast strax under 12 % av besöken som kommer direkt via webbadressen. Webbsidan saob.se har totalt cirka 6 000 besök per dag. Det är samma antal som år 2018.

Vi ser också att den största andelen av alla besök på saob.se görs från mobila enheter och inte från datorer. Detta ställer naturligtvis särskilda krav på gränssnitt och funktionalitet. Svenska Akademiens webbsidor är anpassade för olika typer av skärmar, och det verkar ändamålsenligt att även i framtiden satsa på en flexibel lösning som fungerar för olika plattformar och skärmstorlekar.

3.2.1. Återkommande icke-träffar på saob.se

En analys av webbsidans icke-träffar kan ge viktig information inför SAOB:s revidering. Vi har gått igenom de 531 vanligaste söksträngarna från saob.se år 2021 som inte ledde till någon träff (vilket innefattar alla icke-träffar som förekom i loggfilen minst 50 gånger). Dessa söksträngar kan inordnas i följande kategorier:

1. stavningsvariant/felstavat
2. böjningsvariant
3. flerordsuttryck
4. förkortning
5. engelska (eller engelskinspirerad svenska)
6. tyska
7. annat språk
8. egennamn
9. lemmalucka

Icke-träffarna i kategori 1–4 ovan kan ge indikationer på hur SAOB:s innehåll och funktionalitet kan förbättras. Målet är att den som stavar ett ord fel eller söker på en böjd form ska få relevanta förslag på närliggande ord som kan leda fram till rätt artikel. Vi behöver också göra det lättare för användarna att hitta fram till de flerordsuttryck och förkortningar som faktiskt beskrivs i artiklarna.

Vad gäller kategori 1, ”stavningsvariant/felstavat”, kan icketräffarna dessutom hjälpa redaktionen att identifiera ordformer i SAOB som behöver uppdateras. Att 96 sökningar på ordsträngen *juste* och 84 på *schysst* inte leder till någon träff beror exempelvis inte på att användarna har stavat fel, utan på att SAOB endast anger den ålderdomliga uppslagsformen *just* för det aktuella ordet (som publicerades i ordboken år 1934). Det samma gäller för verbet *sjangsera*, där SAOB anger formen *changera* som den enda möjliga (år 1904). Ordbokens uppslagsformer behöver alltså uppdateras i dessa fall.

Kategori 5–7 utgörs av söksträngar på annat språk än svenska. Förvånansvärt många sökningar avser engelska eller tyska ord. Den tyska frasen *soll erben* har av någon anledning eftersökts i SAOB så mycket som 111 gånger under år 2021. Detta kan vi troligen inte göra så mycket åt. Inte heller kommer kategori 8, ”egennamn”, att läggas till i SAOB:s lemmalista – trots att många användare söker på sådana.

Sökningar på engelskinspirerad svenska kan inte avfärdas lika lätt. Ett exempel är *najs* (eng. *nice*) som har eftersökts i SAOB 87 gånger. Stavningen är försvenskad, och en sökning i moderna tidningstexter visar att ordet delvis också har inlemmats i det svenska böjningssystemet; komparativformen ”najsare” ger 25 träffar i Mediearkivet. Detta talar för att ordet (snart) bör tas upp i svenska ordböcker. Det finns dock ingen särskild anledning att prioritera sådana ord i en historisk ordbok som SAOB. När det gäller den högt prioriterade uppgiften att komplettera lemmalistan med nya

ord är det tveklöst kategori 9, ”lemmalucka”, som är av störst intresse.

På grund av SAOB:s långa utgivningstid är det många ord som saknas i verket, framför allt ord som kommit in i svenskan under 1900-talet (efter att det aktuella alfabetiska spannet hade färdigställts). Att analysera de återkommande icke-träffarna på saob.se kan vara ett sätt att prioritera bland nyordskandidaterna till ordbokens uppdaterade lemmalista. Det är rimligt att ord som många efterfrågar läggs till så snart som möjligt. Tabell 1 ger några exempel på efterfrågade ord, och de ämnesområden som främst berörs.

medicin/hälsa	*aids*, *autism*, *friskvård*
samhälle	*diskriminera*, *global*
datorvärlden	*app*, *dator*, *chatta*, *skanna*
mat	*avokado*, *falafel*, *sushi*
”svåra” ord	*arbitrage*, *ibidem*, *predator*
vardagsliv/vardagsspråk	*cool*, *kuk*, *sunkig*

Tabell 1: Nyordskandidater bland icke-träffar från saob.se fördelade på ämnesområden.

Det är dock viktigt att komma ihåg att endast ord som inte har en homograf i SAOB:s lemmalista kan hittas på detta sätt. Enligt Bäckerud, Nilsson & Sköldberg (2020:95) var den artikel som beskriver pronomenet *hen* en av de mest lästa på ordboksportalen svenska.se år 2018. Detta pronomen finns inte med i SAOB, men strängen *hen* ingår ändå inte bland icke-träffarna eftersom en sökning på denna ordkombination ger träff på det ganska okända substantivet *hen*, som betyder ’brynsten’. På samma sätt har moderna ord som *dagis* och *fika* homografer i SAOB som gör att dessa söksträngar inte dyker upp bland icke-träffarna.

4. Mot ökad tydlighet och överskådlighet

Med en reviderad version av SAOB är förhoppningen alltså att nå ut längre och bredare till användarna, och förutsättningarna för ordboken att bli läst mer är goda mot den bakgrund som nämnts ovan, nämligen att ordboken finns tillgänglig på portalen svenska.se och att dess innehåll är välindexerat av Google.

Det informationstäta och svåröverskådliga användargränssnittet riskerar dock att, så att säga, skrämma bort besökaren i dörren. På svenska.se sticker SAOB:s grafiska framställning ut i förhållande till systerordböckerna, inte minst på grund av det i hög grad förkortade och formelartade metaspråket som är svårt att avkoda och förstå.

Ett ytterligare problem hänger ihop med användarnas förmåga att orientera sig på ordbokens webbsida och deras tålamod vid påträffade hinder. För trots att där redan nu finns instruktioner och hjälpavsnitt så är det många användare som inte läser eller hittar dem. Av de frågor och synpunkter från användarna som redaktionen tar emot blir det tydligt att man ofta efterfrågar information som redan finns på ordbokens hemsida, men som man inte har lyckats hitta eller har haft tid att läsa. Steget verkar i många fall kortare att skicka en fråga till redaktionen än att sätta sig in i det aktuella problemet och läsa på, och detta gör det uppenbart att framställningen i ordboken i några avseenden måste bli betydligt mer direkt och tydlig, framförallt i fråga om information som är relevant för icke-avancerade användare.

I det följande redogörs för ett antal förändringar som redaktionen ser framför sig som har att göra med ökad tydlighet och tillgänglighet för användaren. Gemensamt för de områden som diskuteras här är att de avser förändringar där det innehåll som redan finns i artiklarna utgör utgångspunkten. Det handlar i dessa fall inte i första hand om att lägga till ny information.

4.1. Förkortningar

Som redan påpekats är SAOB:s metaspråk förkortat i stor utsträckning. Redan nu erbjuds läsaren uttydning av samtliga förkortningar som används i ordboken i en lista som på saob.se finns tillgänglig i menyn hjälp>förkortningar. Men att döma av de frågor och synpunkter som kommer in till redaktionen når denna information ofta inte fram till användarna.

Idealt sett kunde förkortningar i en reviderad version av ordboken helt undvikas genom att de aktuella orden i samtliga fall skrevs ut i sin fulla längd: *intr.* > *intransitiv(t)*, *dep.* > *deponens*, *vbalsbst.* > *verbalsubstantiv* osv. I viss utsträckning är det också vägen framåt för SAOB-redaktionen: i en digital ordbok är utrymmesbrist inte ett lika stort problem som i den första upplagans pappersversion, och en hel del förkortningar kan rimligen upplösas. Men även bildskärmar har förstås en viss utrymmesbegränsning, inte minst i fråga om visningslägen hos mobiltelefoner och surfplattor (med vilka majoriteten av besök på saob.se alltså sker; jfr avsnitt 3.2 ovan). Att ersätta varje förkortning i SAOB med ett ord i sin fulla längd är därför inte lämpligt. Istället kommer fokus att riktas mot att förtydliga förkortningar av vardagliga ord som inte (längre) är etablerade och frekventa i allmänspråket: *sht* bör t.ex. hellre skrivas: *synnerhet*, *ss.* > *såsom* och den i ordboken mycket frekventa förkortningen *l.* för *eller* bör ersättas med det mer gängse och lättolkade *el.* För de förkortningar som bevaras bör vägen till uttydning kortas ordentligt: antingen genom att varje förkortning länkas direkt till sin uttydning, med mouse overfunktion eller motsvarande, eller genom att en länk till förkortningslistan placeras på en tydligare position inne i varje artikel.

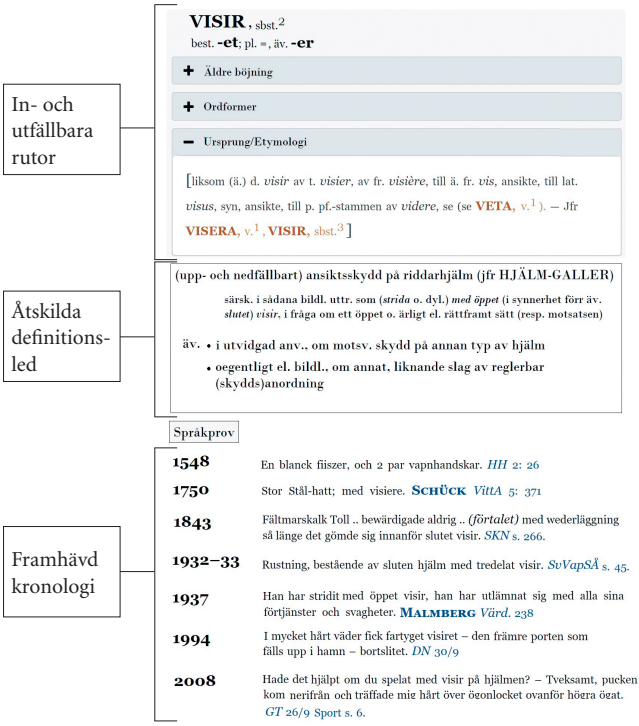
4.2 In- och utfällbara rutor, separerade definitionsled och språkprov

Figur 1 nedan utgör en grov skiss över ett möjligt presentationssätt av artiklar i en reviderad version av SAOB. Skissen tar fasta på och illustrerar ett par principer gällande förtydligad framställning av den information som redan nu finns i ordbokens artiklar. För det första utnyttjas möjligheten att stoppa undan detaljerad information om böjning och olika former i in- och utfällbara rutor i inledningen av artikeln i högre utsträckning än i dagens version av ordboken.[1] Primärt tänker sig redaktionen en lösning där äldre och inte längre brukliga böjningsmönster och -former döljs för att inte stjäla uppmärksamhet från annat. Informationen i rutorna kommer förstås fortfarande att vara sökbar och möjlig att klicka fram för den användare som är ute efter den.

För det andra har definitionstextens olika led separerats i presentationen. Ofta är det betydelsen man är ute efter när man slår upp ett ord i ordboken, och det är därför rimligt att definitionen får en mer framträdande plats i artikeln. SAOB:s definitioner utgör strukturerade texter som normalt inleds med en huvuddefinition följd av ett eller flera under- och sidoordnade definitionsled, vilka ofta markeras med *särsk.* (särskilt) respektive *äv.* (även). Genom att dessa element identifieras inom varje definition och skiljs åt i presentationen kan strukturen tydliggöras för läsaren. I figuren har underordning markerats med indrag och sidoordning med ny rad.

Slutligen har de språkprov som belägger de olika betydelser och användningar som redovisas i definitionen separerats från definitionstexten och försetts med en förtydligande tidslinje i marginalen.

1 I den digitala versionen av SAOB:s första upplaga kan informationskategorierna *ordformer* och *etymologi* och, i längre artiklar, *översikt* visas eller döljas på motsvarande vis med in- och utfällbara rutor.



Figur 1: Idéskiss över nytt presentationssätt i SAOB.

4.3 Artiklar i sammansättningsramsa

Ett betydande överskådlighetsproblem i SAOB gäller beskrivningen av sammansatta ord, och de utgör en väldigt stor del av ordbokens totala innehåll. Problemet här hänger samman med SAOB:s nästalfabetiska struktur, vilken innebär att sammansättningar presenteras ihop i en s.k. ramsa, underordnade simplexordet.[2] Sålunda presenteras exempelvis sammansättningarna *vagn-arbete*, *vagn-knekt* och *vagn-skada* tillsammans inom ramen för artikeln VAGN.

2 Termen *nästalfabetisk* är hämtad från Svensén (2004:441).

För en tryckt ordboks del är det här presentationssättet praktiskt eftersom det är utrymmesbesparande, och redaktören behöver inte upprepa information som gäller relationen till simplexordet vid varje enskild sammansättning. Istället samlas upplysningar som gäller samtliga sammansättningar på ett enda ställe i den överordnade artikeln, och endast efterleden återfinns på alfabetisk plats i ramsan. Men ur läsarens perspektiv är det inte optimalt att behöva söka information om det ord man slagit upp på andra ställen i ordboken. Därför kommer sammansättningsartiklarna i SAOB att göras mer lika huvuduppslagsorden. Strukturen kommer inte att omstöpas helt och hållet, men relevant information om varje sammansättnings relation till förleden, som i dagens upplaga alltså finns på annan plats i ordboken, kommer att flyttas in i varje sammansättningsartikel. Även relationen till efterleden kommer att lyftas fram tydligare och böjningsuppgifter och etymologiparentes kan liksom hos simplexord placeras i in- och utfällbara rutor (jfr avsnitt 4.2 ovan).

I figur 2 och 3 jämförs första upplagans presentation av sammansättningen *vagn-knekt* med en idéskiss över motsvarande artikel i en reviderad upplaga:

-KISTA, -KLASS, se **D.** —

-KNEKT. **(vagn- 1592. vagne- 1593)** [jfr t. *wagenknecht*] (†) person med uppgift att sköta vagn l. betjäna person(er) som färdas med vagn; jfr **knekt 1.** Jacker ått 5 wang kneckter. *KlädkamRSthm* **1592** B, s. 75 a. *KlädkamRSthm* **1593** D, s. 130 a. —

-KOLONN. kolonn (se d. o. **2** *a*) av vagnar (tillhörande militärförband). Så mycket gynnade oss .. lyckan, att vi vid trossen fingo slåss lika tappert som de andra .. Slutligen körde vi bort packet, hvarefter vagnkolonnen satte sig i rörelse. **Sparre** *Findl.* 3: 177 **(1835)**.

Figur 2: Sammansättningen VAGN-KNEKT i SAOB, 1 uppl.

Figur 3: Idéskiss över nytt presentationssätt av sammansättningen VAGN-KNEKT i en reviderad uppl. av SAOB.

5. Nya funktioner och användningsområden

Redaktionen ser även framför sig möjligheten att utöka med funktioner och användningsområden som tidigare helt saknades i SAOB, och det här hänger ihop med att ordboken kommer att övergå från att vara en digital version av en tryckt ordbok till en helt digital databas.

Ur ett mer traditionellt användarperspektiv önskas nya avancerade sökmöjligheter i ordboken. Det ska bli möjligt att söka efter vissa typer av ord (t.ex. från vissa ordklasser, med viss bruklighet eller från ett särskilt ämnesområde).[3] Vidare efterfrågas olika visningslägen för mer avancerade användare, och möjligheten att

3 Det ska påpekas att viss sådan funktionalitet har funnits redan tidigare, i OSA-versionen av ordboken. Men eftersom den versionen innehåller så många maskininläsningsfel är resultaten från sökningar där inte helt tillförlitliga.

spara personliga inställningar och sökningar. Som nämnts tidigare eftersträvas också ”smartare” sökfunktioner – med ett innehåll som taggats i större utsträckning bör sökningar på flerordskombinationer, förkortningar och böjda ord i högre grad leda till adekvata träffar.

Om man lyfter blicken lite, och tänker sig innehållet i den lexikografiska databasen SAOB i andra kontexter än sådana där erfarna användare slår upp ett ord i ordboken, ser redaktionen några olika möjligheter framför sig.

Ett ganska kort steg att ta är att med utgångspunkt i SAOB utveckla material att använda i skolor och på universitetsutbildningar. Som nämnts ovan används portalen svenska.se redan nu flitigt i skolan, och om man vill att SAOB ska komma ännu mer till sin rätt i den kontexten vore detta en naturlig väg framåt, som hittills inte har utnyttjats i någon vidare utsträckning. Undervisningsmaterial om ordbildning, betydelseutveckling och främmande språks påverkan på ordförrådet ligger då nära till hands.

Det faktum att SAOB utgör en del av svenska.se och alltså står sida vid med två andra ordböcker har även fått redaktionen att börja fundera över i vilken utsträckning de olika verken kan harmoniseras och/eller profileras. Framförallt har resonemangen hittills handlat om i vilken utsträckning SAOB:s grammatiska beskrivningar kan harmoniseras med verket *Svenska Akademiens grammatik* (SAG), som också finns tillgänglig i portalen.

Men sedan redaktionen för Svenska Akademiens (renodlade) samtidsordböcker påbörjat ett harmoniseringsarbete kring hur orden i SAOL och SO presenteras (jfr Blensenius 2022, Sköldberg 2022) vore förstås en möjlig vidareutveckling att även involvera SAOB i projektet. Det ska sägas att de olika ordböckernas presentationssätt grundar sig i helt olika principer. För SAOL och SO är utgångspunkten den s.k. *lemma-lexem-modellen* (se t.ex. Allén 1999) och för SAOB:s del är ordets historiska betydelseutveckling avgörande. Meningen är förstås inte att sudda ut SAOB:s specifika

profil i förhållande till sina systerordböcker, men att åtgärda eller åtminstone identifiera principiellt omotiverade skillnader verken emellan vore önskvärt.

Ökad harmonisering kan i ett nästa steg underlätta för användning av de lexikografiska databaserna inom *natural language processing* (NLP). I sitt experiment med DDO och ODS visade Ahmadi et. al. (2021) att s.k. *word sense alignment*-teknik (WSA), dvs. automatisk matchning av betydelser mellan identiska ord, går att tillämpa på lexikografiskt material som utgörs av digitaliserade tryckta ordböcker med sinsemellan stora skillnader och olika principer för presentationsordning. Och med tanke på den iakttagelse som Trap-Jensen (2018:34) gör vore det dumt att inte öppna dörren för ett motsvarande användningsområde för samtliga av Svenska Akademiens ordböcker:

> [W]e are witnessing an increasing need for high-quality lexicographical data. Language technology and artificial intelligence are moving into a phase where the word lists and morphological lexicons developed inside the NLP environment itself are insufficient to meet the demands for developing smarter and more sophisticated products. […] By far the best existing semantic descriptions of language are dictionaries, and for that reason, it is obvious that existing dictionaries are interesting for developers of such applications.

6. Avslutning

Den första upplagan av SAOB är en ordbok i en gammal kostym. Både språket och språkvetenskapen har i stor utsträckning förändrats under ordbokens utgivningsperiod, och användningskontexten ser idag helt annorlunda ut än i slutet av 1800-talet då

det första bandet av den tryckta ordboken gavs ut. Ett stort behov av revidering föreligger därmed.

Även fortsättningsvis är målet att SAOB ska vara en utförlig vetenskaplig ordbok, men den bör samtidigt vara så läsbar och tillgänglig som möjligt också för icke-avancerade användare. Planeringen av revideringsarbetet är nu i full gång. Tack vare indirekt och direkt feedback från användarna och genom interna analyser har den redan kommit ganska långt. Innehållet i ordboken behöver korrigeras, kompletteras och presenteras tydligare. Det måste vidare bli sökbart i större utsträckning – och kanske kan det på sikt även ligga till grund för andra än rent traditionella lexikografiska ändamål. Kort sagt, redaktionen blickar framåt mot nya tider och nya möjligheter.

Litteratur

Ordböcker

DDO = *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab og Gyldendal. <ordnet.dk/ddo> (maj 2022).

ODS = *Ordbog over det Danske Sprog.* Grundlagt af Verner Dahlerup, udg. af Det Danske Sprog- og Litteraturselskab, band 1–28, København 1918–1956. <ordnet.dk/ods> (maj 2022).

OSA = *Om svar anhålles*, äldre digital version av SAOB (band 1–35). <www2.saob.se/osa/index.phtml> (maj 2022).

SAOB = *Ordbok över svenska språket* utgiven av Svenska Akademien 1893–. Lund. <saob.se> och <svenska.se> (maj 2022).

SAOL 14 (2015) = *Svenska Akademiens ordlista.* 14 uppl. Stockholm: Norstedts. <svenska.se> (maj 2022).

SO (2021) = *Svensk ordbok utgiven av Svenska Akademien.* 2 uppl. <svenska.se> (maj 2022).

Annan litteratur

Abel, Andrea & Christian M. Meyer (2013): The dynamics outside the paper: user contributions to online dictionaries. I: Iztok Kosem et al. (eds.): *Proceedings of the eLex 2013 conference: Electronic lexicography in the 21st century: thinking outside the paper*. 17–19 October 2013. Tallinn: Institute of the Estonian Language/Trojina, Institute for Applied Slovene Studies, 179–194.

Ahmadi, Sina, Sanni Nimb, John P. McCrae & Nicolai H. Sørensen (2021): Towards Automatic Linking of Lexicographic Data: the case of a Historical and a Modern Danish Dictionary. I: Zoe Gavriilidou, Maria Mitsiaki & Asimakis Fliatouras (eds.): *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion, Vol I*. 17-19 September 2021. Alexandroupolis: Democritus University of Thrace, 63–72.

Allén, Sture. (1999) [1981]: The Lemma-Lexeme Model of the Swedish Lexical Database. I: *Modersmålet i Fäderneslandet. Ett urval uppsatser under fyrtio år av Sture Allén*. (Meijerbergs arkiv för svensk ordforskning 25.) Göteborg: Göteborgs universitet, 268–278.

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

Blensenius, Kristian (2022): Mot en harmonisk lemma-lexem-modell. Föredrag vid ”Den 16:e konferensen om lexikografi i Norden”. Lund 2022.

Bäckerud, Erik, Pär Nilsson & Emma Sköldberg (2020): Så används Svenska Akademiens ordböcker på nätet. Implicit och explicit feedback från användarna. I: *Nordiska studier i lexikografi* 15. Helsingfors: Nordisk förening för lexikografi, 91–101.

Larsson, Lennart (2020): Total bruklighetsinskränkning i SAOB – nu och i framtiden. I: *Nordiska studier i lexikografi* 15. Helsingfors: Nordisk förening för lexikografi, 187–194.

Mediearkivet <app.retriever-info.com> (maj 2022).

Rosqvist, Bodil & Bo A. Wendt (2020): Inför den framtida revideringen av SAOB. I: Daniel Sävborg, Eva Liina Asu & Anu Laanemets (red.): *Studier i svensk språkhistoria* 15. *Språkmöte och språkhistoria.* Tartu 2020, 242–253.

Sköldberg, Emma (2022): ”Varför står det olika i SAOL och i SO?” Om bearbetning av omotiverade skillnader mellan Svenska Akademiens samtidsordböcker. Föredrag vid ”Den 16:e konferensen om lexikografi i Norden”. Lund 2022.

Svensén, Bo (2004): *Handbok i lexikografi*. Stockholm: Norstedts Akademiska Förlag.

Tarp, Sven (2008): *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography.* (Lexicographica Series Maior 34.) Tübingen: Max Niemeyer Verlag.

Trap-Jensen, Lars (2018): Lexicography between NLP and Linguistics: Aspects of Theory and Practice. I: Jaka Čibej, Vojko Gorjanc, Iztok Kosem & Simon Krek (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. 17–21 July 2018. Ljubljana: Ljubljana University Press, Faculty of Arts, 25–37.

Pär Nilsson
Ordboksredaktör, fil. lic.
Svenska Akademiens ordboksredaktion
Dalbyvägen 3
SE-224 60 Lund
par.nilsson@svenskaakademien.se

Bodil Rosqvist
Bitr. huvudredaktör, fil. dr.
Svenska Akademiens ordboksredaktion
Dalbyvägen 3
SE-224 60 Lund
bodil.rosqvist@svenskaakademien.se

COR-S – den semantiske del af Det Centrale OrdRegister (COR)

Sanni Nimb, Bolette S. Pedersen, Nathalie Carmen Hau Sørensen, Ida Flörke, Sussi Olsen & Thomas Troelsgård

We present the formal lexicon *COR-S*, which constitutes the semantic part of a Danish computational lexicon project called *COR*. *COR-S* is based on linked data, but apart from transferring, adjusting, and validating the information from existing Danish lexicons and dictionaries, the goal is also to compile an AI suitable sense granularity level. Based on the fine-grained sense inventory of *Den Danske Ordbog* (DDO), senses are clustered, partly by hand according to a set of principles, and partly by means of automatic NLP methods.

1. Baggrund og introduktion til COR

COR står for Det Centrale OrdRegister for dansk og er et nystartet sprogteknologisk ordbogsprojekt der har til formål at etablere en leksikalsk ressource som er egnet til anvendelse i kunstig intelligens og andre teknologiske applikationer der arbejder med dansk sprog. Projektet blev igangsat i 2021 i et samarbejde imellem på den ene side to af de ledende udviklere af danske ordbøger, Dansk Sprognævn og Det Danske Sprog- og Litteraturselskab (DSL), og på den anden side en sprogteknologisk partner, Center for Sprogteknologi (CST) ved Københavns Universitet. Det løber frem til udgangen af 2023 og udgør en del af den sprogteknologisatsning der er blevet igangsat i forbindelse med den danske regerings AI-strategi fra 2019, og hvor det påpeges at der er et behov for at styrke indsatsen omkring frit tilgængelige danske sprogressourcer.

En af grundtankerne i COR er at skabe analogi mellem Det Centrale PersonRegister (CPR) og Det Centrale OrdRegister. Alle

oplysninger om danske ord kobles således til samme stabile og fremtidssikrede register, både de semantiske oplysninger der allerede nu udvikles i COR-S og fremtidige leksikalske data. Kongstanken er at man ved at stille et indekseret system med unikke og stabile id-numre på alle ordformer til rådighed kan sikre en mere effektiv deling af danske leksikalske ressourcer. Frit tilgængelige oplysninger om stavning, bøjning og formelle betydningsbeskrivelser koblet til en central del af ordene vil gøre det langt nemmere for danske offentlige institutioner og virksomheder at arbejde med dansk sprogforståelse (jf. AI-strategiens målsætning). COR giver fx mulighed for at forbedre sprogcentrerede AI-systemer der oprindeligt er udviklet til engelsk, idet de leksikalske beskrivelser tager afsæt i anerkendt, lokalt forankret viden om dansk sprog og kultur som matcher det samfund som systemerne skal fungere i.

I COR-projektet indarbejdes og forenkles ordbogsmateriale og sprogressourcer der tidligere er udviklet ved de tre samarbejdende institutioner, som alle oplever en stigende efterspørgsel på ordbogsdata med basale oplysninger om udtale, bøjning og betydning beskrevet på en standardiseret og kompatibel måde. Offentlige institutioner og virksomheder der i dag arbejder med danske, digitale sprogdata, finder det udfordrende og tidskrævende at afklare hvad der findes, hvor det findes, og hvor tilgængeligt det i praksis er, blandt andet pga. mangel på sproglig ekspertise i virksomheden og pga. ressourcernes ofte fragmenterede natur. Ofte rejses spørgsmålet om hvordan de spiller sammen med virksomhedens egen fagterminologi, og i praksis fører det til at man i mange tilfælde udelukkende arbejder tekstbaseret og undlader at inddrage vigtig viden fra leksikalske ressourcer i sine sprogteknologiske komponenter. Men da de tekstmængder der er til rådighed, ofte er for sparsomme til at opnå gode resultater, giver det rigtig god mening også at inddrage leksikalsk viden, særligt hvis den er tilgængelig i en overskuelig og formaliseret form.

Artiklens fokus er den semantiske del af COR-projektet, COR-S, der lanceres sidst i 2023. I næste afsnit beskriver vi eksisterende bagvedliggende sprogressourcer der udnyttes i arbejdet med at udvikle COR-S, og vi giver eksempler på indholdet af en ordbogsindgang baseret på overførsel af data fra disse ressourcer. I afsnit 3 beskriver vi den leksikografiske fremgangsmåde i projektet, og i afsnit 4 de metoder der udvikles med henblik på at automatisere en del af arbejdet med at opnå et betydningsinventar for polyseme lemmaer der er håndterbart i sprogteknologi.

2. Bagvedliggende ordbogsressourcer

I COR-S samles en lang række formaliserede semantiske oplysninger om danske lemmaer der i dag kun er offentligt tilgængelige i adskilte ressourcer. Oplysningerne er gennem mange år opbygget i et tæt samarbejde mellem DSL og CST om at udvikle semantiske ressourcer til anvendelse i forskning i automatisk sprogforståelse. Ressourcerne er baseret på internationale standarder, og samarbejdet blev indledt med udviklingen af det danske WordNet *DanNet* i 2004 (Pedersen 2009 et al.), kort efter at den trykte udgave af *Den Danske Ordbog* (Hjorth & Kristensen 2003-2005, herefter DDO) var færdiggjort. Hver ny ressource har undervejs i forløbet muliggjort den næste og bidraget med afgørende ny information (Pedersen et al. 2018a, Pedersen, Nimb & Olsen 2021). Også *Den Danske Begrebsordbog* (herefter Begrebsordbogen), udgivet af DSL i 2015 (Nimb et al. 2015), er udviklet efter dette princip. Begrebsordbogen bygger videre på informationer i både DDO og DanNet, og efterfølgende blev dens indhold og emnestruktur kombineret med oplysninger i DDO og udnyttet til at udvikle to andre semantiske ressourcer. Den første var et FrameNet-leksikon der angiver en eller flere semantiske rammer for størstedelen af DDO's verber (se Nimb et al. 2017, Nimb 2018), og nogle år efter fulgte et senti-

mentleksikon der angiver om et lemma har positiv eller negativ konnotation (Nimb et al. 2022, Pedersen, Nimb & Olsen 2021).

Fælles for ressourcerne er at de alle er koblet til DDO's unikke og stabile betydningsnumre og dermed også til selve DDO-lemmaet. Det er det der nu udnyttes til fulde i COR-projektet, idet DDO's lemmaer kobles til *Retskrivningsordbogens* lemmaer der udgør basis for COR-registret. Den trykte DDO er siden 2005 videreudviklet som onlineordbog, og de data der er tilføjet siden første udgave, indgår også i ressourcerne. Med udgangspunkt i DDO-betydningsnumrene kan vi i dag kombinere alle typer af data på kryds og tværs: DDO-artiklens mange oplysninger vedrørende betydningen, DanNets oplysninger om ontologisk type og semantiske relationer, FrameNet-leksikonets semantiske rammer, oplysninger vedrørende positiv eller negativ konnotation fra sentimentleksikonet og endelig oplysninger om beslægtede ord, nøgleord og emne fra Begrebsordbogen.

Fælles for ressourcerne er også at de, modsat DDO, alle er opbygget ud fra en ontologisk tilgang som traditionelt anvendes i arbejdet med at opbygge formelle semantiske leksika (se Geeraerts 2002, Pustejovsky 1995). DanNet strukturerer således betydninger af DDO-lemmaer i en række semantiske relationer, primært i over- og underbegrebsrelationer. Ud fra betydningernes placering i træstrukturen kan man fx udlede automatisk om de er tæt på hinanden eller ej. WordNets er udviklet for en lang række sprog i verden ud fra samme format og standard som oprindelig blev etableret ved Princeton University (Fellbaum 1998), se hjemmesiden for Global WordNet Association. WordNets udgør grundlæggende ordbogsressourcer i international forskning inden for udviklingen af metoder til automatisk sprogforståelse. Det samme gør sig gældende for FrameNets, der findes for engelsk, svensk og en lang række andre sprog. De baserer sig på en standard der er udarbejdet (og stadig videreudvikles) ved University of California, Berkeley, USA (Ruppenhofer et al. 2016), se også FrameNets hjem-

meside. I projektet fastlægges og navngives en lang række semantiske rammer og deres tilhørende semantiske roller for engelsk. Korpustekster opmærkes med disse oplysninger så man dermed efterhånden får etableret et leksikon med de lemmaer der "udløser" en ramme. For dansk blev leksikonet i stedet udarbejdet ud fra en opmærkning af semantiske grupper af verber og verbalsubstantiver i Begrebsordbogen idet valensmønstre fra DDO samtidig blev koblet til de enkelte forekomster.

Den ontologiske tilgang i udviklingen af både DanNet og det danske FrameNet-leksikon har som konsekvens at ikke alle betydninger af et givet DDO-lemma nødvendigvis er repræsenteret i dem. Fokusset har ikke ligget på polysemi og betydningsinventar, men i stedet på taksonomier, begreber og relationer mellem betydninger på tværs af lemmaer. Da COR-projektets grundidé er at koble sproglige ressourcer sammen via indekserede lemmalister, må vi nødvendigvis vende tilbage til det semasiologiske udgangspunkt fra DDO og forholde os til alle DDO-betydninger. En formel ressource med fuld betydningsdækning for lemmaerne vil samtidig give nye muligheder for forskning i automatisk entydiggørelse af ordbetydninger, en af de helt store udfordringer i automatisk analyse af sprog. Vi ved fra tidligere forskningsprojekter at DDO's betydningsinventar er meget finkornet og ikke umiddelbart velegnet til formålet (Pedersen et al. 2016, Pedersen 2018, Pedersen et al. 2018b). Automatiske metoder er afhængige af at et lemmas forskellige betydninger udviser distributionelle forskelle i korpora. Betydningerne må i brug omgive sig med temmelig forskellige naboord for at kunne skelnes automatisk fra hinanden. En af de væsentlige opgaver i projektet er at opbygge et betydningsinventar med dette for øje, uden dog at gå på kompromis med de betydningsskel der tydeligt opfattes af mennesker. COR-S bliver med andre ord en forenklet udgave af DDO, hvor kun de væsentligste betydninger af lemmaer er repræsenteret, vel at mærke udtrykt ved hjælp af værdier inden for et begrænset formelt inventar (on-

tologisk type, semantisk ramme) samt et overbegreb i form af en præcis ordbetydning fra COR-S, ikke blot en tekststreng. Samtidig bliver der givet forklarende oplysninger til den menneskelige læser og it-bruger af COR-S. I tabel 1 ses et eksempel på en COR-S-ordbogsindgang med formaliserede betydningsoplysninger der kan trækkes direkte ud fra eksisterende data i DanNet, FrameNet-leksikonet og Begrebsordbogen.

Verb *bemærke*	**Betydning 1**	**Betydning 2**
Ontologisk type	Act+Mental	Act+Communication
Overbegreb	opfatte_COR_1	ytre_COR_1
Semantisk ramme	Becoming_aware	Mention
Definition	blive opmærksom på; lægge mærke til	gøre opmærksom på; nævne
Stikord	få øje på, observere	nævne, omtale
Eksempel (fra DDO)	*Flere naboer bemærkede en kraftig banken på et stuevindue …*	*Ja, fru Nielsen er flink, bemærkede Linda adspredt*

Tabel 1: Et eksempel på et verbum i COR-S hvor de formelle oplysninger om betydning stammer fra eksisterende sprogteknologiske ordbøger der er koblet til DDO.

Den ontologi der anvendes i COR-S, er i udgangspunktet magen til DanNets (og EuroWordNets), se Pedersen et al. (2009) og Vossen (1999). Den er dog forenklet, ikke kun hvad angår selve betegnelserne (’3rdOrderEntity’ hedder fx i stedet ’Abstract’), men også hvad angår omfanget af typer. Mange typer i DanNet er sammensat af adskillige betydningselementer (fx både ’Purpose’ og ’Social’ i typen ’Dynamic+Agentive+Purpose+Social’). I COR-S er de forenklet, baseret på hvor ofte de sammensatte typer reelt er anvendt i DanNet, og leksikografen må i stedet vurdere hvilket betydningselement der er vigtigst. Både *teselskab* og *hverv* har fx ovennævnte type i DanNet, men i COR-S har *teselskab* fået tildelt

typen 'Act+Social', mens *hverv* har fået tildelt 'Act+Purpose'. Et skel i DanNet-ontologien mellem uafsluttet og afsluttet handling/hændelse er helt fjernet i COR-S; skellet var oprindeligt møntet på romanske sprog i EuroWord-ontologien, men det er sjældent leksikaliseret i dansk. I alt er antallet af ontologiske typer reduceret med 36 % fra 204 i DanNet til 130 i COR-S. Se eksempler i tabel 2.

DanNet	COR-S	Eks.
UnboundedEvent	Event	*ske*, *hænde*, *foregå*
BoundedEvent		
UnboundedEvent+Agentive	Act	*gøre*, *handle*, *handling*
BoundedEvent+Agentive		
Dynamic+Agentive		
3rdOrderEntity+Mental+Purpose	Abstract +Purpose	*formål*, *mål*
3rdOrderEntity+Mental+ Purpose+Manner		
BoundedEvent+Agentive+ Purpose+Possession	Act+ Possession	*overdrage*, *give*
BoundedEvent+Agentive+ Purpose+Possession+Social		

Tabel 2: Eksempler på ontologiske typer i DanNet og COR-S.

Overbegrebet i tabel 1 overføres fra DanNet og tilrettes semiautomatisk til det modsvarende COR-S-betydningsnummer når det ligger fast ved projektets afslutning. Semantisk ramme overføres direkte fra FrameNet-leksikonet der beskriver rammer for en eller flere betydninger af 5.300 verber og 6.490 verbalsubstantiver i DDO. I alt er der anvendt 671 forskellige værdier fra Berkeley FrameNet, og de kan slås op i en ordbog på projektets hjemmeside, hvor deres betydning og tilhørende semantiske roller beskrives. Definitionerne overføres fra DanNet (hvor de består af "klippede" definitioner fra DDO). De suppleres med stikord i form af ordet umiddelbart til venstre for DDO-betydningen i Begrebsordbogen

samt det nærmeste nøgleord til venstre. Stikordene udtrækkes med samme algoritme som anvendes i funktionen *Ord i nærheden* i DDO (Nimb, Sørensen & Troelsgård 2018), se figur 1.

Figur 1: Stikordene *få øje på* og *observere* for betydning 1 af verbet *bemærke*, taget fra *Ord i nærheden*, ordnet.dk/ddo.

I tilfælde af uanvendelige stikord fjernes de i redigeringsprocessen. Endelig overføres eksemplet i tabel 1 fra citatmaterialet i DDO.

2.1. Hvilket ordforråd?

Det er et krav at de "væsentligste" betydninger i dansk skal være repræsenteret i første version af COR-S. Faste udtryk medtages ikke, og vi fokuserer på de åbne ordklasser, primært substantiver, verber og adjektiver. Vores lemmaselektion bygger på viden om det danske ordforråd som vi har opnået i arbejdet med andre ordbøger og ressourcer. For det første medtager vi de danske lemmaer der via DanNet i mindst én betydning er udpeget som ækvivalent til et af de 5.000 centrale begreber i Princeton WordNet (Pedersen et al. 2019). Der er tale om 4.600 DDO-lemmaer som vi betegner CBC-lemmaer (forkortelse for 'Core/Base Concepts'; for udvælgelse af disse se Global WordNet Associations hjemmeside). Nogle eksempler er *abe*, *acceptere*, *adgang*, *ekspert*, *elegant*, *knække*, *spise*. Tre fjerdedele af CBC-lemmaerne er polyseme i DDO, og selvom de kun udgør ca. 3,5 % af ordbogens lemmaer, dækker de ca. 11 % af betydningerne i den, endda uden at medregne betydninger fra de mange faste udtryk som en del af lemmaerne indgår i. For det andet medtager vi alle DDO-lemmaer der i mindst én betydning

optræder som nøgleord i Begrebsordbogen, dvs. er fremhævet som indledende overskrift for en semantisk gruppe af nærsynonymer og/eller synonymer i et af de 888 navngivne tematiske afsnit. På den måde sikrer vi en bred dækning af emner i COR-S-ordforrådet, og vi sikrer at fremtidige brugere af COR-registret med stor sandsynlighed kan relatere deres egne ord til et COR-lemma, fx inden for et fagligt område. Omkring 11.500 lemmaer optræder som nøgleord i Begrebsordbogen, dog er ca. 3.100 af dem allerede udvalgt som CBC-lemma, men i alt opnår vi på denne måde ca. 13.000 "væsentlige" og ofte polyseme lemmaer. Første version af COR-S indeholder også mange af de øvrige polyseme lemmaer i DDO hvoraf mindst én betydning er med i DanNet. Vi har manuelt opmærket og sammenlagt 2.600 af disse på linje med CBC-ordene, mens vi påregner at behandle 5.000 automatisk, se afsnit 4. Desuden omfatter første version af COR-S alle monoseme DDO-ord der er i både DanNet og *Retskrivningsordbogen*, i alt 18.000 lemmaer. Disse kan forholdsvis nemt overføres fra DanNet og kan fungere som sikre forankringer for automatiske semantiske analyser. I alt vil første version af COR-S omfatte mindst 35.000 lemmaer.

3. Leksikografisk fremgangsmåde

De færreste polyseme lemmaer i DDO opfører sig som verbet *bemærke* ovenfor, hvor vi anser begge betydninger for at være væsentlige og berettiget til at blive repræsenteret i COR-S-ressourcen. Mange har i stedet en eller flere ikke-centrale betydninger eller betydninger der kan anses som en indsnævret (eller udvidet) variant af en hovedbetydning. I nogle tilfælde er en sådan variant beskrevet som en ny hovedbetydning for at undgå alt for dybe betydningshierarkier i DDO. En stor del af det manuelle leksikografiske arbejde i COR består derfor i at analysere de enkelte lemmaers

betydninger med henblik på en enklere repræsentation. Arbejdet foretages i flere trin og optimeres ved inddragelse af automatiske metoder som beskrives i afsnit 4.

Første trin er selve udvælgelsen af de leksikografiske informationer der er relevante ved analysen af betydningsinventaret for hvert lemma, se tabel 3. Vi har her inddraget oplysninger fra DDO (fx definition, brugs- og fagmarkeringer og valensmønster), Begrebsordbogen (fx naboord og ordets evt. status og frekvens som nøgleord), DanNet (overbegreb, ontologisk type) og FrameNet-leksikonet (semantisk ramme). Derudover er der beregnet et pointtal for hver betydnings "tyngde" ud fra hvor mange citater, kollokationer og andre oplysningstyper i det hele taget der er knyttet til betydningen i DDO. Der er seks leksikografer involveret i arbejdet, og der arbejdes i online regneark; nogle er lavet til de enkelte leksikografer, andre er fælles, men de har alle ens opsætning.

Verbet *blomstre*	**semant. ramme**	**nøgle-ord**	**point**	**hovedbet. = 1, underbet. = 2**
bet.1. 'have blomster der er sprunget ud'		0	56	1
bet. 1.a 'trives og udfolde sig, være el. komme i god udvikling'	Thriving	2	72	2
bet.1.b 'være sund og smuk'		0	17	2

Tabel 3: Eksempel på oplysninger for verbet *blomstre*. De to underbetydninger 1.a og 1.b. lægges sammen til én i COR-S.

Arbejdet med at udvælge og sammenlægge DDO-betydninger, herunder fravælge betydninger man mener er for perifere eller sjældne, foregår ud fra en række principper og udføres af både studentermedhjælpere og erfarne leksikografer. Fra tidligere projekter med manuel opmærkning af leksikalske data har vi gode

erfaringer med at udarbejde detaljerede regler for at sikre at alle annotører, både studentermedhjælpere og erfarne leksikografer, arbejder ud fra samme retningslinjer. I arbejdet med at forenkle betydningsstrukturen udledte vi på baggrund af opmærkning af 25 % af CBC-ordene en række principper der baserer sig på de tilgængelige leksikografiske oplysninger. Betydninger der er markeret som sjældne eller historiske i DDO, bør fx fravælges, det samme gælder betydninger med meget lavt pointtal i det datasæt der er opstillet. Indsnævrede og udvidede underbetydninger lægges sammen med deres hovedbetydning, hvorimod overførte betydninger bevares. Konkrete betydninger bevares så vidt muligt, også selv om de er indskrænkede eller udvidede underbetydninger. Vi er også meget opmærksomme på at sikre en række faste principper vedr. systematisk polysemi (se fx Pustejovsky 1995) så den samme type polysemi behandles ens for alle ord. Fx bibeholdes begge betydninger ved mønstret dyr vs. madvare (fx *kylling*, *kalv*), mens de lægges sammen ved bygning vs. institution (fx *skole*). Ved mønstret proces vs. resultat bevarer vi begge betydninger når resultatet er konkret (som i *byggeri*), men når resultatet er abstrakt, lægger vi derimod de to betydninger sammen (som i *udtalelse*). I alt har vi registreret og opstillet regler for 35 forskellige mønstre, og information om mønstret vil komme med i ressourcen.

Der er udarbejdet skemaer med beskrivelser af de enkelte ontologiske typer og eksempler, herunder også en grafisk illustration til at overskueliggøre ontologien. Ikke kun ontologiske typer og semantiske rammer, men også overbegreber betragtes i øvrigt som lukkede inventarer. En liste over alle overbegreber der er anvendt i DanNet sorteret efter frekvens sikrer at tildelingen af overbegreber til nye betydninger strømlines og holdes inden for lemmaer der i forvejen er udvalgt til COR-S.

Hvad angår oplysninger om ontologisk type og overbegreb, kompliceres genbrug af data fra DanNet når betydninger lægges sammen, og det leksikografiske arbejde består derfor også i at ju-

stere de overførte værdier så de passer på den sammenlagte betydning. Samme justering skal i øvrigt foretages for synonymer fra DDO og for de positive/negative konnotationsværdier fra sentimentleksikonet. Synonymer tilføjes evt. i stedet som et særskilt modul til COR-registret.

Derudover skal de polyseme lemmaer forsynes med oplysninger ved relevante betydninger der ikke i forvejen er med i DanNet. Endelig skal eksisterende DanNet-oplysninger om ontologisk type og overbegreb ved de monoseme lemmaer valideres og tilpasses den nye, forenklede COR-S-ontologi.

Datasættet med polyseme CBC-ord spiller en central rolle i COR-S. De manuelle sammenlægninger der annoteres i datasættet, danner nemlig også udgangspunkt for de automatiske metoder til sammenlægning af betydninger ved resten af det polyseme ordforråd (se afsnit 5). Vi validerer derfor hinandens arbejde, dels ved at notere svære tilfælde der efterfølgende tjekkes af endnu en leksikograf, dels ved systematisk at validere 2 % af alle lemmaer med 6 eller færre DDO-betydninger i datasættet. Annotørenigheden er på 88 %, hvilket regnes for højt når der er tale om semantiske opmærkninger. Desuden er alle lemmaer med mere end 6 betydninger i DDO gennemgået af mindst to leksikografer, se fx verbet *støtte* i tabel 4, hvor man konfererede om sammenlægningerne. Resultatet af arbejdet er en meget klar reduktion af den finkornede betydningsinddeling i DDO. Antallet af betydninger er reduceret med 43 %, fra et gennemsnit på 4,3 betydninger pr. CBC-lemma i DDO til 2,4 betydninger pr. lemma i COR-S.

DDO-bet.	***støtte*, vb.**	COR-bet.
1	'yde moralsk, økonomisk eller anden hjælp og bistand'	1
1.a	'give sin tilslutning til; bakke op; gå ind for'	1

1.b	'underbygge yderligere; gøre mere troværdig eller overbevisende'	2
2	'bære eller holde noget oppe så det ikke falder ned eller vælter'	3
2.a	'hjælpe nogen med at holde sig oprejst, rejse sig, bevæge sig af sted el.lign. ved at lade vedkommende holde fast i eller hvile med sin vægt mod en'	1 (= 'yde hjælp') eller 3 (= 'holde/støtte fysisk')?
3	'hvile med sin vægt mod eller på noget; læne sig op ad'	4
3.a	'lade en legemsdel hvile mod eller på et fast underlag el.lign. så den holdes oppe'	4

Tabel 4: DDO-underbetydninger lægges ofte sammen med DDO-hovedbetydninger. Tvivlstilfælde diskuteres med de øvrige leksikografer, her 2.a af verbet *støtte*.

4. Automatiserede metoder

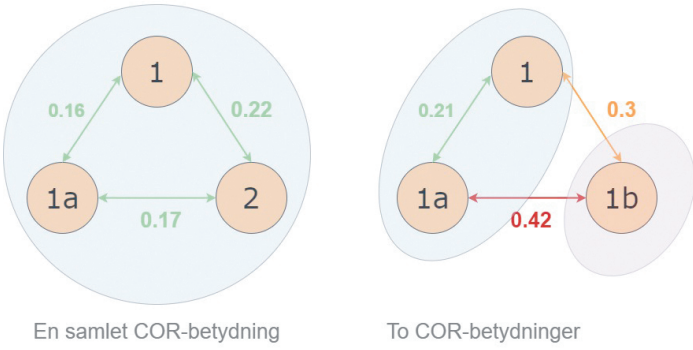
Selvom de leksikografiske principper skaber et godt grundlag for sammenlægning af betydninger, er den manuelle sammenlægning stadig en omstændelig og tidskrævende proces. Vi eksperimenterer derfor med at automatisere en del af processen, særligt for de knap så polyseme lemmaer (jf. Pedersen et al. 2022 for en fuld teknisk beskrivelse af disse eksperimenter). Til denne automatisering kan vi genbruge vores håndannotationer som både trænings- og valideringskorpus, med andre ord fungerer CBC-datasættet samt de øvrige håndopmærkede polyseme lemmaer som vores "guldstandard".

Vi undersøger tre tilgange til automatisk sammenlægning: en regelbaseret tilgang der er baseret på de leksikografiske principper og to statistiske tilgange med tekstdata fra DDO. Fælles for dem er at vi bruger den samme metode. Først beregner vi en afstandsscore

mellem alle kombinationer af et lemmas betydninger. Det er her tilgangene afviger mest fra hinanden, da vi bruger forskellige modeller til at bestemme semantisk nærhed. I de statistiske tilgange bruger vi trænede word embeddings der udregner en vektorrepræsentation ud fra distributionen af et lemmas forekomster i et træningskorpus[1]. Disse vektorrepræsentationer sammenlignes med henblik på at beregne en afstand imellem dem. I den regelbaserede tilgang får resultater der afspejler de udarbejdede principper, i stedet den bedste score. I andet trin benytter vi en algoritme til at bestemme hvilke betydninger der lægges sammen på baggrund af afstandsscoren. Når mere end to betydninger lægges sammen, sikrer vi dermed at de kun lægges sammen hvis afstanden mellem alle betydningerne er tilstrækkelig lille. Et eksempel på automatisk sammenlægning ses på figur 2 som viser to lemmaer der begge har tre betydninger som udgangspunkt.

Hvis alle afstandsscorerne er lave (til venstre i figur 2), kan alle betydningerne lægges sammen. Det er tilfældet for et lemma som *tøj* hvor definitioner for alle tre betydninger indeholder lemmaet *stof* og andre lemmaer relateret dertil (fx *beklædningsgenstande*, *klæde*). Lemmaet *sport* opfører sig derimod som eksemplet til højre på figuren hvor en betydning har mindst én høj afstandsscore til de andre. Betydning 1.b er nemlig i dette tilfælde overført, hvilket også kommer til udtryk i citatet (*for mange drenge er det en sport at skrive bilnumre op*). Citaterne for betydning 1 og 1.a omhandler dyrkning af sportsgrene, hvorimod citatet fra 1.b beskriver ordet *sport* om en sjov aktivitet eller leg. Betydning 1 og 1.a kan derfor lægges sammen, mens 1.b får sin egen betydning.

1 Word embeddings tager udgangspunkt i den distributionelle hypotese (Firth 1957, Harris 1954). Ifølge denne kan man udlede et ords semantik fra den omkringliggende kontekst. Dette kan afbildes statistisk som en vektor i et vektorrum via en word embedding-model, fx word2vec (Mikolov, Yih & Zweig 2013). Ord der ligger tæt på hinanden i vektorrummet, deler distributionel information og kan derfor antages at have betydninger der ligner hinanden.



Figur 2: Sammenlægning baseret på afstandsberegning.

4.1. Eksperiment 1: Regelbaseret tilgang

Den regelbaserede tilgang udregner semantisk nærhed efter de leksikografiske principper, jf. afsnit 3. To betydninger anses for at være semantisk tæt på hinanden hvis tre kriterier er opfyldt: (1) de hører under samme hovedbetydning, (2) ingen af betydningerne er overførte ifølge DDO, og (3) de har samme ontologiske type ifølge DanNet. Det sidste kriterium er betinget af at begge betydninger findes i DanNet.

Fordelen ved denne tilgang er at den er teoretisk og praktisk ligetil, dog med den ulempe at vi ikke har eksplicitte principper for hvornår betydninger på tværs af hovedbetydninger kan sammenlægges.

4.2. Eksperiment 2: word embeddings – statistiske ordprofiler

Traditionelle word embedding-modeller kan i udgangspunktet ikke adskille betydninger. Flertydige ord vil derfor få én samlet vektor der indeholder distributionel information om alle betyd-

ningerne. For at kunne udnytte de leksikalske informationer i COR-S har vi brug for at kunne splitte word2vec-vektorerne i betydninger baseret på DDO (jf. fx Olsen, Pedersen & Sayeed 2020). Konkret bruger vi en word2vec-model trænet af DSL på basis af DSL's korpus (Sørensen & Nimb 2018). For hver betydning vi har udtrukket fra DDO, beregner vi en kombineret word2vec-vektor ud fra definitioner og citater i DDO. I eksemplet ved figur 2 er det netop overlappet mellem definitioner (fx *tøj*) og citater (fx *sport*) der dannede baggrund for at beregne afstanden mellem betydninger. Afstanden måles ved hjælp af cosinus og læren om trekanters vinkler.

4.3. Eksperiment 3: kontekstualiserede embeddings

Kontekstualiserede embeddings er en nyere og mere kompleks metode som omgår problemet med flertydige repræsentationer ved at skabe en vektor for hver token (streng) i en sætning. Da kontekstualiserede embeddings repræsenterer et givent lemma i en specifik kontekst, kan vi antage at vektorerne her i højere grad afspejler betydninger.

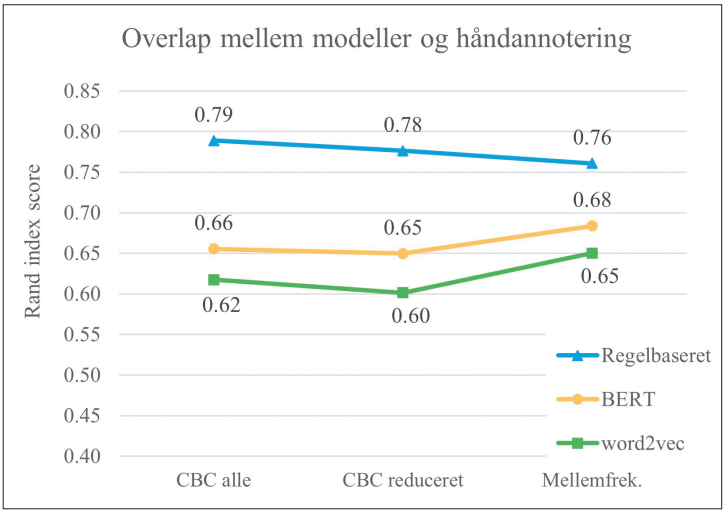
Vi bruger en såkaldt BERT-model (Devlin et al. 2019)[2], der er fortrænet af firmaet Certainly på et dansk tekstkorpus på cirka 1,6 milliarder ord fra Common Crawl, Danish OpenSubtitles, Danish Wikipedia og anden tekst fra internettet. Man kan yderligere finindstille modellen til et bestemt formål ved hjælp af en ekstra træningsopgave, og i vores tilfælde tilpasser vi modellen på vores håndannoterede datasæt. Inputtet er ligesom ved word2vec-eksperimentet definitioner og citater fra DDO. En ekstra fordel ved denne tilgang er at vi direkte kan anvende outputtet fra modellen

2 BERT står for Bidirectional Encoder Representations from Transformers og er en tilgang der er mere fleksibel end fx den tidligere såkaldte word2vec-tilgang, idet den processerer teksten forfra og bagfra på samme tid og på den måde indfanger konteksten i beregningerne på en mere nuanceret måde.

som en afstandsscore mellem betydninger og på den måde afgøre om betydningerne ligger semantisk tæt på hinanden.

4.4. Foreløbige resultater

I figur 3 ses en score for hver tilgang anvendt på tre forskellige dele af de håndannoterede data, nemlig et sæt med alle de centrale CBC-lemmaer, et med kun de polyseme CBC-lemmaer der har 5 eller færre DDO-betydninger, og et med ikke-centrale polyseme lemmaer i DDO (dvs. lemmaer der ikke er i CBC-udvalget) der tilsvarende har 5 eller færre betydninger. Den regelbaserede tilgang udviser foreløbig de bedste resultater med over 0,7 på alle datasæt.



Figur 3: Score for de tre tilgange sammenlignet med håndannoteringen.

Af de statistiske modeller fungerer BERT bedst; modellen opnår konsekvent bedre scorer end word2vec-modellen. Dog er forskellen mindst hvad angår de almindelige (ikke-centrale) polyseme lemmaer med mellem 2 og 5 betydninger. Her opnår begge modeller deres bedste score. Det viser også at statistiske modeller har

svært ved at håndtere meget finkornede betydningsadskillelser, og det skyldes bl.a. et klassisk problem med dataknaphed. Kun ét citat og én definition fra DDO er simpelthen ikke fuldt tilstrækkeligt datamateriale til statistisk beregning. Et udsnit af de automatiske sammenlægninger er gennemgået manuelt, og sandsynligvis vil kun de lemmaer der har 4 eller færre betydninger i DDO, kunne indsættes direkte i COR-S, de øvrige skal behandles manuelt.

5. Konklusioner

Det er en stor udfordring at samle leksikalske data fra de mange forskellige eksisterende ordbogsressourcer og ikke mindst at forenkle det betydningsinventar som har dannet grundlag for deres tilblivelse. Det kræver en del leksikografisk arbejde, særligt for stærkt polyseme ord, men de automatiske metoder vi har udviklet i projektet, giver gode resultater for lemmaer med 4 eller færre betydninger i DDO.

Det leksikografiske arbejde har givet os nye indsigter i det danske ordforråd, fx hvad angår mønstre af systematisk polysemi, og hvad angår fordelingen af ontologiske typer blandt monoseme ord. De grundige analyser af betydningsinventaret og de manuelle opmærkninger vil efterfølgende kunne anvendes i mange sammenhænge i redigeringen af DDO. Vi har fx registreret en del betydninger der i dag er blevet sjældne og gammeldags, 25 år efter at den første udgave af DDO blev redigeret. Det forenklede COR-S-betydningsinventar (der stadig er koblet til DDO-betydningerne) vil kunne danne grundlag for DSL's arbejde med at udgive en forenklet udgave af DDO, fx til skolebrug.

Planen er at COR-S-ressourcen efter projektets afslutning skal opdateres årligt med nye ord baseret på *Retskrivningsordbogens* løbende udvidelse. Vi håber at der bliver mulighed for at arbejde videre med et modul der beskriver de mange faste udtryk i DDO når

første version af COR-S er færdig; disse er naturligvis vigtige i en sprogteknologisk ressource. Også syntaktiske oplysninger fra *STO*, en ordbog med formaliserede syntaktiske oplysninger om en stor del af *Retskrivningsordbogens* og DDO's lemmaer, skal på længere sigt kobles på. Eksterne brugere af COR-registret vil forhåbentlig finde stor nytte i at tilkoble domænespecifikke ordbøger og i den sammenhæng få gavn af at almenordforrådet i forvejen er dækket af COR-S.

6. Litteratur

Ordbøger og digitale ressourcer

Begrebsordbogen = *Den Danske Begrebsordbog*. Sanni Nimb (hovedred.), Henrik Lorentzen, Liisa Theilgaard & Thomas Troelsgård (2015). Det Danske Sprog- og Litteraturselskab og Syddansk Universitetsforlag.

DanNet, se: <cst.ku.dk/projekter/dannet/>, <andreord.nors.ku.dk> (april 2022).

DDO = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab, se: <ordnet.dk/ddo> (april 2022).

FrameNet, se: <framenet.icsi.berkeley.edu/fndrupal/> (april 2022).

Global WordNet Association, se: <globalwordnet.org> (april 2022).

Princeton Wordnet, se: WordNet Base Concepts.

Retskrivningsordbogen. Dansk Sprognævn, se <dsn.dk/ordboeger/retskrivningsordbogen/> (april 2022).

STO = Sprogteknologisk Ordbase. Anna Braasch et al. (red.) København: Center for Sprogteknologi. <cst.dk/cgibin/defisto> (april 2022).

WordNet Base Concepts, se: <globalwordnet.org/resources/gwa-base-concepts/> (april 2022).

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.

Fellbaum, Christiane (ed.) (1998): *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Firth, John Rupert (1957): *Studies in Linguistic Analysis.* Special Volume of the Philological Society. Oxford: Blackwell.

Geeraerts, Dirk (2002): The theoretical and descriptive development of lexical semantics. I: Leila Behrens & Dietmar Zaefferer (eds.): *The Lexicon in Focus. Competition and Convergence in Current Lexicology*. Frankfurt: Peter Lang Verlag, 23-42.

Harris, Zellig (1954): Distributional structure. I: *Word* 10 (23), 146-162.

Hjorth, Ebba & Kjeld Kristensen (red.) (2003-2005): *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab og Gyldendal.

Mikolov, Tomas, Wen-tau Yih & Geofffrey Zweig (2013): Linguistic regularities in continuous space word representations. I: *Proceedings of the 2013 conference of NAACL: Human language technologies*, Atlanta, Georgia, 746-751.

Nimb, Sanni (2018): The Danish FrameNet Lexicon: method and lexical coverage. I: *Proceedings of the International FrameNet Workshop at LREC 2018*, Miyazaki, Japan, 51-55.

Nimb, Sanni, Sussi Olsen, Bolette S. Pedersen & Thomas Troelsgaard (2022): A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. I: *Proceedings of the 13th LREC conference*, Marseille, Frankrig, 2826-2832.

Nimb, Sanni, Nicolai H. Sørensen & Thomas Troelsgård (2018): From standalone thesaurus to integrated related words in the Danish Dictionary. I: *Proceedings from Euralex 2018*, Ljubliana, Slovenien, 916-923.

Nimb, Sanni, Anna Braasch, Sussi Olsen, Bolette S. Pedersen, Anders Søgaard (2017): From Thesaurus to FrameNet. I: *Electronic Lexicography in the 21st century: Proceedings of eLex 2017 conference*, Leiden, Holland, 1-22.

Olsen, Ida R., Bolette S. Pedersen & Asad Sayeed (2020): Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources. I: *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, Marseille, Frankrig, 45-52.

Pedersen, Bolette Sandford (2018): Semantisk processering og leksikografi. I: Ásta Svavarsdóttir, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdís Úlfarsdóttir (red.): *Nordiske Studier i leksikografi* 14. Reykjavík: Nordisk forening for leksikografi, 18-28.

Pedersen, Bolette S., Nathalie C. H. Sørensen, Sanni Nimb, Ida Flörke, Sussi Olsen, Thomas Troelsgård (2022): Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open-Source COR Lexicon. I: *Proceedings of the 13th LREC Conference*, Marseille, Frankrig, 51-60.

Pedersen, Bolette S., Sanni Nimb og Sussi Olsen (2021): Dansk betydningsinventar i et datalingvistisk perspektiv. I: *Danske Studier* 2021. Odense: Syddansk Universitetsforlag & Universitets-Jubilæets danske Samfund, 72-106.

Pedersen, Bolette S., Sanni Nimb, Ida Rørmann Olsen, Sussi Olsen (2019): Merging DanNet with Princeton WordNet. I: *Proceedings of the 10th Global WordNet Conference2019 Proceedings*, Wroclaw, Poland: Oficyna Wydawnicza Politechniki Wroclawskiej, 125-134.

Pedersen, Bolette S., Sanni Nimb, Sussi Olsen & Nicolai H. Sørensen (2018a): Combining Dictionaries, Wordnets and other

Lexical Resources – Advantages and Challenges. I: *Globalex Proceedings 2018*, Miyasaki, Japan, 102-105.

Pedersen, Bolette S., Manex Aguirrezabal Zabaleta, Sanni Nimb, Sussi Olsen & Ida Rørmann Olsen (2018b): Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. I: *Proceedings of Global WordNet Conference 2018*. Singapore: Global WordNet Association, 182-189.

Pedersen, Bolette S., Anna Braasch, Anders Johannsen, Héctor Martínez Alonso, Sanni Nimb, Sussi Olsen, Anders Søgaard & Nicolai Hartvig Sørensen (2016): The SemDaX Corpus – sense annotations with scalable sense inventories. I: *Proceedings of the 10th LREC conference.* Portorož, Slovenien: European Language Resources Association (ELRA), 842-847.

Pedersen, Bolette S., Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen & Henrik Lorentzen (2009): DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. I: *Language Resources and Evaluation* 43, 269-299.

Pustejovsky, James (1995): *The Generative Lexicon: A Theory of Computational Lexical semantics*. Cambridge: MIT Press.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker & Jan Scheffczyk (2016): *FrameNet II: Extended Theory and Practice*. <framenet.icsi.berkeley.edu/fndrupal/the_book> (april 2022).

Sørensen, Nicolai Hartvig & Sanni Nimb (2018): Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings. I: *Proceedings from Euralex 2018*. Ljubliana, Slovenien: Ljubljana University Press, 819-824.

Vossen, Piek (1999): *EuroWordNet General Document*. <archive.illc.uva.nl//EuroWordNet/docs.html> (april 2022).

Sanni Nimb, ledende redaktør, ph.d.
Ida Flörke, assisterende redaktør, cand.mag.
Thomas Troelsgård, seniorredaktør, cand.mag.
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København
{sn, if, tt}@dsl.dk

Bolette Sandford Pedersen, professor, ph.d.
Nathalie Carmen Hau Sørensen, videnskabelig assistent, cand.mag.
Sussi Olsen, videnskabelig medarbejder, cand.mag.
Center for Sprogteknologi
Københavns Universitet
Emil Holms Kanal 2
2300 København S
{bspedersen, nmp828, saolsen}@hum.ku.dk

Brukarmedverknad i utvikling av nettsida ordbøkene.no

Margunn Rauset

The two Norwegian dictionaries *Bokmålsordboka* and *Nynorskordboka* have recently launched a common search interface at ordbokene.no. During the development of the new website the University of Bergen has emphasized user interaction. Valuable information and feedback have been obtained from in-depth interviews with students from different age groups and their tutors, as well as from emails, social media and pop-up surveys on the old as well as the new webpage.

1. Bakgrunn

Bokmålsordboka (BOB) og *Nynorskordboka* (NOB) har vore tilgjengelege på nett i nærare 30 år, og ordbok.uib.no har vore ei velkjend og mykje brukt side. Sidan 2016 har Språksamlingane ved Universitetet i Bergen (UiB) forvalta desse ordbøkene, og styringsgruppa for Språksamlingane bestemte i 2017 at dataa og funksjonane i samlingane skulle overførast til ny teknologi. I 2020 var tida komen for å utvikle ei ny nettside for ordbøkene, og eit utviklarteam sett saman av fagleg leiar for Språksamlingane, IT-utviklarane knytte til Språksamlingane, representantar frå IT-avdelinga og eg som leiar for ordbokredaksjonen har stått for gjennomføringa. Etter ein kort gjennomgang av kvifor vi valde å lage ei ny side, ordbøkene.no, heller enn å forbetre den gamle, gjer artikkelen greie for korleis utviklarteamet på ulike stadium har kommunisert med brukarane og fått innspel frå ekspertar på digital tenesteutvikling for å identifisere behov og velje løysingar på grunnlag av brukaråtferd på nett i dag.

1.1. Motivasjonen til UiB for å lage ny side

BOB og NOB, som med ei fellesnemning ofte blir kalla standardordbøkene, har sidan 2018 gått gjennom den mest omfattande innhaldsrevisjonen sidan dei første trykte utgåvene kom i 1986 (Rauset 2019, Selback 2020). Redaksjonen i Revisjonsprosjektet (2018–2024) har fått gjennomslag for at standardordbøkene heretter ikkje skal trykkjast, så vi redigerer no for eit heildigitalt format, der vi ikkje treng ta plassomsyn. Det betyr mellom anna at redaktørane tek inn mange nye ord, ein del nye tydingar, fleire faste uttrykk og fleire bruksdøme.[1]

Språkrådet og UiB eig standardordbøkene i fellesskap, og UiB har ansvar for å drifte og utvikle den datatekniske sida. No er det fleire grunnar til at UiB meiner at det oppdaterte innhaldet bør presenterast i ei oppdatert innpakning. Ein ting er at vi over tid har fått tilbakemeldingar frå skuleverket om at yngre brukarar oppfatta det gamle grensesnittet som så utdatert at dei ikkje ville bruke nettsida, men heller valde andre og ofte lite kvalitetssikra ressursar. Den spesielle statusen til BOB og NOB som den viktigaste kanalen Språkrådet har ut til språksamfunnet for å formidle dei offisielle, fulle og oppdaterte normene i bokmål og nynorsk, gjer dette særleg alvorleg. Elevar, lærarar og læremiddelforfattarar er nemleg, saman med tilsette i offentleg sektor, dei som er forplikta til å følgje dei offisielle normene; for alle andre er dei meir som ei rettesnor å rekne (Fretland & Søyland 2013:18–19).

Rundt ein tredjedel av søka i standardordbøkene skjer på mobiltelefonar, nettbrett og andre handhaldne einingar. Sjølv om den gamle visningssida ordbok.uib.no har hatt responsivt design i nokre år, var det potensial for å la innhaldet bli endå betre tilpassa storleiken på skjermen det blir søkt frå.

Alle norske offisielle nettsider pliktar frå vinteren 2023 å ha universell utforming. Nettsida ordbok.uib.no blei i 2021 meld inn

1 Sjå nettsida til Revisjonsprosjektet (jf. litteraturlista) for meir info.

for Diskrimineringsnemnda – eit nøytralt forvaltingsorgan som avgjer klagar på diskriminering, trakassering og gjengjelding – grunna manglande universell utforming. Klagen, som framleis er under handsaming, gjaldt mangelfull funksjon på sida ordbok.uib.no for blinde og svaksynte brukarar. Saka har medverka positivt til at vi har dette perspektivet langt framme i utviklinga av ei ny side.

Standardordbøkene og nettsida ordbok.uib.no er ein del av Språksamlingane, som UiB overtok frå Universitetet i Oslo i 2016 (Kyrkjebø & Myking 2017). Dei datatekniske løysingane i Språksamlingane hadde ei rekkje svake sider, mellom anna manglande dokumentasjon, delvis manglande kjeldekode og utdatert teknologi. Til dels utgjorde dette ein stor risiko med stort sannsyn for alvorlege hendingar og for at desse kunne få alvorlege konsekvensar (PwC 2017). UiB har difor gått for å utvikle ein heilt ny dataarkitektur. Målet er at den nye tekniske løysinga skal vere brukarvennleg, tilgjengeleg via opne grensesnitt (API) og forankra i moderne arkitekturprinsipp.[2] Dette har òg utløyst behov for ei ny nettside for ordbøkene som køyrer mot den nye arkitekturen.

2. Identifisering av behov

2.1. Innleiande bidrag frå Netlife

Forsommaren 2020 engasjerte UiB design- og teknologibyrået Netlife for å identifisere sterke og svake sider ved den gamle nettsida og hjelpe oss med å utvikle ei funksjonell nettside basert på brukarbehov. Dei laga designskisser, slik at vi tidleg kunne gå for

2 Det er prinsipielt viktig for norske styresmakter, Språkrådet og UiB at ordbokdataa er så opne som mogleg, og frå 03.02.21 har ordbokinnhaldet vore tilgjengeleg på open lisens via visnings-API-et (Språksamlingane 2021). Innhaldet kan brukast til kva formål som helst, inkludert kommersielle, og på halvanna år har 150 personar og miljø vendt seg til Språksamlingane for å få tilgang.

ein profil med omsyn til fargar, ikon, typografi, luft på sida osv. Parallelt såg dei på brukarstatistikkane til ordbøkene og involverte utviklarteamet, ordbokredaksjonen og Språkrådet i diskusjonen om kva perspektiv som er viktige frå eigar- og redaksjonshald.

Basert på at standardordbøkene er så viktige i skuleverket, definerte UiB elevar, studentar og undervisarar på ulike trinn i utdanningsløpet som primærmålgruppe for den nye sida. Frå starten har vi vore medvitne om at andre brukargrupper kan ha andre behov og ønske, men tanken var at om vi utviklar ei nettside som er intuitiv, tiltalande og velfungerande for undervisningssektoren, vil det vere ei god side for fleirtalet av dei andre brukarane òg.

Hausten 2020 gjennomførte Netlife halvannan times djupneintervju med to ungdomsskuleelevar, ein norsklærar i ungdomsskulen, ein elev i vidaregåande skule, ein masterstudent på universitetet og ein universitetslektor i norsk som framandspråk. Med utgangspunkt i ein spesialutvikla intervjugaid sat Netlife saman med éin og éin av intervjuobjekta. Utover å observere brukaråtferda på ordbok.uib.no fekk dei snakka om generell brukaroppleving og kva som var greitt eller vanskeleg å finne ut av på ordboksida.

Eit viktig funn i denne kvalitative undersøkinga var at mange opplevde nettsida som lite intuitiv, og at det difor var nødvendig med grundig opplæring i bruk av denne spesifikke nettsida, ikkje berre ordbokbruk generelt. Gjennomgåande blei bøyingskodane (som «v1», «m2» eller «adj.») i artikkelhovuda omtalte som eit problem. Mange forstod ikkje at dette var klikkbare lenkjer som kunne opne fulle bøyingstabellar, og dei klarte heller ikkje å dra nytte av den systematiske grupperinga av bøyingsmønster som kodane representerer.

Ein del elevar sa at dei aldri hadde høyrt om eller brukt *Bokmålsordboka* og *Nynorskordboka*, men då dei gjekk inn på ordbok.uib.no, var reaksjonen at dei hadde brukt sida mange gongar. To årsaker til at somme elevar sa at dei valde vekk ordbok.uib.no, var at dei opplevde sida som gammaldags og 90-talsaktig, og at ho

ikkje tilbyr omsetjing mellom bokmål og nynorsk. Visse andre opne nettordbøker tilbyr omsetjing, men der er kvalitetssikringa av innhald og normer svak.

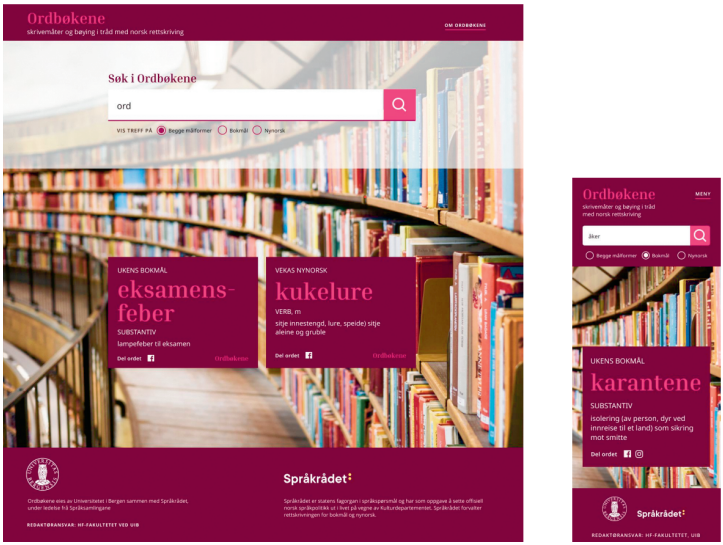
2.2. Viktige prioriteringar og tidlege avklaringar

Den vidare utviklinga av designet bygde både på dei første avklaringane med UiB og Språkrådet og på innspela frå brukarane i djupneintervjua. Eit hovudbodskap frå Netlife var at standardordbøkene må vere mykje lettare å kjenne igjen og skilje frå andre ordbøker. Dei la vekt på at dette ikkje handlar om merkevarebygging for Språkrådet og UiB, men om at brukarane må få vite at innhaldet på denne sida er påliteleg og korrekt. I ei tid der kjeldekritikk blir stadig viktigare, og der maskingenerert ordbokinnhald er like tilgjengeleg som grundig kvalitetssikra ordbøker, som våre, må det vere krystallklart formidla kven som står bak nettsida, og at det er her ein finn dei oppdaterte og fulle offisielle normene.

Tilrådinga frå Netlife var difor å velje namn, fargar og utforming som gjer at ordboksida skil seg ut, samstundes som ho er tiltalande – og gjer det klart kvifor dette er sida ein må bruke om ein skal skrive rett. Fellesnemninga Ordbøkene og url-en ordbokene.no er lette å hugse, og nemninga er allereie kjend frå appen vår Ordbøkene, som kom i 2017 og er mykje nytta. Frå redaktørhald er det eit poeng at den bundne fleirtalsforma får fram at BOB og NOB er to separate ordbøker som utfyller kvarandre, ikkje ei tospråkleg ordbok mellom bokmål og nynorsk.

Heilt frå dei første designskissene blei det lagt stor vekt på at ein må informere tydleg om kva type informasjon ein gjev. Dette harmoniserer godt med den redaksjonelle avgjerda om å no lage heildigitale ordbøker, der vi ikkje tek plassomsyn, men kan leggje oss i selen for å utforme eit pedagogisk oppsett. Eit av dei sentrale elementa er å innføre overskriftene «opphav», «uttale», «tyding», «døme» og «faste uttrykk», for å lette navigeringa.

Eit anna perspektiv vi har hatt med oss frå starten, er at innhaldet på handhaldne einingar kan og bør presenterast annleis enn på store skjermar. Figur 1 er ei Netlife-skisse av korleis velkomstsida til ordbøkene.no kunne sjå ut på høvesvis pc-skjerm og mobilskjerm.



Figur 1: Systematisk arbeid med responsivt design.

Søk er udiskutabelt den primære brukaroppgåva på sida, og det må vere like lett på små som på store skjermar. Det er utfordrande å utforme søkjefunksjonalitet og ei visning av treffa som fungerer like godt for to ordbøker med ulike skriftspråk samstundes, og det er noko som blei lagt stor vekt på å løyse på ein god måte.

At det skal vere lett å finne fram på den nye sida utan opplæring, har vore ein hovudprioritet. Kodane med grammatisk informasjon blei erstatta av ein eigen knapp som eksplisitt heiter «vis bøying», sjå figur 2. I tillegg har knappen eit +-teikn, slik at det ikkje kan vere til å misforstå at ein kan trykkje der for å få meir informasjon.

Figur 2: Utforming av bøyingsknapp i BOB.

Dersom ein klikkar på denne knappen, dukkar det opp ein tabell som viser den fulle bøyinga til ordet.

3. Brukarkontakt etter lansering

Det er IT-utviklarane knytte til Språksamlingane som med utgangspunkt i skissene frå Netlife laga nettsida. Redaksjonen bidrog med testing og kom med ytterlegare ønske. Då UiB kunne lansere ein betaversjon 1. oktober 2021, oppmoda vi brukarane om å gje oss tilbakemeldingar på e-post. Dette formidla vi både gjennom botnteksten på nettsida («*Bokmålsordboka* og *Nynorskordboka* viser skrivemåte og bøying i tråd med norsk rettskriving. Språkrådet og Universitetet i Bergen står bak ordbøkene. Gi oss gjerne tilbakemelding på ordbok.beta@uib.no.»), i universitetsavisa *På Høyden* (Ryen 2021), i NRK (Svedal 2021) og i sosiale medium (Facebook-sidene til ordbøkene og til Språkrådet).

3.1. Organisering av tilbakemeldingane

Med tanke på dokumentasjon, og for å halde styr på e-postane og svar på dei, oppretta utviklarteamet ei side i programvara Git, der meldingar som går på funksjonalitet og utforming av ordbøkene.no, blir lagde inn. Det blir konsekvent gjort for kritiske spørsmål og alt som konkret kan hjelpe oss i vidareutviklinga av nettsida, men ikkje nødvendigvis for tilbakemeldingar frå folk som berre gjev tommel opp for den nye sida. I tillegg til e-postane har vi òg oppretta Git-saker på relevante tilbakemeldingar og

spørsmål som har kome i sosiale medium. Erfaringa er at det er enklare å finne tilbake til gamle saker og å halde oversikt når ein får samla alt på éin stad.

Vi har fått så mykje ut av e-postutvekslingane med brukarane at alle som har sendt e-post om ordbøkene.no, fram til nyleg har fått eit personleg svar. Har det vore fleire e-postrundar med same brukaren om same tema, legg vi det i same Git-sak. No meiner vi at vi har kome så langt med utviklinga av sida at vi har lagt inn automatisk svar om at alle meldingar blir lesne og tekne stilling til, slik at vi meir kan velje kva e-postar vi brukar tid på å manuelt svare på.

Vel eit halvår etter lanseringa av betaversjonen var i underkant av 200 saker registrerte i Git. Ein overordna konklusjon er at ordbøkene har mange tilhengjarar som engasjerer seg sterkt i korleis ressursen blir utvikla. Med ei slik mengde innspel blir det viktig å prioritere. Når det er noko mange brukarar etterlyser og kommenterer, har det fått større vekt enn om éin enkeltbrukar saknar eller ikkje likar noko. Samstundes er det ei kompetent gruppe med IT-utviklarar og leksikografar som står på mottakarlista av e-postane, og sjølvsagt vurderer vi verdien av innspela. Ofte gjev svaret på e-postane seg sjølv, og underteikna har kunna svare brukarane på direkten, ein del avklaringar tek vi på e-post for å fortløpande kunne kvittere ut sakene, og meir kompliserte spørsmål har utviklarteamet diskutert på fellesmøta som vi har ein gong i veka.

3.2. Negativ respons

Mange e-postar inneheld konkrete forslag til forbetringar, og det har vore til stor nytte. I den første perioden handla mange om den grafiske utforminga og at det var behov for opprydding i skrifttypar og -storleik, linjeavstand, avrunding på hjørna på dei ulike felta osv. Her var det ein del kyndige grafikarar som uttalte seg, og mykje av kritikken dei kom med, var rimeleg.

Eit stadig tilbakevendande tema er rosafargen. Som ein brukar skreiv på sosiale medium: «På pluss-sida: Det er eit stort framsteg at det står 'sjå bøying'. Det burde gjera det mykje lettare å finne bøyingsformene. På minussida: Treng det vere så rosa? Ein annan fargebruk ville gi meir tyngde – trur eg, då.» Det vanlegaste argumentet mot fargen er at han er slitsam for auget og verkar forstyrrande. Vi har teke innspela på alvor i den forstand at vi mellom anna har lagt inn kvit som bakgrunnsfarge i artiklane og brukt mindre rosa i bøyingstabellane. Når vi likevel har valt å halde på fargepaletten, er det fordi vi har gjort eit medvite val om å vilje skilje oss ut i ordbokverda med den utforminga sida har fått. Vi trur at ein del av motviljen handlar om at fargebruken og totalpakken er annleis, og vi ønskjer å sjå det an om brukarane ikkje aksepterer dette når dei først har kome over «rosasjokket».

Det er ikkje til å stikke under stol at ein del brukarar totalt sett ikkje likar den nye sida og endringane, som desse tilbakemeldingane vitnar om: «Har alltid likt disse ordbøkene og har aldri ønska noen ny versjon. Den nye må man venne seg til… ☹» og «Elendig side. Fatter ikke hvorfor dere fjerner den gamle som er helt fantastisk og som har fungert i alle år». Med tanke på at ordbøkene har vore på nett sidan 1994 og mykje av funksjonaliteten har vore nokså lik sidan, er det forståeleg at somme reagerer. Dei fleste ønskjer å kjapt finne svaret på det dei lurer på, ikkje å bruke tid på å finne ut av korleis ein søkjer.

For dei som kjende ordbok.uib.no inn og ut etter mange års bruk, var sida enkel å bruke og effektiv ved at ho var så informasjonstett, slik denne brukaren set ord på det: «Må si at grensesnittet i forrige versjon var nydelig å jobbe i, og svært intuitivt. Det er derfor ubegripelig at dette ikke kunne videreføres ved ny lansering.» Vi ønskjer så klart ikkje å irritere brukarane våre, og det er ein balansegang å utforme sida slik at ho er intuitiv for nye brukarar og i tråd med dagens standardar, samstundes som ho skal vere

attkjenneleg for dei mange storbrukarane av sida gjennom år, som ikkje bør oppfatte at dei får ei dårlegare teneste.

3.3. Døme på funksjonalitet vi har lagt til

Ofte er det små detaljar som avgjer kor smidig ei side fungerer. Vi har mange døme på forbetringar som har kome etter brukarinnspel; til dømes etterlyste ein brukar ein X i søkjefeltet, som fjernar førre søk. Ein annan brukar ønskte ein snarveg til søkjefeltet, og det har hen fått ved at ein no kan trykkje Shift + 7. Då vi lanserte betaversjonen, stod ikkje markøren i søkjefeltet når ein kom inn på sida, men eit slikt autofokus blei raskt etterlyst og lagt til. Med unntak av autofokus var ingen av desse funksjonane tilgjengelege på den gamle sida, og utviklarane har lagt vekt på å velje løysingar som brukarane kjenner frå andre nettsider og søkjemotorar som Google.

Det er fleire som gjev uttrykk for at det er for mykje luft på den nye sida, og at det gjer det vanskelegare å få oversikt over innhaldet, særleg ved søk som gjev treff i fleire artiklar og i lange artiklar. Dette har vi forsøkt å kome i møte med å lage ein funksjon der ein kan komprimere artiklane, sjå figur 3 og skyve-knappen lengst til høgre under søkjefeltet.

Komprimeringsfunksjonen er tilgjengeleg for dei som ønskjer det, til dømes om ein gjer store søk som alle adjektiv som startar på *e*, som figur 3 viser eit utsnitt av. At det no står kor mange treff eit søk gjev i kvar ordbok, er òg ein brukar å takke, som etterlyste det.

Mange unngår helst å lese bruksrettleiingar, og dei skreiv heller e-post til oss enn å gå til menyen oppe i høgre hjørnet på nettsida for å sjekke om informasjonen låg der. Figur 4 viser dei fire lenkjene vi la i botnteksten for å tilgjengeleggjere informasjonen ytterlegare: «Om ordbøkene», «Hjelp til søk», «Innstillingar» og «Kontakt oss».

Figur 3: Skyve-knapp som gjev komprimert artikkelvisning.

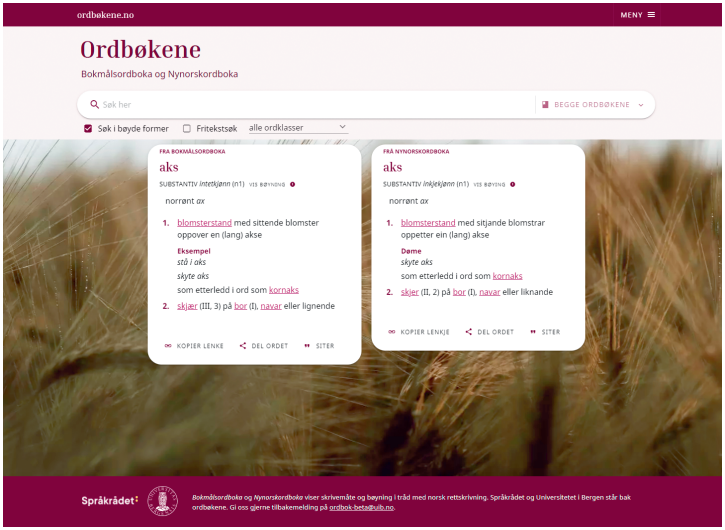
Figur 4: Ny utgåve av botnteksten frå mars 2022.

I tillegg har vi lagt inn dei oftast stilte spørsmåla i e-postane under posten «Kontakt oss». Erfaringa er at vi får færre e-postar som etterlyser funksjonalitet som allereie finst, etter at vi gjorde dette meir eksplisitt. Det er betre og meir effektivt både for brukarane og for oss som skal svare på tilbakemeldingane.

Her må det òg nemnast at vi aldri hadde trudd at vi skulle leggje inn igjen dei gamle bøyingskodane, men at folk sakna dei, er noko av det vi har fått aller flest tilbakemeldingar om. Har ein først lært seg kva dei betyr, ser ein kjapt kva bøyingsmønster eit lemma har, og ein treng ikkje klikke seg inn i bøyingsskjemaet. Vi har fått nokre riktig fornøgde brukarar når vi fortel at dei kan gå til innstillingar og leggje kodane til igjen. Dette hugsar nettlesaren til neste gong du besøkjer nettsida. Vi opplever det som ei fornuftig tilnærming at standardsøket er så enkelt som mogleg, men at dei

avanserte brukarane får fleire valmoglegheiter i innstillingane no enn før. Til dømes ser vi på det som eit stort framsteg at brukarane no kan velje om dei ønskjer metateksten på sida på bokmål, nynorsk eller engelsk. Det gjev mindre visuell støy enn å konsekvent vise alt på bokmål og nynorsk samstundes, slik det var på ordbok.uib.no.

Den funksjonaliteten som ligg fast under søkjefeltet, handlar om utviding eller avgrensing av søket, jf. figur 5.



Figur 5: Framsida av ordbøkene.no i mars 2022.

I dag kan ein hake av for å søkje i bøygde former (noko søkjeloggane viser at mange gjer) og fritekstsøk, og ein kan avgrense søket på ordklasse.[3] Alt dette er funksjonalitet etterlyst av brukarane. Tilpassingar av visninga, til dømes om du alltid ønskjer å få opp bøyingstabellane for å spare eit klikk, er lagt under innstillingar. Figur 5 viser dessutan at ein no kan kopiere lenkja til enkeltarti-

3 Søkjestatistikken har vore offentleg tilgjengeleg sidan ordbøkene.no tok over som offisiell visningsside 26. januar 2022.

klar, at ein kan dele ordet i appar som er tilgjengelege på eininga ein brukar, og at ein kan få opp forslag til korleis ein kan sitere den konkrete artikkelen.

4. Kvantitativ brukarundersøking

Via djupneintervjua med elevar/studentar og undervisarane deira i 2020 og dei i hovudsak erfarne ordbokbrukarane som skriv e-post til oss, har vi fått god innsikt i kva einskildpersonar meiner om nettsidene. Vi har lagt stor vekt på å lytte til desse innspela, men møter vi krava og forventingane til dei breie brukarmassane ved ein slik metodikk?

I samarbeid med Netlife utforma vi ei oppsprettsundersøking (eng. *pop-up survey*) til ordbok.uib.no i analysereiskapen Task Analytics. Det sentrale spørsmålet til brukarane var om dei fekk utført den oppgåva dei kom til nettsida for å løyse, men vi la inn ein del andre spørsmål òg. I løpet av rundt seks veker i november og desember 2021 svarte 7000 brukarar på undersøkinga, og av desse la over 2400 inn ein fritekstmerknad. Ikkje overraskande viste det seg at den vanlegaste oppgåva brukarane kom til nettsida for å løyse, var å sjekke bøying og annan grammatikk. Denne oppgåva oppgav litt under ein tredjedel (32,4 %) av brukarane. 24,2 % av brukarane kom for å sjekke tydinga til eit ord, og 22,4 % ønskte å sjekke skrivemåten.

Redaksjonen hadde lista opp tolv ulike oppgåver som vi rekna for dei mest sannsynlege at brukarane ønskjer å finne svar på. Med tanke på at standardordbøkene har som ei av sine viktigaste oppgåver å informere om offisielt normert skrive- og bøyingsmåte av norske ord, er det svært tilfredsstillande at dei skårar på topp (med over 90 % vellykka oppgåveløysing) i akkurat dei to kategoriane: «bøying og annan grammatikk» og «skrivemåte». Mellom 60 og 90 % (som ifølgje Netlife blir rekna som akseptable verdiar i Task

Analytics) av respondentane fann det dei var ute etter på områda tyding, bruksdøme, synonym, etymologi, samanlikning av bokmål og nynorsk og faste uttrykk. Sjå tabell 1 under for nøyaktige verdiar. Under 60 %, og det blir rekna som kritisk, av respondentane fekk utretta det dei kom for når oppgåva var omsetjing frå bokmål til nynorsk, omsetjing frå nynorsk til bokmål, informasjon om ordbøkene og det diffuse «anna». At såpass mange som 4,3 % av respondentane kom til denne sida primært for å få omsetjingshjelp, når ordbok.uib.no ikkje tilbaud denne hjelpa, er interessant og stadfestar at dette er eit behov.

Av annan nyttig informasjon kom det fram at 20 % av dei som svarte på undersøkinga, har eit anna morsmål enn norsk. Det er noko redaktørane må vere medvitne om, ikkje minst med tanke på definisjonsutforming, behov for bruksdøme og utfyllande informasjon om orda.

Det finst ikkje sikre tal for kor mange som har nynorsk som det primære skriftspråket sitt, men i skuleverket har i underkant av 12 % nynorsk som opplæringsspråk i 2021/2022 (Utdanningsdirektoratet 2022). Trass det svarte 26 % av respondentane at dei har nynorsk som hovudmål. Standardordbøkene er med andre ord svært viktige for nynorskskrivande, og det er ein av altfor få digitale ressursar som er like gode for begge skriftspråka. I fritekstfeltet er det òg fleire som trekkjer fram kor viktig parallellvisninga av skriftspråka er.

Statistikken på spørsmålet «Kven er du?» kom til dels noko overraskande på oss. I minkande rekkefølgje fordelte respondentane seg på denne måten: Eg skriv på jobben / er sakshandsamar (25 %), språkinteressert (22 %), undervisar (16 %), skribent eller omsetjar (9 %), ungdomsskuleelev (7 %), anna (7 %), elev på vidaregåande skule (4 %), student på andre fag enn språkfag (4 %), student på språkfag (3 %), student på norskopplæring (3 %). Med andre ord var posisjonen til ordbok.uib.no sterk blant dei vaksne

og truleg allereie nokså normsikre brukarane, men nettsida hadde ikkje den posisjonen blant elevar på ungdomsskulen og i vidaregåande opplæring som vi meiner ho burde ha.

Det er ei nyttig innsikt å ha med seg at for dei brukargruppene ordbok.uib.no faktisk hadde, var det lett å få gjort det dei skulle, truleg fordi dei hadde brukt nettsida over tid og kjende henne godt. Eller som ein brukar skreiv i fritekstfeltet: «Denne ordboka er flott utforma og brukarvenleg, og har i mange år vore ein fast følgesven gjennom arbeidsdagen. Skal de endra på noko på denne sida, så gjer det varsamt.»

Fleire tilbakemeldingar i fritekstfeltet stadfester inntrykket frå djupneintervjua om at lærarar tykte gamlesida var oversiktleg og god, men at elevane deira måtte få grundig opplæring og tykte ho var vanskelegare å bruke enn ein del andre nettordbøker. Det flest brukarar konkret etterlyste i fritekstfeltet, er koplingar mellom dei to ordbøkene som gjev hjelp til omsetjing mellom bokmål og nynorsk. Det store fleirtalet av dei som skreiv ein merknad, nytta likevel berre høvet til å takke for eit kvalitetssikra og fritt tilgjengeleg produkt som dei set stor pris på.

26. januar 2022 gjekk den nye nettsida frå å vere betaversjon til å bli hovudversjon, og i mars la vi ut ei tilsvarande oppsprettsundersøking på ordbøkene.no.[4] Tabell 1 oppsummerer kor mange prosent som fekk svar på det dei kom for, på gammal og ny side – og differansen mellom dei i prosentpoeng. Ein ønskjer flest mogleg grøne verdiar, som indikerer at over 90 % fekk svar på det dei kom for. Svarte verdiar er ok, men raude er kritiske.

4 I dag får brukarane beskjed om at innhaldet på ordbok.uib.no ikkje lenger blir oppdatert, og vi oppmodar dei til å bruke ordbøkene.no. Sjølv om fleire brukarar har bede om at gamleversjonen fortset å vere tilgjengeleg, gjer den manglande universelle utforminga at vi må stengje sida i første halvdel av 2023.

Oppgåve	gammal	ny	dif.
bøying og annan grammatikk	92,5 %	91,6 %	– 0,9
tyding	86,5 %	86,6 %	+ 0,1
skrivemåte	90,6 %	89,2 %	– 1,4
anna	59,9 %	54,7 %	– 5,2
døme på korleis ordet er brukt	77,6 %	76,1 %	– 1,5
synonym	64,5 %	67,1 %	+ 2,6
omsetjing frå bokmål til nynorsk	56 %	59,9 %	+ 3,9
etymologi	68 %	72,2 %	+ 4,2
samanlikning av bokmål og nynorsk	85,4 %	75 %	– 10,4
faste uttrykk	70 %	57,4 %	– 12,6
omsetjing frå nynorsk til bokmål	50,9 %	61,3 %	+ 10,4
Informasjon om ordbøkene	57,9 %	73,1 %	+ 15,2

Tabell 1: Samanlikning av kor mange som fekk utført det dei kom for, på gammal og ny ordbokside.

Rekkefølgja på oppgåvene i tabell 1 er styrt av kor mange brukarar som kryssa av for at dette var det dei primært kom for å gjere i den første oppsprettsundersøkinga. Dei største endringane ser vi, naturleg nok, på spørsmål knytte til funksjonalitet. Den største endringa, med + 15,2 prosentpoeng, er at fleire fann informasjon om ordbøkene. 10,4 prosentpoeng fleire fekk hjelp med omsetjing frå nynorsk til bokmål, 4,2 prosentpoeng fleire fann etymologien til ord (kanskje fordi vi tydlegare har skilt han ut med eiga overskrift?), og 3,9 prosentpoeng fleire fekk hjelp til omsetjing frå bokmål til nynorsk. På den sistnemnde oppgåva er vi så godt som oppe på svart nivå. At den nye nettsida gjev betre omsetjingshjelp, kan henge saman med at vi har lagt til ein «meinte du?»-funksjon basert på Levenshtein-distanse, som hjelper til der det er ortografisk likskap, men ikkje overlapp mellom skriftspråka.

Ei oppgåve som ordbøkene.no skårar merkbart dårlegare på enn gamlesida, er samanlikning av bokmål og nynorsk. 10,4 prosentpoeng færre får hjelp til denne oppgåva. Samanlikning av skriftspråka går typisk føre seg ved at ein søkjer i begge ordbøkene og får opp artiklane frå dei to ordbøkene side ved side i kvar sin kolonne. Standardoppsettet på både pc og mobil er framleis at ein søkjer i begge ordbøkene, og på pc fungerer parallellvisninga på same måte som på gamlesida. I fritekstfeltet er det fleire som skriv at dei opplever mobilvisninga som mindre oversiktleg, då resultata i bokmål og nynorsk blir viste meir om kvarandre enn tidlegare når det er meir enn to treff. Fleire respondentar etterlyser at ein sjølv kan styre rekkefølgja i mobilvisninga, som denne: «Fint om ein kunne velgja anten 'Begge, nynorsk øvst' eller 'Begge, bokmål øverst'.»

Aller størst nedgang, med 12,6 prosentpoeng, ser vi på kor mange som finn dei faste uttrykka dei er på jakt etter. Dette kan mogelegvis forklarast med at vi har tatt eit fagleg fundert val om å flytte uttrykka frå ulike tydingar og samle dei til slutt i artiklane. Kanskje ein del brukarar ikkje har oppfatta det enno – eller kanskje dei opplever det som uheldig? Ei anna forklaring kan vere at vi i undersøkingsperioden eksperimenterte med kompakt artikkelvisning ved meir enn to treff, og at dei faste uttrykka aldri blei viste i den kompakte artikkelen. Då var det ikkje alle brukarane som oppfatta at dei måtte klikke på «vis artikkel» for å sjå heile innhaldet.

I merknadsfeltet er det mange som gjev uttrykk for at ordbøkene.no er lettare å bruke enn ordbok.uib.no: «Den nye ordboka har blitt veldig elegant og brukarvennleg! Det må ligge eit enormt arbeid bak dette» og «Denne versjonen er mye bedre enn den gamle! Søkefeltet spesielt er bedre». Andre uttrykkjer skepsis: «Den nye nettsiden er penere, men mindre intuitiv og brukervennlig» og «Layouten var bedre før på ordbok.uib.no. Skjermen er for vid, og altfor mye skrolling for å finne det en leter etter». Med dette bak-

teppet svarer brukarane forbausande likt i dei to undersøkingane på kor lett eller vanskeleg det var å utføre oppgåva dei kom til nettsida for. Tabell 2 viser at på ein skala frå 1 til 7, der 1 betyr at det var svært enkelt å gjere det ein kom for, og 7 at det var svært vanskeleg, kryssa 82 % av respondentane av på 1 og 2 (alternativa som i høgast grad signaliserer at det var lett) for den gamle sida, medan 85 % haka av for dei same nivåa på den nye sida. Med skepsisen mange uttrykte for endringar, trur eg det må reknast som eit godt resultat nokre få månader etter overgang til ny side.

	1 – veldig lett	**2**	**3**	**4**	**5**	**6**	**7 – veldig vanskeleg**
Gammal side	65	17	6	4	3	2	2
Ny side	65	20	6	3	3	2	1

Tabell 2: Kor lett eller vanskeleg det var for brukarane å gjere det dei kom for.

Samanlikna med tilbakemeldingane vi fekk rett i etterkant av lanseringa av betaversjonen, er det påfallande få i den siste oppsprettsundersøkinga som er opptekne av fargebruken. Vi meiner også sjølve at vi har kome fram til ein betre balanse enn i dei tidlege designskissene, og somme brukarar nemner dette no som noko positivt: «Fargepaletten på sida er aldeles nydeleg.» Framleis får likevel somme andre assosiasjonar enn intendert: «Designet ser ut som nettbutikken til kondomeriet.»

Av andre moment som blir trekte fram, er «meinte du?»-funksjonen: «Fint at det kjem opp 'liknande treff', slik at eg fann ordet eg leitte etter (eg var usikker på skrivemåte).» Vi hadde ikkje venta å få kritikk for denne funksjonen, men ein gjeng scrabblespelarar kallar det juks når det dukkar opp forslag om ein har søkt på noko som ikkje ligg inne i ordboka. Enkelte uttrykkjer at dei fryktar at forslaga kan føre til misforståingar når brukarar får treff når dei skriv inn noko feil (sjå òg Dominczak, Antonsen & Trosterud i dette nummeret). For mange brukarar gjev det lite meining at ein

må kjenne skrivemåten til eit ord for å slå det opp i ei ordbok, så det er ikkje eit alternativ å fjerne funksjonaliteten igjen, men vi må passe på at markeringane er eintydige og klare. Somme kommenterer at dei likar at informasjonskategoriane er tydeleg skilde frå kvarandre med overskrifter, andre set pris på den nye utforminga av bøyingstabellane, og atter andre roser at sida har færre kodar og forkortingar.

Det er framleis omsetjarfunksjonen (dvs. koplingane mellom bokmål og nynorsk) som flest etterlyser, òg i ordpar med semantisk, men ikkje ortografisk likskap:

> Det hadde ofte vore bra om bm/nn-omsetjing var oppført, særleg i dei bokmålsorda som ikkje er godkjende for nn. Det hadde vore bra for mange elevar og studentar og generelt dei som har nynorsk som sidemål. Faktisk ville det ha vore veldig bra for mange av oss med nynorsk som hovudmål også.

At det er så mange som etterlyser dette, gjer at vi må ta det på alvor, men det vil krevje ressursar om vi skal få det til med tilfredsstillande kvalitet. Med andre ord er det ikkje ei oppgåve dagens redaksjon kan løyse i Revisjonsprosjektet, men det bør vere ei framtidig utviklingsoppgåve.

Andre ønske som går igjen, er fleire oppslagsord, fleire bruksdøme – ikkje minst døme som viser preposisjonsbruk –, fleire synonym, moglegheit til å logge inn for å markere favorittord og lage personlege ordlister, meir informasjon om uttale og gjerne lydfiler, og ny utgåve av appen i tråd med nettsida. Det er med andre ord rikeleg med oppgåver å gripe fatt i både i innhald og i form. Og arbeidet med innhaldsrevisjonen held fram, forhåpentleg òg etter at Revisjonsprosjektet er ferdig i 2024.

Sjølv om det finst enkelte kritiske røyster, er fleirtalet av merknadane svært positive til ordbøkene og dei endringane den

nye sida fører med seg. Vi kan nesten ikkje ønskje oss noka betre tilbakemelding enn denne: «Dette er ikke bare en av de viktigste nettsidene i Norge – det er en av de beste. Rask, reklamefri og med mye lærerik informasjon.» Men det blir likevel toppa av at prosentdelen som svarte at dei er ungdomsskuleelevar, har stige frå 7 % i undersøkinga på ordbok.uib.no til 11 % i undersøkinga på ordbøkene.no. Vi skal ikkje plage brukarane med oppsprettsundersøkingar i tide og utide, men vi ønskjer å følgje med på over tid om vi får fleire unge brukarar til å nytte standardordbøkene.

5. Oppsummering

Fleire brukarar har omtala den nye nettsida ordbøkene.no som «eit skikkeleg løft» for BOB og NOB, i tydinga 'klar forbetring', men arbeidet har òg vore eit løft i tydinga 'krafttak' for UiB. Det har blitt lagt ned store ressursar i utviklingsarbeid, og gjennom heile prosessen har vi vore i tett dialog med ordbokbrukarane. Vi knytte til oss ekspertar på digital tenesteutvikling som hadde klare råd om kva metodikk vi burde bruke for å skaffe informasjon om ulike brukargrupper. I tillegg til å gjennomføre ei brukarundersøking for oss utforma dei designskisser som IT-utviklarane knytte til Språksamlingane bygde på då dei laga sida ordbøkene.no.

Vi gjorde eit val om at utdanningssektoren skulle vere ein viktig premissleverandør for grunnstrukturane på sida til standardordbøkene, og så har dei meir erfarne ordbokbrukarane kome med innspel i andre kanalar seinare. Begge delar har vore avgjerande for at vi har fått ei side som er lett å ta i bruk og navigere på, men der dei som ønskjer det, kan gjere ei rekkje innstillingar og tilpassingar. Eit viktig mål har vore at sida skulle ha funksjonalitet og utforming som møter dagens krav og forventingar, samstundes som overgangen til nytt system ikkje skulle bli for stor for som dei som fram til no har utført nærare 60 millionar søk i året på ordbok.uib.no og i appen.

Litteratur

Ordbøker

BOB = *Bokmålsordboka*. Språkrådet og Universitetet i Bergen. <ordbøkene.no> (mars 2022).

NOB = *Nynorskordboka*. Språkrådet og Universitetet i Bergen. <ordbøkene.no> (mars 2022).

Annan litteratur

Fretland, Jan Olav & Aud Søyland (2013): *Rett og godt. Handbok i nynorskundervisning*. Oslo: Samlaget.

Kyrkjebø, Rune & Johan Myking (2017). Språk på flyttefot – Språksamlingane til Bergen. I: *Årbok for Universitetsmuseet i Bergen*. Bergen: Universitetet i Bergen, 91–94.

PwC (2017): Språksamlingene UiB: Teknisk vurdering. Upublisert.

Rauset, Margunn (2019): *Bokmålsordboka* og *Nynorskordboka* – einegga, toegga eller siamesiske tvillingar? I: *LexicoNordica* 26, 155–175.

Ryen, Pål Adrian Clausen (2021): Populære ordbøker i ny drakt. *På høyden* 22.10.2021. <pahoyden.no/ordbokene/populaere-ordboker-i-ny-drakt/114168> (mars 2022).

Selback, Bente (2020): «Å nei, det ordet er ikkje lov på nynorsk!» Eller ...? Om parallell redigering av to norske ordbøker. I: Caroline Sandström, Ulla-Maija Forsberg, Charlotta af Hällström-Reijonen, Maria Lehtonen & Klaas Ruppel (red.): *Nordiska studier i lexikografi* 15. Helsingfors: Nordisk förening för lexikografi, 297–305.

Språksamlingane (2021): Lisens for Bokmålsordboka og Nynorskordboka. <uib.no/ub/fagressurser/spesialsamlingene/142334/lisens-bokm%C3%A5lsordboka-og-nynorskordboka> (mars 2022).

Svedal, Maria Gunnarsdotter (2021): No skal ordbøkene bli enklare å bruka. NRK 31.10.2021. <nrk.no/vestland/no-skal-ordbokene-bli-enklare-a-bruka-1.15709338> (mars 2022).

Andre nettsider

Gammal nettvisning av BOB og NOB, tilgjengeleg fram til første halvdel av 2023. <ordbok.uib.no> (mars 2022).

Ordbøkene på Facebook. <facebook.com/ordbokene> (mars 2022).

Revisjonsprosjektet. <revisjonsprosjektet.no> (mars 2022).

Språkrådet på Facebook. <facebook.com/sprakradet.no> (mars 2022).

Søkjestatistikken til ordbøkene.no. <oda.uib.no/olog/> (august 2022).

Utdanningsdirektoratet. Fakta om grunnskolen 2021/22. <udir.no/tall-og-forskning/statistikk/statistikk-grunnskole/analyser/fakta-om-grunnskolen/fakta-om-grunnskolen/#farre-elever-velger-nynorsk-som-opplaringsform> (april 2022).

Margunn Rauset
forskar, ph.d.
Universitetet i Bergen
Institutt for lingvistiske, litterære og estetiske studium
Postboks 7805
NO-5020 Bergen
margunn.rauset@uib.no

Chatbots, dialogdesign og leksikografi?

Henrik Køhler Simonsen

This article discusses lexicography and chatbots and explores how lexicographic data can be combined with interaction and dialogue design to help healthcare professionals communicate in L2 (Danish or Swedish). The article draws on a literature review, a test of 30 chatbots and data from 4 test protocols. The insights uncovered, led to the formulation of new models on how to develop lexicographically designed chatbots. The article presents a theoretical framework for SKANDIBOT, which is an AI-based chatbot with a needs-oriented interaction design.

1. Indledning og problemformulering

Leksikografien står, som bekendt, over for et paradigmeskifte.[1] Konventionelle ordbøger har i længere tid været udfordrede af bl.a. oversættelsessystemer og AI-skriveassistenter, som påvist af f.eks. Simonsen (2020a, 2020b og 2021).

Samtidig finder leksikografiske data og leksikografiske metoder og processer også i stigende grad anvendelse i nye sammenhænge. En af disse nye sammenhænge kunne være sammen med chatbots, som her ses som en applikation, som man kan tale eller skrive med. Google Assistant eller Siri er eksempler på chatbots.

Anvendelsen af chatbots er steget i både virksomheder og i uddannelsessektoren. Det ses f.eks. på antallet af publikationer om chatbots i uddannelse og i sprogundervisning, som påvist af Wollny et al. (2021), og Smutny & Schreiberova (2020) peger endvidere

1 Der skal rettes en tak til NORDPLUS for støtte til projekt SKANDIBOT (NPLA-2021/10023), som artiklen her til dels er baseret på. Der skal også rettes en tak til projektpartnerne Olga Viberg fra Kungliga Tekniska Högskolan samt Thomas Troelsgård og Karen Skovgaard-Petersen fra Det Danske Sprog- og Litteraturselskab.

på, at chatbots er meget velegnede i sprogundervisningen. Endelig diskuteres en række tidlige overvejelser om SKANDIBOT-projektet i Simonsen & Viberg (2022).

Artiklen hviler på en hypotese om, at kombinationen af chatbots, leksikografisk design, leksikografiske data og behovsbaseret interaktions- og dialogdesign vil kunne løse en række kendte udfordringer. For det første bliver brugernes oplevelse og interaktion med en chatbot bedre, når den er baseret på leksikografisk behovsdesign, dvs. når chatbotten tager udgangspunkt i brugerens kommunikative, kognitive og operationelle behov og præsenterer de relevante typer af leksikografiske data for at tilfredsstille dette behov. For det andet løser kombinationen en anden for leksikografi så velkendt udfordring – nemlig brugerens uforudsigelige søgeadfærd og datatilgang.

På trods af årtiers forskning i disse spørgsmål har leksikografien stadig ikke helt fundet løsningen på, hvordan brugere finder det, de egentlig har brug for. Her kan et behovsbaseret interaktions- og dialogdesign, nøje udvalgte og tilpassede leksikografiske data og kunstig intelligens hjælpe brugeren med at identificere det, som vedkommende reelt leder efter ved at hjælpe med at afdække behovet vha. leksikografisk dialogdesign.

Artiklen trækker på teoretiske overvejelser om chatbots, interaktions- og dialogdesign samt leksikografi (f.eks. Wollny et al. 2021, Smutny & Schreiberova 2020, Thomas 2020, Fryer et al. 2020, Hughes 2022, Cooper et al. 2007, Sharp et al. 2007, McCarthy & Wright 2007, Bergenholtz & Tarp 1995, Tarp 2010 og Agerbo 2017).

Artiklen foreslår en række modeller, som kan anvendes til at udvikle læringseffektive og socialrelationelle chatbots vha. leksikografisk metode, leksikografisk interaktions- og dialogdesign samt kuraterede leksikografiske data. Ved faste, kuraterede datasæt forstås her statiske data, som er indtastet, redigeret og kvalitetssikret af f.eks. ordbogsredaktører, og ved dynamiske, ikke-kuraterede

datasæt forstås data, der løbende ændrer sig, og som ikke er blevet kurateret, redigeret eller kvalitetssikret.

2. Metode og data

Denne artikel diskuterer, hvorvidt og hvordan behovsbaseret interaktions- og dialogdesign, leksikografiske data og chatbots kan kombineres.

Metodisk hviler artiklen på fire metodiske elementer. For det første er der blevet gennemført en struktureret litteraturanalyse vha. EBSCOHost, som er en meget anerkendt database med videnskabelig litteratur, hvor relevante bidrag om chatbots er blevet selekteret og analyseret. Denne analyse resulterede i mere end 30 artikler om chatbots, som efterfølgende blev analyseret, og indsigter fra denne litteraturanalyse er blevet inddraget i artiklen

For det andet er der blevet gennemført en empirisk analyse og test af 30 udvalgte chatbots, som blev testet på udvalgte parametre. De 30 chatbots blev kategoriseret efter pris, AI-muligheder, designmuligheder, platformsmuligheder og kodningskompleksitet. Indsigter fra analysen af de mange chatbots er inddraget i artiklen.

For det tredje hviler artiklen på indsigter fra og erfaringer med design af AI-baserede chatbots i tre ERASMUS+-støttede projekter: IDEAS, DS4AIR og SKILLABILITY, som alle tre handler om anvendelsen af chatbots i undervisning og læring, og disse erfaringer har ligeledes spillet en vigtig rolle i udviklingen af SKANDIBOT.

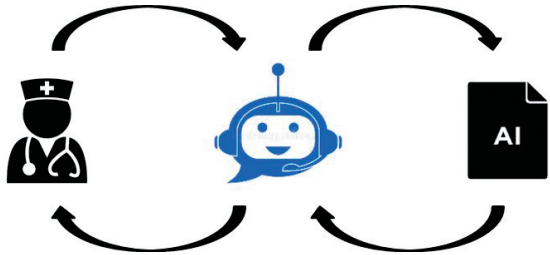
Endelig blev der gennemført en lille brugertest med to sundhedsprofessionelle og to læringskonsulenter, som blev bedt om at teste en tidlig version af chatbotten og samtidig svare på en række spørgsmål og lave en kort protokol. De fire protokoller er ligeledes blevet anvendt som en del af det empiriske grundlag i artiklen.

3. Intenderede brugere og leksikografiske behov

SKANDIBOT-projektet har til formål at udvikle en chatbot, som kan anvendes af personer med anden sproglig baggrund end dansk eller svensk. SKANDIBOT-chatbotten henvender sig primært til sundhedsprofessionelle, f.eks. sygeplejersker eller social- og sundhedsassistenter i den danske og svenske sundhedssektor og fokuserer på at give brugerne:

- kommunikativ hjælp (skrive og tale dansk og svensk)
- kognitiv hjælp (tilegne sig viden på dansk og svensk)
- operativ hjælp (få hjælp til at løse en konkret opgave).

For at kunne udvikle SKANDIBOT var det således først nødvendigt at udvikle en teoretisk model, der beskriver den særlige leksikografiske mediatorrolle. Som det fremgår af figur 1, manifesterer den leksikografiske mediatorrolle sig vha. chatbotten og dens behovstilpassede, leksikografiske interaktions- og dialogdesign, hvilket diskuteres senere i afsnit 5.



Figur 1: Den leksikografiske mediatorrolle.

Figur 1 viser, at brugeren konsulterer chatbotten, som i øvrigt virker på både mobiltelefon og computer, for at få hjælp til at kommunikere, få mere viden eller få konkret hjælp til løsning af en opgave. Chatbotten fortolker derefter brugerens spørgsmål og hensigter, og starter på basis af den indbyggede AI en leksikografisk dialog

med brugeren for at afdække den konkrete hensigt med spørgsmålet. Chatbotten henter/genererer derefter et svar på basis af de faste, kuraterede datasæt, som er indtastet i selve chatbotten, eller den søger i enten faste, kuraterede datasæt fra hhv. DSL (2022) og Språkbanken (2022) eller i dynamiske, ikke-kuraterede datasæt på f.eks. en organisations server, herunder kliniske retningslinjer etc.

4. Chatbots, dialogdesign og leksikografi

Forestillingen om at have en virtuel assistent, der kan hjælpe med alt, er ikke ny. Allerede i 1990'erne lancerede Microsoft den virtuelle assistent CLIPPY. CLIPPY blev dog hurtigt fjernet fra Windows, fordi den ikke tilbød relevant hjælp. Der er dog sket en del siden, og chatbots anvender nu også kunstig intelligens.

Petrović & Jovanović (2021) diskuterer den vigtige rolle, som chatbots spiller i sprogindlæring, og sprogindlæring med chatbots defineres af Huang et al. (2021) som:

> …the use of a chatbot to interact with students using natural language for daily practice (e.g., conversation practice), answering language learning practice and conducting assessment and providing feedback.

Denne forståelse dækker til dels også anvendelsen af chatbots i SKANDIBOT-projektet.

Ifølge Wollny et al. (2021) er antallet af publikationer om chatbots i sprogundervisning siden 2016 steget markant, og formålet med langt de fleste af disse chatbots er at understøtte de studerendes læring, gøre deres læring mere effektiv eller at øge de studerendes motivation. Anvendelse af chatbots i sprogundervisningen er formentlig steget betydeligt, fordi chatbots kan kommunikere med brugerne på deres eget sprog, jf. f.eks. Fryer et al. (2020), og

fordi de kan hjælpe brugerne med at finde det ønskede. Endvidere opstiller Thomas (2020:787) otte forskellige områder, hvor chatbots kan anvendes i sprogundervisningen, og SKANDIBOT-projektet gør især brug af "Customised learning", "Appealing methods of online education" og "Competent in language".

Endelig diskuterer Smutny & Schreiberova (2020) fire typiske interaktionsmetoder, hvor SKANDIBOT gør brug af de tre første interaktionsmetoder:

A. Button-based decision-tree hierarchy access
B. Keyword recognition-based access
C. Contextual utilize access
D. Voice-enabled access

Chatbots er ganske velegnede til at starte den dialog, som er så vigtig for brugerne, og forsøg med at gøre chatbots mere personlige og "empatiske" har ført til overraskende resultater. Hughes (2022) beskriver en chatbot på et stort amerikansk universitet, som i starten var helt ubrugelig. Det var først, da holdet bag chatbotten Ekhobot gav den en anden, mere empatisk og personlig chatbot-karakter, at den blev populær hos de studerende. Hughes (2022) konkluderer, at netop chatbottens dialogdesign og empati er afgørende for, om folk vil bruge chatbots.

Dette fører til den anden meget vigtige teoretiske byggesten, nemlig chatbottens interaktions- og dialogdesign. Lige præcis dialogen er helt afgørende for brugeren, og også her kan leksikografisk teori og metode spille en vigtig rolle. Begrebet interaktionsdesign vandt indpas i 1980'erne, i takt med at flere og flere interaktive teknologier blev lanceret. Et af de afgørende teoretiske bidrag inden for interaktionsdesign er Cooper et al. (2007), som i deres bog definerer interaktionsdesign som "the practice of designing interactive digital products, environments, systems, and services". Begrebet dialogdesign defineres endvidere af W3Computing (2022) således:

> Dialog is the communication between the computer and a person. Well-designed dialog makes it easier for people to use a computer and lessens their frustration with the computer system. Recall the elements of the TAM (technology acceptance model) indicating that perceived usefulness and perceived ease of use will lead first to an intention to use the system and eventually to using it.

Det handler således om den kommunikation, der foregår mellem mennesket og i dette tilfælde chatbotten, og det handler om i så vid udstrækning som muligt at designe en dialog, som afdækker og forstår brugerens hensigter. I citatet ovenfor omtales TAM (technology acceptance model), som fastslår, at jo mere brugervenlig en teknologi opleves, jo større er sandsynligheden for, at den pågældende teknologi anvendes og accepteres.

Sharp et al. (2007) beskriver endvidere begrebet affektiv interaktionsdesign således: "elements that influence user emotional responses and motivational, learning, creative, social and persuasive influences" og knytter dermed an til Hughes (2022). Det affektive dialogdesign synes at være meget relevant for SKANDIBOT. Endelig beskriver McCarthy & Wright (2007), hvordan teknologi skal ses som en oplevelse: "the user experience must take into consideration the emotional, intellectual, and sensual aspects of our interactions with technology", hvilket i stor udstrækning minder om mange af elementerne i den digitale forretningsmodel, som beskrives af Weill & Woerner (2018). Som det vil fremgå nedenfor, er dialogdesignet en afgørende faktor – især hvis det kombineres med en behovsorienteret leksikografisk tilgang.

Endelig er leksikografisk teori og metode naturligvis helt centralt for udviklingen af den teoretiske model, som ligger til grund for udviklingen af SKANDIBOT. For det første hviler den leksikografiske chatbotmodel på dele af funktionsteorien, som beskrevet af Bergenholtz & Tarp (1995). Især er funktionerne L2-produktion

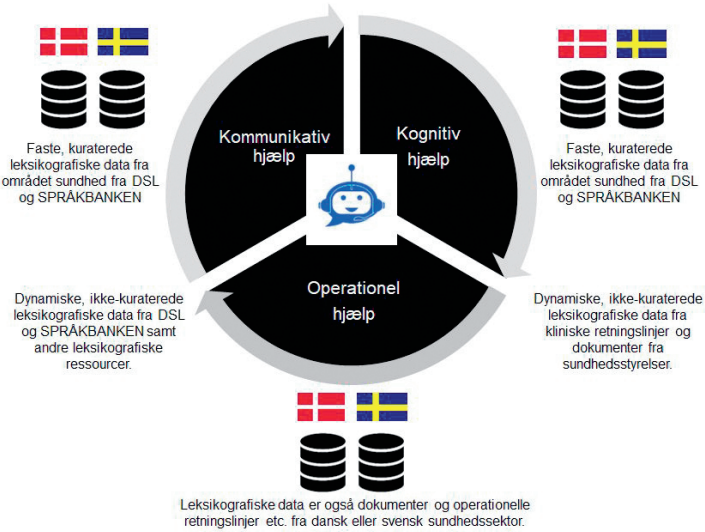
og L2-kognition helt essentielle for SKANDIBOT, fordi det handler om at hjælpe brugerne med hhv. at kunne kommunikere på dansk og svensk og at tilegne sig viden. Der trækkes også på Simonsen (2002:435-436), som foreslår en kombinationsmodel for funktioner, funktionaliteter og datalokation.

Den leksikografiske chatbotmodel anvender endelig overvejelser fra Tarp (2010) og Agerbo (2017), som diskuterer den operative brugersituation. En sådan situation er der tale om, når et opslagsværk hjælper brugeren med at løse en konkret opgave. Tarp (2010:177) skriver følgende om den operative situation: "...to give instructions on how to proceed in specific situations, e.g. in relation to the operation of machines and other instruments", hvilket synes at være relevant for SKANDIBOT. Agerbo (2017:385-387) diskuterer den operative funktion og medgiver, at der stadig er mange ubesvarede spørgsmål om, hvor f.eks. grænserne går mellem den operative funktion og den kognitive funktion. I SKANDIBOT-projektet handler det om at kunne tilbyde en bruger konkret hjælp til f.eks. at betjene et blodtryksapparat eller lægge et drop etc.

Kombinationen af ovenstående overvejelser og de empiriske indsigter har ført til en række nye teoretiske overvejelser, som placerer leksikografi helt centralt i udviklingen af nye leksikografisk designede chatbots.

5. Behovs- og dialogtilpasset chatbotdesign

SKANDIBOT-chatbotten er opbygget på basis af følgende leksikografiske funktions- og dataarkitektur, som tilfredsstiller de tre definerede brugerbehov.



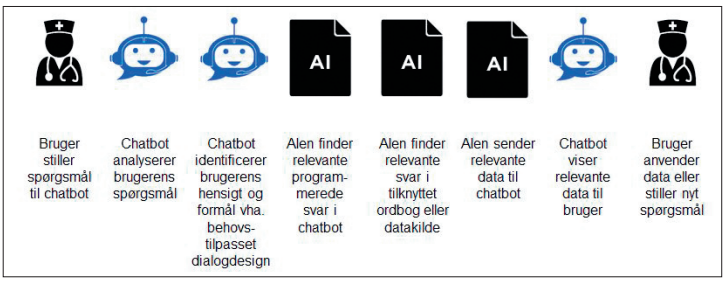
Figur 2: Leksikografisk funktions- og dataarkitektur.

Som det fremgår af modellen i figur 2, tilbyder chatbotten brugerne kommunikativ, kognitiv og operativ hjælp ved at trække på først og fremmest faste, kuraterede leksikografiske data fra DSL (2022) og Språkbanken (2022), samt fra dynamiske, ikke-kuraterede data fra dokumentsamlinger fra DSL (2022), Språkbanken (2022) og øvrige leksikografiske ressourcer. Chatbotten trækker også på andre dynamiske, ikke kuraterede data fra f.eks. det pågældende hospitals eller sundhedsstyrelses server, f.eks. kliniske retningslinjer, SOP'er (Standard Operational Procedures) eller andre dokumenter.

Med anvendelse af AI er der reelt tale om to brugere – nemlig den humane bruger, som interagerer med chatbotten, og den kunstige intelligens, som interagerer med de tilsluttede datasæt. Denne dyadiske brugerrolle bevirker, at der er opstået en ny leksikografisk mediatorrolle, som konstitueres vha. chatbotten. Mediatorrollen indebærer, at chatbotten først afdækker den humane brugers fak-

tiske intentioner vha. en inkluderende og behovsorienteret dialog, og dernæst kommunikerer den med AI'en, som så leder efter svar. Denne leksikografiske mediatorrolle blev illustreret i figur 1.

En tredje teoretisk overvejelse vedrører en model for den behovsopfyldelse, som sker vha. de leksikografiske data. Figur 3 viser et forslag til en model, som illustrerer den humane brugers interaktion med chatbotten, chatbottens interaktion med AI'en, AI'ens interaktion med chatbotten og til sidst chatbottens interaktion med den humane bruger.



Figur 3: Leksikografisk behovsopfyldelse med chatbot og AI.

Figur 3 viser hvordan den humane bruger konsulterer chatbotten, hvis fundament er leksikografisk og behovstilpasset dialogdesign, til at afdække brugerens faktiske intention. AI'en tager derefter over og finder svar fra enten faste, kuraterede datasæt (programmeret i chatbotten) eller dynamiske ikke-kuraterede datasæt (ekstraheret fra eksterne datasæt) og sender svaret retur til chatbotten, som tager svaret videre til den humane bruger.

Metasproget i SKANDIBOT er i øvrigt engelsk for at forbedre de intenderede brugeres interaktion med chatbotten, ligesom chatbotten præsenterer engelske ækvivalenter i den leksikografiske chatbot-artikel.

Kombinationen af de leksikografiske funktioner, som realiseres af udvalgte leksikografiske data fra nøje udvalgte leksikografiske datakilder vises i tabel 1, 2 og 3 herunder.

Funktion	Data	Datakilder
L_2-produktion	• L_{UK} ækvivalent • L_2-eksempel • L_2-kollokationer • L_2-synonymer • L_2-udtale • Link til L_2-tekst, video eller andet	• Faste og kuraterede leksikografiske L_2-datasæt fra DSL og Språkbanken • Dynamiske og ikke-kuraterede leksikografiske L_2-datasæt fra tilknyttede eksterne datakilder

Tabel 1: L_2-produktion, data og datakilder.

Funktion	Data	Datakilder
L_2-kognition	• L_{UK} ækvivalent • L_2-definition • L_2-eksempel • Link til L_2-tekst, video, podcast eller andet	• Faste og kuraterede leksikografiske L_2-datasæt fra DSL og Språkbanken • Dynamiske og ikke-kuraterede leksikografiske L_2-datasæt fra tilknyttede eksterne datakilder

Tabel 2: L_2-kognition, data og datakilder.

Funktion	Data	Datakilder
L_2-operation	• L_2-manual • L_2-SOP • L_2-retningslinjer • L_2-proceskort	• Faste og kuraterede L_2-dokumenter fra internt dokumentarkiv. • Dynamiske og ikke-kuraterede L_2-dokumenter fra eksternt dokumentarkiv

Tabel 3: L_2-operation, data og datakilder.

De leksikografiske valg skal dog suppleres af et behovstilpasset leksikografisk dialogdesign, hvilket giver leksikografien en ny dimension. Dialogdesign har til formål at tænke en bruger ind i dialogen, og denne type dialog er afgørende for, hvordan chatbotten bliver modtaget og anvendt, som påvist af Hughes (2022). Det behovstilpassede dialogdesign søger at tilfredsstille brugerens egeninteresse og bygger på relevans og belønning og på en interesse for bruge-

rens "pains and gains", jf. Osterwalder et. al (2014). Et leksikografisk dialogdesign bygger her på behovstilpassede data kombineret med de tre dialogvalg chatbot-tilgang, chatbot-personlighed og chatbot-tone.

Chatbot-tilgangen angiver, hvorvidt der anvendes en push eller pull-tilgang, dvs. hvorvidt chatbotten aktivt skal tilbyde hjælp, eller om brugeren selv skal spørge efter hjælp.

Chatbot-personligheden konstitueres i denne artikel vha. femfaktormodellen, som beskriver de fem psykologiske dimensioner. De fem personlighedsmæssige egenskaber eller karaktertræk er "Openness, conscientiousness, extraversion, agreeableness and neuroticism", som beskrevet af Costa et al. (2003). Modellen anvender de danske betegnelser for disse fem personligheder. Graden angives på en skala fra 1-5, hvor 5 er meget høj.

Chatbot-tonen er den "tone of voice", som chatbotten anvender i sit dialogsprog, og defineres som "a quality, feeling, or attitude expressed by the words that someone uses in speaking or writing" af Britannica.com (2022). Ifølge Nielsen Norman Group (2022) anvendes typisk "*humor, formality, respectfulness and enthusiasm*", og modellen herunder anvender også disse fire chatbot-toner.

Den nødvendige kombination af disse tre dialogvalg med de definerede leksikografiske funktioner og data fra tabel 1, 2 og 3 vises herunder i tabel 4, 5 og 6.

Funktion	Tilgang	Personlighed	Tone
L_2-produktion	Push	Åbenhed 5 Udadvendthed 5 Venlighed 5 Neuroticisme 1	Humor 4 Formel stil 2 Entusiasme 5

Tabel 4: Dialogvalg i forbindelse med L_2-produktion.

I forbindelse med L_2-produktion bliver den leksikografiske behovsopfyldelse realiseret ved, at de leksikografiske data værdiberi-

ges at en aktiv chatbot, som 1) tilbyder sig aktivt til brugeren, 2) har en personlighed, som er præget af maksimal åbenhed, udadvendthed, venlighed og rolighed, og 3) at det anvendte sprog i dialogen kendetegnes ved humor og entusiasme.

Funktion	Tilgang	Personlighed	Tone
L_2-kognition	Pull	Åbenhed 5 Udadvendthed 5 Venlighed 5 Neuroticisme 1	Humor 4 Formel stil 4 Entusiasme 5

Tabel 5: Dialogvalg i forbindelse med L_2-kognition.

Ved L_2-kognition bliver den sundhedsprofessionelles behov for viden realiseret ved, at de leksikografiske data værdiberiges af en passiv chatbot, som 1) skal kaldes frem af brugeren, 2) har en personlighed, som er præget af maksimal åbenhed, udadvendthed, venlighed og rolighed, og 3) at det anvendte sprog i dialogen kendetegnes ved humor, formel stil samt entusiasme.

Funktion	Tilgang	Personlighed	Tone
L_2-operation	Pull	Åbenhed 5 Udadvendthed 5 Venlighed 5 Neuroticisme 1	Humor 2 Formel stil 5 Entusiasme 3

Tabel 6: Dialogvalg i forbindelse med L_2-operation.

Ved L_2-operation spørger den sundhedsprofessionelle om konkret hjælp til løsning af en opgave. Dette behov realiseres ved, at de leksikografiske data værdiberiges af en passiv chatbot, som 1) skal kaldes frem af brugeren, 2) har en personlighed, som er præget af maksimal åbenhed, udadvendthed, venlighed og rolighed, og 3) at det anvendte dialogsprog er karakteriseret af især formel stil, idet der her er tale om instruktioner.

Der vil nu blive redegjort for, hvordan disse teoretiske overvejelser er blevet realiseret i SKANDIBOT.

6. SKANDIBOT – en leksikografisk chatbot

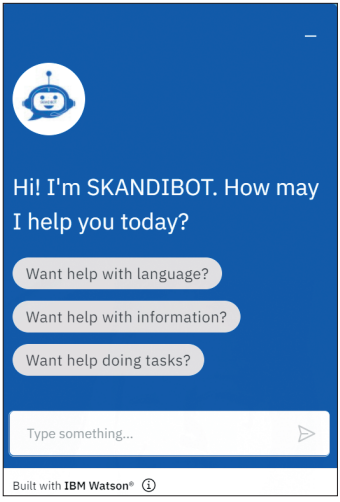
For at vælge den rigtige platform blev 30 udvalgte chatbots testet. Testen benyttede selektionskriterier som f.eks. AI-anvendelse, sværhedsgrad i programmering, publikations-muligheder (web, Messenger, WhatsApp), pris og mulighed for at udvikle dialogdesign.

Valget faldt på IBM Watson Assistant, som i meget vid udstrækning tilfredsstillede alle ønsker til SKANDIBOT. De intenderede brugere af chatbotten er sundhedsprofessionelle med anden sproglig baggrund end dansk eller svensk, og det var vigtigt, at chatbotten kan anvendes på især mobiltelefoner.

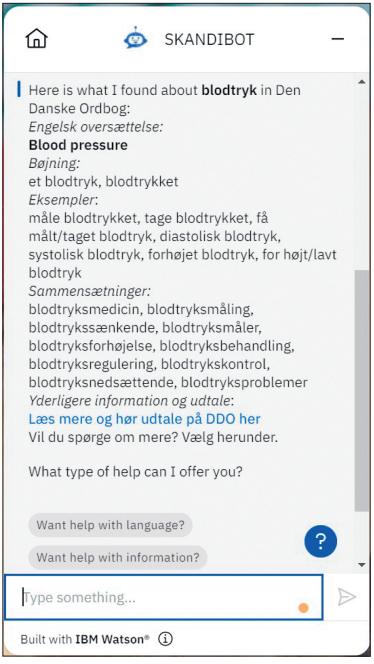
Som det fremgår af figur 4, præsenterer SKANDIBOT først sig selv og beder brugeren vælge mellem tre typer af hjælp. For eksemplets skyld antages, at brugeren først vælger "Want help with language" og derefter søger efter enten "blodtryk" eller den engelske ækvivalent "blood pressure".

SKANDIBOT returnerer derefter en chatbotartikel, som vist i figur 5. SKANDIBOT har hentet faste, kuraterede data fra DSL (2022), og hensigten i den færdige version er, at SKANDIBOT har en personlighed, der udtrykker maksimal imødekommenhed og udadvendthed og anvender en chatbot-tone, som er gennemsyret af humor og entusiasme.

Udviklingsarbejdet i SKANDIBOT-projektet er stadig i gang, og der udestår stadig meget arbejde. Men de fire små brugertests, som blev gennemført i foråret 2022 med deltagelse af to sundhedsprofessionelle med hhv. norsk og nepalesisk baggrund og to danske læringskonsulenter viste, at designet synes at være på rette vej. En testperson udtalte i en af de fire protokoller:



Figur 4: Velkomstskærm.



Figur 5: Chatbot-artikel.

> I think the design is fine. It is easy to get an overview and it is easy to find answers. I also think that the options are relevant and there is a normal, friendly greeting. I like the direct questions and direct answers.

En anden testperson udtalte sig om den operative funktion og udtrykte et klart ønske om at kunne tilgå guidelines, sundhedsprotokoller og endda kontaktoplysninger.

> The options are relevant. I understand that this must be limited to what the chatbot is able to answer. This is not Google. I think that you need some options regarding rules and protocols, and contact information.

Endelig peger en tredje testperson på et helt konkret ønske til indhold: ”Perhaps it would be a good idea to offer information to users about the healthcare sector at all levels. That is – how the healthcare sector is organized”.

Undersøgelsen er på ingen måde dækkende, men peger dog på en række områder, hvor SKANDIBOT kan forbedres. I foråret 2023 vil der blive gennemført en større undersøgelse med deltagere fra både Danmark og Sverige, hvilket formentlig vil give flere indsigter.

Der udestår stadig en række udfordringer, som der skal findes teoretiske og praktiske løsninger på. For det første er mulighederne for at designe selve chatbot-artiklen ganske begrænsede, hvilket også til dels fremgår af figur 5. For det andet er der behov for en endnu højere grad af tilpasning af det leksikografiske indhold, så det matcher brugernes behov. Endelig er der brug for at udvikle en stemmestyret tilgang, jf. f.eks. Smutny & Schreiberova (2020), som vil give de sundhedsprofessionelle mulighed for at interagere med SKANDIBOT vha. stemmestyring.

7. Konklusion og perspektiver

Denne artikel præsenterede en række teoretiske overvejelser og modeller, som kan anvendes til at udnytte samspillet mellem chatbots, dialogdesign og leksikografi.

For det første præsenterede artiklen en model for den leksikografiske mediatorrolle og diskuterede i den forbindelse den dyadiske tilgang til de to typer af leksikografiske data og til de to typer af brugere, nemlig den humane bruger og AI-brugeren.

For det andet præsenterede artiklen en model til behovs- og dialogtilpasset chatbotdesign og en model for leksikografisk behovsopfyldelse med chatbot og AI.

Endelig præsenterede artiklen en række teoretiske overvejelser, som kan anvendes til at realisere den triadiske kombination mellem chatbots, dialogdesign og leksikografi. Overvejelserne indeholdt en konkret metode til at kombinere leksikografiske funktioner, behovstilpassede leksikografiske data og nøje udvalgte leksikografiske datakilder med et behovstilpasset dialogdesign baseret på strategiske dialogvalg om chatbot-tilgang, chatbot-personlighed og chatbot-tone.

Det anføres, at kombinationen af leksikografiske data, leksikografisk interaktions- og dialogdesign og chatbots kan bidrage til at løse kendte kommunikative, kognitive og operative udfordringer for brugeren.

Projektet er pr. 1. juni 2022 foran planen, men vil i løbet af 2022 blive omkalfatret og nye partnere vil tilgå, så der er stadig meget at gøre. Nogle af de vigtigste opgaver bliver at teste chatbotten med flere brugere i både Sverige og Danmark, samt videreudvikle koblingen mellem de leksikografiske ressourcer og chatbottens AI.

Litteratur

Agerbo, Heidi (2017): Monofunctional and polyfunctional information tools with an operative function. I: *Lexicographica - International Annual for Lexicography / Internationales Jahrbuch für Lexikographie*, 361-390.

Bergenholtz, Henning & Sven Tarp (1995): *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. Amsterdam: Benjamins.

Britannica.com (2022): *The Britannica Dictionary*. <britannica.com/dictionary/tone> (januar 2022).

Cooper, Alan, Kaye Reimann & Leiben Keezer (2007): *About Face 3: The Essentials of Interaction Design*. Indianapolis, Indiana: Wiley.

Costa, Paul & Robert McCrae (2003): *NEO PI-R Manual*. PsykologiErhverv, Virum. Oversat og bearbejdet af Henrik Skovdahl Hansen og Erik Lykke Mortensen.

DSL (2022): Det Danske Sprog- og Litteraturselskab. <dsl.dk/> (januar 2022).

Fryer, Luke, David Coniam, Rollo Carpenter & Diana Lăpușneanu (2020): Bots for language learning now: Current and future directions. I: *Language Learning & Technology*, 24(2), 8-22.

Huang, Weijiao, Khe Foon Hew & Luke Fryer (2021): Chatbots for language learning – Are they really useful? A systematic review of chatbot-supported language learning. I: *Journal of Computer Assisted Language Learning*, 1-21.

Hughes, Tara (2022): No cheat code: Exploring the intersection of Empathy and Technology. <er.educause.edu/articles/2022/2/no-cheat-code-exploring-the-intersection-of-empathy-and-technology> (januar 2022).

McCarthy, John & Peter Wright (2007): *Technology as Experience*. MIT Press.

Nielsen Norman Group (2022): The Four Dimensions of Tone of Voice. <nngroup.com/articles/tone-of-voice-dimensions/> (januar 2022).

Osterwalder, Alexander, Yves Pigneur, Gregory Bernada & Alan Smith (2014): *Value Proposition Canvas: How to create products and services customers want.* Hoboken: John Wiley & Sons, Inc.

Petrović, Jasna & Mladjan Jovanović (2021): The Role of Chatbots in Foreign Language Learning: The Present Situation and the Future Outlook. I: Pap E. (eds.): *Artificial Intelligence: Theory and Applications. Studies in Computational Intelligence*, vol 973. Springer, Cham, 1-18.

Sharp, Helen, Yvonne Rogers & Jenny Preece, (2007): Interaction Design. I: *Beyond Human–Computer Interaction* (2nd ed.). John Wiley & Sons, 181-217.

Simonsen, Henrik Køhler (2002): *TeleLex - Theoretical Considerations on Corporate LSP Intranet Lexicography: Design and Development of TeleLex – an Intranet-based Lexicographic Knowledge and Communications Management System*. Ph.d.-afhandling, Århus: Handelshøjskolen i Århus.

Simonsen, Henrik Køhler (2020a): Augmented Writing: nye muligheder og nye teorier. I: Caroline Sandström, Ulla-Maija Forsberg, Charlotta af Hällström-Reijonen, Maria Lehtonen og Klass Ruppel (red.): *Nordiska studier i lexikografi* 15. Helsingfors: Nordisk förening för lexikografi, 307-315.

Simonsen, Henrik Køhler (2020b): Augmented Writing Needs Lexicography. I: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras (eds.): *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. I. Democritus University of Thrace, 509-514.

Simonsen, Henrik Køhler (2021): AI Writers in Language Learning. I: Maiga Chang, Nian-Shing Chen, Demetrios G. Sampson & Ahmed Tlili (eds.): *Proceedings IEEE 21st International Conference on Advanced Learning Technologies*, Online, 238-240.

Simonsen, Henrik Køhler & Olga Viberg (2022): Supporting Professional Second Language Learners through SKANDIBOT: A Lexicographical Design Approach. I: Maiga Chang, Nian-Shing Chen, Demetrios G. Sampson & Ahmed Tlili (eds.): *Proceedings IEEE 22nd International Conference on Advanced Learning Technologies* (under udgivelse).

Smutny, Pavel & Petra Schreiberova (2020): Chatbots for learning: A review of educational chatbots for Facebook Messenger. I: *Computers & Education* 151, 1-11.

Språkbanken (2022): SpråkbankenText. <spraakbanken.gu.se/> (januar 2022).

Tarp, Sven (2010): Lexicography in the Information Age. I: *Lexikos* 17 (AFRILEX-reeks/series 17: 2007), 170-179.

Thomas, Hephzibah (2020): Critical Literature Review on Chatbots in Education. I: *International Journal of Trend in Scientific Research and Development (IJTSRD)* Volume 4, Issue 6, September-October 2020, 786-788.

W3computing (2022): Guidelines for Dialog Design. <w3computing.com/systemsanalysis/guidelines-dialog-design-hci/> (januar 2022).

Weill, Peter & Stephanie Woerner (2018): *What's Your Digital Business Model? Six Questions to Help You Build the Next-Generation Enterprise.* Boston, Massachusetts: Harvard Business Review Press.

Wollny, Sebastian, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger & Hendrik Drachsler (2021): Are We There Yet? – A Systematic Literature Review on Chatbots in Education. I: *Frontiers in Artificial Intelligence* July 2021, Vol. 4, 1-18.

Henrik Køhler Simonsen
Ekstern lektor, PhD, MA, MBA
Copenhagen Business School
Dalgas Have 15
2000 Frederiksberg
hks.msc@cbs.dk

Andra upplagan av *Svensk ordbok*: förutsättningar och redaktionella val

Emma Sköldberg

In the article, the editor-in-chief of the second edition of *Svensk ordbok utgiven av Svenska Akademien* (SO2, 2021) gives an overall picture of, e.g., the technical conditions, financial framework and agreements with the financier which have guided the work with the edition. Furthermore, examples are provided of some of the lexicographical work initiatives that have taken place prior to the second edition and the motives behind these, as well as the priorities that have been necessary.

1. Inledning

Senvåren 2021 publicerades andra upplagan av *Svensk ordbok utgiven av Svenska Akademien* (SO), som tillsammans med *Svenska Akademiens ordlista* (SAOL) utgör Svenska Akademiens båda samtidsordböcker.

Förutsättningarna för lexikografiskt arbete har förändrats en hel del sedan den första upplagan av SO (SO1)utarbetades för att utkomma 2009. I det följande redogör jag, i egenskap av huvudredaktör, för några viktigare omständigheter som i hög grad påverkat redaktionens val och prioriteringar i det redaktionella arbetet med den andra upplagan (hädanefter SO2).

2. Förutsättningar

SO utarbetas och vidareutvecklas inom ramen för ett samarbete mellan Svenska Akademien och Göteborgs universitet. Detta samarbete regleras bl.a. av innehållet i ett samarbetsavtal från år 2010

som gäller i femtio år, dvs. till och med år 2060. Arbetet som har lett fram till SO2 har i stora drag styrts av innehållet i samarbetsavtalet, men i princip har ramarna för arbetet satts av innehållet i de löpande ansökningar om finansiering som skickats till och beviljats av Svenska Akademien. Akademien, främst dess språkkommitté, har varit en viktig diskussionspartner under arbetets gång och viktigare förändringar inför SO2 har redaktionen förankrat hos uppdragsgivaren.

SO2 bygger på en både lång och väl utvecklad lexikografisk tradition (se vidare bl.a. Malmgren 1992, 2009; Malmgren & Sköldberg 2013). SO1, och den lexikaliska databas som ordboken genereras ur, har många styrkor. Men samtidsordböckers innehåll måste ständigt uppdateras. Detta är nödvändigt för att den aktuella ordboken ska spegla det samtida språket och för att användarna av ordboken ska uppleva att den är i takt med tiden. Svagheter av mer teknisk karaktär i SO1 har också kommit i dagen efter att den tryckta SO1 publicerats i elektronisk form (se nedan).

Vidare har övriga ordboks-Sverige och -Norden förändrats mycket sedan SO1 publicerades 2009. Tidigare fanns det t.ex. flera förlag som utgav ordböcker av olika slag och i olika medier. Nu finns det istället andra aktörer vilka enbart arbetar med digitala resurser och för nya syften, exempelvis Språkbanken Text med dess utvecklande av språkteknologisk forskningsinfrastruktur. Under arbetets gång är det sådana aktörer som SO2-redaktionen har förhållit sig till eller valt att samarbeta med.

Den tryckta versionen av SO1 kompletterades förvisso av ordboksappar 2015 och som nätversion på ordboksportalen svenska.se 2017, men SO1 var i första hand en tryckt ordbok. Numera är den digitala ordboken det primära, och SO2 publiceras enbart elektroniskt, i form av appar och på webben. Detta faktum har påverkat redaktionens arbetssätt samt erbjudit nya möjligheter och utmaningar. E-ordboken är som bekant mer dynamisk och flexibel än den tryckta. Idag råder det också en ny standard vad gäller löpande

uppdateringar av ordböcker. Nordiska verk som *Den Danske Ordbog* (DDO) och *Det Norske Akademis ordbok* (NAOB) uppdateras löpande. Eftersom SO numera är en digital ordbok kommer också den att uppdateras regelbundet.

Möjligheten till löpande uppdateringar har medfört att tänkesättet ”release early, release often” präglar det lexikografiska arbetet mer än tidigare. Den tidigare eftersträvade fullständigheten och konsekvensen beträffande artiklarnas innehåll har alltmer luckrats upp. Exempelvis är det i skrivande stund (1 juli 2022) förhållandevis enkelt att finna ordboksartiklar i DDO vilka saknar någon eller några av de förväntade informationskategorierna. Till skillnad från artiklar som **soja** och **sol**[1] saknar bl.a. artikeln **sojamælk** uttalsuppgifter, såväl utskrivna som upplästa. I artikeln **sojasovs/sojasauce** ges ingen uppgift om årtal för första belägg i skrift. Adjektiven **solflimrende** och **solgylden** saknar betydelsebeskrivningar och där finns det heller inte några språkprov. I en digital ordbok är det lättare att efter hand komplettera artiklar med mer information, i den mån redaktionen önskar att göra detta.

Sedan arbetet med SO1 pågick, och ordboken publicerades i tryck, har, som redan antytts, också ordboksportalen svenska.se tillkommit. I portalen kan man söka i SAOL, SO och *Svenska Akademiens ordbok* (SAOB) samtidigt. Det är i detta sammanhang som de allra flesta av dagens ordboksanvändare möter SO. Användarstatistik visar att antalet sökningar på portalen ökar hela tiden. Tack vare svenska.se, och apparna, når SO2 också ut till en annan och bredare användargrupp än den tryckta SO1 och dess föregångare gjorde. Exempelvis torde andelen användare som är inlärare av svenska, och som befinner sig på olika kompetensnivåer, ha ökat högst betydligt.

I och med att de tre ordböckerna visas bredvid varandra i samma gränssnitt på ordboksportalen har deras respektive funktioner och perspektiv kommit att diskuteras ingående inom projektgruppen. Med tanke på att SAOL främst är normativ finns det ett värde

i att renodla SO2 till en mer deskriptiv ordbok. Även arbetsfördelningen mellan samtidsordböckerna och den historiska SAOB har dryftats under de löpande diskussionerna mellan redaktionen och uppdragsgivaren. Det är kostsamt att utarbeta ordböcker och under samtalens gång har det ifrågasatts om Akademiens ordböcker ska tillhandahålla överlappande informationskategorier, t.ex. uppslagsordens uttal och etymologi.

Det samtida språkbruket, som SO har till uppgift att spegla, förändras hela tiden. Det har dessutom hänt mycket när det gäller tillgången på korpusar sedan åren före 2009 och publiceringen av SO1. Innehållet i Språkbanken Texts konkordansverktyg Korp har t.ex. utökats kvantitativt men också breddats innehållsligt, exempelvis genom korpusar som bygger på sociala medier (Språkbanken Text/Korp 2022). Detta har haft konsekvenser för arbetet med SO2, bl.a. vid granskningen och bedömningen av befintliga språkexempel i SO1.

Slutligen bedrivs svensk språkvård i dag inte på exakt samma sätt som när SO1 publicerades. Detta märks t.ex. i de svar och av de rekommendationer som avdelningen Språkrådet inom den statliga myndigheten Institutet för språk- och folkminnen (Isof) ger i sin frågelåda. Språkrådets svar är mer resonerande än tidigare, och sådana svar kräver visst utrymme. Korta och mer kategoriska direktiv av ett mer traditionellt slag och av typen ”det är rätt och det är fel”, är mindre vanliga. Svaren på inkomna språkfrågor publiceras i Språkrådets frågelåda som också uppdateras kontinuerligt (Frågelådan i svenska 2022).

3. Redaktionella val

Arbetet med att uppdatera en ordbok med fler än 65 000 uppslagsord är en stor och krävande uppgift. Liksom SO1 och dess föregångare har SO2 styrkor men också svagheter och rena brister.

Svenska Akademien är en mycket generös finansiär av arbetet med ordböcker. Trots detta har prioriteringar i SO-redaktionens arbete varit nödvändiga. I samråd med Akademien har redaktionen därför valt att ge företräde för vissa arbetsuppgifter i förhållande till andra. Exempelvis har redaktionen utnyttjat den digitala ordbokens fördelar och kraftigt utvecklat den interna sökstrukturen i form av ett stort antal nytillagda klickbara länkar, vilka underlättar för användarna att röra sig inom ordboken. I många betydelsebeskrivningar har det centrala definitionsordet länkats vidare till den egna definitionen, såsom ordet *akrobatik* i definitionen ”person som (yrkesmässigt) ägnar sig åt akrobatik” under uppslagsordet **akrobat**. Ett annat exempel är verbet **dräpa** med definitionen ”döda genom dråp”. I just det fallet har substantivet *dråp* gjorts om till en klickbar länk i SO2. Men utan tvekan kan framtida versioner av SO utvecklas betydligt på denna punkt. Det gäller även t.ex. sökfunktioner på portalen svenska.se.

Vidare har arbetet med att renodla SO:s deskriptiva karaktär prioriterats. Detta har bl.a. lett till att informationskategorin ”Stilruta”, som lades till inför SO1, inte ingår i SO2 (se vidare om dessa rutor i bl.a. Malmgren 2009:19). Anledningen är att ett stort antal stilrutor innehåller information som inte (längre) är aktuell eller kortfattade konstateranden av slaget ”X går bra att skriva”, ”Y måste anses felaktigt” och ”Z är inte korrekt”. Konstateranden som t.ex. att *flata* och *bög* är ”fullt brukbara som neutrala beteckningar, kanske de bästa som finns” stämmer enligt SO2-redaktionens bedömning helt enkelt inte (se stilruta under **flata** i SO1; se även bl.a. Petersson & Sköldberg 2020 om hanteringen av kontroversiella ord i samband med uppdateringen av ordboken). Påpekanden som ”*Eftersom* bör inte följas av *att*. Det är alltså inte bra att skriva **han kom inte eftersom att han var sjuk*” (som återfinns i anslutning till artikeln **eftersom** i SO1), passar heller inte in i en deskriptiv ordbok som behandlar modern svenska. I just detta fall har SO2-redaktionen istället valt att uppdatera artikeln **eftersom**

genom tillägg av dels en ny konstruktionsangivelse (*eftersom att* SATS <vardagligt>), dels ett vardagligare språkprov (*affären är stängd eftersom att ägarna har semester*). Slutligen kan den stilruta som återfinns i SO1 vid lemmat **euro** nämnas. Det står:

> Att uttala *euro* som **jorå* i svensk kontext är lika omotiverat som det skulle vara att säga **Joropa* för *Europa* och **terapjot* för *terapeut*. Vårt uttal euro svarar mot tyskarnas *åj´rå* och fransmännens *örå´*.

Man kan dock konstatera att uttalet [jo´rå] är vanligt förkommande i svenskan av i dag. Uttalsvarianten är därför tillagd i SO2. Det kan också noteras att användare som söker rekommendationer om enskilda ord, ofta kan finna sådana i normativa SAOL. Mer ingående och nyanserade resonemang går att finna i Språkrådets redan nämnda frågelåda.

Att ordbokens deskriptiva funktion nu renodlas får också konsekvenser för de verbalsubstantiv som återfinns i verbartiklarna (se vidare om verbalsubstantiven i Malmgren 1992:489–490). Ett stort antal sådana substantiv har nu mönstrats ut. När det gäller denna informationskategori har SO2-redaktionen valt att informera användarna, som får antas ha mycket skiftande svenskkunskaper, om sådana avledningar som *faktiskt* används i svenskan, och inte om både sådana som *faktiskt* används och sådana som är teoretiskt *möjliga*. Avledningar av det senare slaget återfinns för övrigt i SAOB som ligger strax bredvid SO2 på svenska.se. För att konkretisera har verbalsubstantiv som är vanligt förekommande i samtida texter behållits i SO2:s verbartiklar. Det gäller exempelvis *tjatande* och *tjat* (i artikeln **tjata**), *manifesterande*, *manifestering*, *manifestation* (i artikeln **manifestera**) samt *misstänkliggörande* (i artikeln **misstänkliggöra**). Inför SO2 har däremot verbalsubstantiv som *antikiserande* (i ***antikisera***), *evaporerande* (i **evaporera**) och *bankrutterande* respektive *bankruttering* (i **bankruttera**)

mönstrats ut. Dessa substantiv används extremt sällan i moderna texter. Exempelvis förekommer de bara en handfull gånger vardera i samtliga korpusar som ingår i Korp (i dagsläget på drygt 14 miljarder tokens). Fler långsökta verbalsubstantiv av detta slag kommer för övrigt att mönstras ut inför nästa uppdatering, t.ex. *hankande* (i **hanka sig**), *illfänande* (i **illfänas**) och *piruetterande* samt *piruettering* (i **piruettera**) av det enkla skälet att dessa substantiv inte hör hemma i en ordbok som beskriver samtida svenskt språkbruk.

Att välja ut nya uppslagsord som ska ingå i den aktuella ordboken är en viktig uppgift för ordboksredaktioner. Några exempel på nyinlagda SO2-uppslagord är: *agil*, *barnfri*, *elcykel*, *filterbubbla*, *gaddning*, *hemmasittare*, *köttfri*, *lyocell*, *mem/meme*, *mående*, *nätläkare*, *ortorexi*, *plantbaserad*, *preppa*, *responsiv*, *topsa*, *viral*, *vobba*, *wow*, *xerofil*, *ålderism*, *äggdonator* och *överäta*. Vad gäller det stora antal pandemirelaterade ord, vilka började uppträda i samhället från och med första kvartalet 2020, har redaktionen bakom SO2 i många fall valt att avvakta något. Härigenom kan redaktionen se vilka av dessa ord som verkligen etableras i svenskan och vilka av de mer fackspråkliga orden som inlemmas i allmänspråket. Med tanke på planerade uppdateringar av SO2, är det förhållandevis enkelt att på sikt införliva sådana pandemirelaterade ord som, med ett lite längre tidsperspektiv, bör ingå i verket.

Vidare har det inom redaktionen funnits en önskan att SO2 ska spegla fler sorters språkbruk än vad SO1 gör och att SO2 ska ge en bild av det språk som används på 2020-talet. Detta har lett till att ett mycket stort antal otidsenliga språkprov i SO1 har uppdaterats eller ersatts av nya. De språkprov som har mönstrats ut (för att oftast ersättas av nya exempel) är av relativt olika slag, vilket framgår av exempel som *skriva ut uppsatsen på en ordbehandlare* (under lemmat **ut**), *man får 10% (i) rabatt om man betalar kontant* (under **rabatt**) och *han sade ”du” till själve landshövdingen* (under **¹själv**). Som en följd av att bl.a. tekniken har utvecklats, kontant-

hanteringen har gått ner och att det sociala avståndet mellan olika kategorier i samhället har minskat, upplevs ordboksexempel som dessa som föråldrade. Bland språkproven i SO1 finns det också en rad uttalanden som präglas av sexism eller könsstereotypier, t.ex. *hans tjej var alltid villig, håhå, vad är det för en karl som inte kan laga bilen* och *han är ett beskedligt kräk som aldrig vågar säga emot sin fru* (se artiklarna **villig**, **håhå** och **kräk** i SO1). Till exemplen av detta slag kan man även föra språkprovet *mesig tjejfotboll* (under **mesig**) som fått stor negativ uppmärksamhet inte minst i svensk press och i sociala medier (se vidare Sköldberg 2020).

Det betydande arbetet med språkexemplen har bedömts vara av största vikt eftersom exemplen är så centrala och då ett illa valt eller förlegat språkprov, ur användarsynpunkt, kan fördärva en i övrigt informativ ordboksartikel. Arbetet med språkexemplen har därför prioriteras i förhållande till exempelvis arbete med att lägga till etymologier till nyinlagda ord. Här kan det även nämnas att just arbetsuppgiften med etymologier inte omfattas av det femtioåriga samarbetsavtalet mellan Svenska Akademien och Göteborgs universitet och inte heller beaktas i beviljade ansökningar om finansiering för arbetet med SO2. Det kan också konstateras att det även i SO1 finns uppslagsord (inklusive lånord) som saknar etymologiska uppgifter, bl.a. **absent**, **community**, **podda** och **stadsjeep**. Etymologiska uppgifter kan läggas till efter hand i den digitala ordboken. Eftersom SAOB kommer att uppdateras efter fullbordandet av den första upplagan, kommer också många etymologier som inte ingår i SO att vara tillgängliga via den ordboken.

Vid en genomgripande förändring av presentationen av uttalsuppgifterna inför publiceringen av SO2 föll uppgifter om hur vissa ord (såsom *dator*) uttalas i pluralis bort. Detta är en klar försämring, men uppgifterna finns i databasen, och målet är att de åter ska vara på plats efter nästa uppdatering av SO2. Bristen berör emellertid i första hand användare av SO2-apparna. Användare av svenska.se har tillgång till de aktuella uttalsuppgifterna i SAOL.

Slutligen har SO2-redaktionen valt att vidareutveckla de uppgifter som ges om olika slags ordkombinationer i ordboken (jfr Malmgren 2009:18). Vad gäller kollokationer har vissa ordboksartiklar med många kollokationer numera en egen sektion med rubriken ”KOLLOKATIONER” innehållande just den typen av ordförbindelser. Exempel på sådana ordboksartiklar är **avtal**, **argument** och **läkemedel**. Man kan diskutera om den valda rubriken är för ogenomskinlig för ordbokens målgrupp, men ordet går trots allt att slå upp i verket.

Vissa gränsdragningsproblem kan också infinna sig när man som lexikograf ska skilja mellan olika typer av ordkombinationer m.m. SO2-redaktionen har, detta till trots, gjort bedömningen att lösningen med en egen kollokationssektion med rubrik leder till att mer ovana ordboksanvändare lättare uppmärksammar kollokationerna. Tack vare den nya sektionen i artiklar med många kollokationer kan användarna också lättare överblicka artiklarnas innehåll (se vidare Sköldberg 2022; se även Loenheim & Hult 2020:36–38 om värdet av rubriker i ordboksartiklar).

Gällande konstruktionsuppgifterna har arbetet inför SO2 präglats av ”fortsatt strävan mot ett tydligt beskrivningsspråk och användarvänlighet” (Blensenius 2019:207). I SO1 ges endast samlad konstruktionsinformation i slutet av huvudbetydelsen (se vidare om valensuppgifter i bl.a. SO1 i Malmgren & Toporowska Gronostaj 2009). Den information som ges är inte sällan bristfällig. När det gäller vissa underbetydelser saknas uppgifter helt. I en verbartikel som **regna** i SO1 återfinns endast konstruktionsuppgiften ”regna” och man kan fråga sig hur användarna ska konstruera en exempelmening som *förtroendeuppdragen formligen regnade över honom* (vilken återfinns som språkprov i en bildlig underbetydelse) utifrån den angivelsen.

I många andra fall är den samlade konstruktionsinformationen i slutet av huvudbetydelsen mycket svårtolkad på grund av sin komplexitet. Ordboksanvändarna förväntas då koppla en viss

konstruktionsuppgift till rätt delbetydelse och det är lätt att inse att detta system kan vålla problem bland användarna, i synnerhet för inlärare. Som exempel kan angivelsen under verbet **hemställa** nämnas: *hemställa* (*hos/till ngn/ngt*) (*om*) *ngt/att+V/SATS*. Angivelsen kan ses som en sammanfattning av mer än tio olika konstruktionsuppgifter. Ett annat exempel är konstruktionsangivelsen under verbet **ljuga** i SO1 som lyder: *ljuga* (*om ngn/ngt/SATS*) (*för ngn*), *ljuga* (*ihop ngt/SATS*) (*för ngn*), *ljuga* (*med ngt/SATS*), *ljuga* (*sig till ngt*). Också den utgör en sammanfattning av ett stort antal konstruktionsangivelser. I SO2-gränssnittet kan emellertid mer komplexa uppgifter expanderas så att man ser vilka de olika konstruktionerna är. Vidare har konstruktionsuppgifterna placerats högre upp i ordboksartiklarnas mikrostruktur. Uppgifterna fördelas också på huvudbetydelser och underbetydelser i ett mycket stort antal fall och dessutom har uppgifter om subjekt lagts till. Redaktionen har också arbetat vidare med att komplettera aktuella konstruktionsuppgifter med illustrativa språkprov (se vidare Blensenius 2019).

Att granska, revidera och komplettera hela verket när det gäller kollokationer och valensuppgifter är också ett mycket omfattande arbete. Redaktionen var medveten om att arbetet med kollokationerna och konstruktionsuppgifterna inte var helt genomfört när SO2 publicerades. Att SO numera är en digital ordbok, som ska uppdateras mer löpande, torde kunna leda till att uppgifterna om ordkombinationer kan bli mer kompletta i framtiden. Utifrån det tidigare nämnda tänkesättet ”release early, release often” valde redaktionen att publicera den andra upplagan av SO år 2021 istället för att vänta tills hela ordboken var uppdaterad på samtliga punkter. Jämfört med konstruktionsredovisningen i SO1, där informationen i vissa fall är ofullständig och ibland svår att förstå, framstår det nya presentationssättet som tydligare och mera lättillgängligt för användarna. Redaktionens bedömning är att det, ur användarnas perspektiv, hade varit sämre att bibehålla det gamla systemet

än att introducera det nya presentationssättet, även om det ännu inte är genomfört till punkt och pricka.

4. Slutord

Det har nu gått närmare 15 år sedan arbetet med första upplagan av SO pågick för fullt. Mycket har hänt på dessa år. Samhället har förändrats liksom språkbruket och språkvårdens karaktär – förändringar som naturligtvis måste avspeglas i den nya upplagan av SO.

Varje redaktion gör, under ledning av sin huvudredaktör, sina val och prioriteringar. Dessa baseras på de senaste forskningsrönen, på tekniska förutsättningar, ekonomiska ramar och inte minst överenskommelser och löpande diskussioner med finansiären. Alla redaktioners mål är naturligtvis att göra den aktuella ordboken, i detta fall SO, så bra som möjligt. SO2-redaktionen har velat förvalta det lexikografiska arvet från SO1 och dess föregångare och samtidigt modernisera ordbokens innehåll och utnyttja de möjligheter som erbjuds idag, inte minst tack vare digitala medier.

Just de digitala publikationsformerna har ändrat förutsättningarna för arbetet med *Svensk ordbok* i grunden. Till skillnad från de tryckta föregångarna genererar inte SO2 några försäljningsintäkter för sin huvudman. En given uppgift för ordboksredaktioner allmänt har varit att leverera underlag för säljande argument inför marknadsföringen av det nya verket – det skulle innehålla fler ord och nya informationskategorier jämfört med föregångaren. Att detta även har gällt SO framgår exempelvis då den tidigare huvudredaktören, Sven-Göran Malmgren (1992, 2009), i positiva ordalag beskriver vilka informationskategorier m.m. som successivt har lagts till i SO:s olika föregångare, från *Svensk ordbok* (SOB 1986) till SO1 (2009).

Utvecklingen av den nya upplagans digitala karaktär har stått i fokus och åtkomststrukturen är nu markant förbättrad med avancerad länkning mellan artiklarna. Kategorier som fokuseras i SAOL respektive SAOB har tonats ned. Dessa ändringar har inte genererat fler uppslagsord. Inte heller renodlingen av ordbokens deskriptiva karaktär genom utmönstring av normativa inslag är i samklang med tidigare försäljningsargument. De kan därför, mot bakgrund av den tidigare rådande synen, rent av uppfattas som försämring av ordboken. Men utgångspunkten att en ny upplagas kvalitet ska bedömas efter hur många nya ord och vilka nya informationskategorier har lagts till är förlegad och inte tillämplig på digitala verk. Jakten på nya uppslagsord m.m. kan lätt ta fokus från det viktiga arbetet med att se över befintliga artiklar och anpassa innehållet i dem så att det speglar det samhälle vi har idag. Strykningar av olika slag, såsom uteslutningen av otidsenliga språkprov och utmönstring av normativa informationskategorier som inte längre är i takt med språkvårdens direktiv eller det allmänna språkbruket, är en viktig del av redaktionens kvalitetsarbete med den nya upplagan. För att en ordboks innehåll ska vara i takt med tiden kan det – åtminstone periodvis – vara viktigare att redaktionen uppdaterar befintliga artiklar än att den fokuserar på att författa nya.

Avslutningsvis konstaterar jag att SO2 självfallet kommer att utvecklas och att digitala medier erbjuder goda möjligheter att publicera förbättrade versioner av SO. Den nuvarande SO-redaktionen noterar, liksom tidigare redaktioner, de synpunkter som användare av ordboken lämnar. All slags feedback från användarna är viktig för arbetet inför nästa uppdatering.

Litteratur

Ordböcker

DDO = *Den Danske Ordbog*. <ordnet.dk/ddo> (juli 2022).
NAOB = *Det Norske Akademis ordbok*. <naob.no/> (juli 2022).
SAOB = *Svenska Akademiens ordbok* (1898–). Lund.
SAOL = *Svenska Akademiens ordlista*. 14 uppl. 2015. <svenska.se/saol> (juli 2022).
SO1 (2009) = *Svensk ordbok utgiven av Svenska Akademien*. Första upplagan. Stockholm: Norstedts i distribution.
SO2 (2021) = *Svensk ordbok utgiven av Svenska Akademien*. Andra upplagan. Tillgänglig som appar för iOS och Android och via <svenska.se/so> (juli 2022).
SOB (1986) = *Svensk ordbok*. Stockholm: Esselte studium.

Övrig litteratur

Blensenius, Kristian (2019): Revision av konstruktionsuppgifter i *Svensk ordbok utgiven av Svenska Akademien*. I: *LexicoNordica* 26, 203–223.
Frågelådan i svenska (2022): Institutet för språk och folkminnen (Isof). <frageladan.isof.se/> (juli 2022).
Loenheim, Lisa & Ann-Kristin Hult (2020): *Framtidens Lexin: Forskningsöversikt*. (Rapporter från Språkrådet 12.) Stockholm: Institutet för språk och folkminnen.
Malmgren, Sven-Göran (1992): From Svensk ordbok ('A Dictionary of Swedish') to Nationalencyklopediens ordbok (The Dictionary of the National Encyklopedia). I: Hannu Tommola, Krista Varantola, Tarja Salmi-Tolonen & Jurgen Schopp (eds.): *Proceedings of the 5th EURALEX International Congress.* Tampere: Tampereen Yliopisto, 485–491.

Malmgren, Sven-Göran (2009): Från *Nationalencyklopedins ordbok* (1995–96) till *Svensk ordbok utgiven av Svenska Akademien* (2009). Med tillbaka- och sidoblickar. I: *LEDA-Nyt* 47, 14–20.

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

Malmgren, Sven-Göran & Maria Toporowska Gronostaj (2009): Valensbeskrivning i svenska ordböcker – och några andra. I: *LexicoNordica* 16, 181–196.

Petersson, Stellan & Emma Sköldberg (2020): To discriminate between discrimination and inclusion: a lexicographer's dilemma. I: Zoe Gavriilidou, Maria Mitsiaki & Asimakis Fliatouras (eds.): *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion.* Vol. I. Alexandroupolis, 381–386.

Sköldberg, Emma (2020): En mer inkluderande lexikografi – bortom frågan om kvinnor och män. I: Janne Bondi Johannessen & Kristin Hagen (red.): *Leksikografi og korpus. En hyllest til Ruth Vatvedt Fjeld.* (Oslo Studies in Language 11(1)). Oslo, 7–30.

Sköldberg, Emma (2022): Phraseological theory, evidence in corpora and lexicographical practice. On collocations in a monolingual dictionary of Swedish. I: Kristian Blensenius (ed.): *Valency and constructions. Perspectives on combining words.* (Meijerbergs arkiv för svensk ordforskning 46.) Göteborg, 155–182.

Språkbanken Text/Korp. <spraakbanken.gu.se/korp/> (juli 2022).

svenska.se = Svenska Akademiens ordboksportal. <svenska.se/> (juli 2022).

Emma Sköldberg
professor
Institutionen för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
Box 200
SE-405 30 Göteborg
emma.skoeldberg@svenska.gu.se

Dannelsen af en tosproglig ordbog med hjælp af sprogteknologiske metoder

Þórdís Úlfarsdóttir & Steinþór Steingrímsson

The article discusses a new approach in Icelandic bilingual lexicography where English as a target language (TL) is produced by using technological means. In order to generate the English equivalents, various TLs from online dictionaries are used as pivot languages, among other methods. Icelandic phrases and examples are translated by translation tools. Thereafter, lexicographers do necessary post-processing of the TL, monitor the quality of the work and finalise each article.

1. Indledning

Árni Magnússon-instituttet for islandske studier (AMI) har i årenes løb udgivet ordbøger eftersom leksikografi falder ind under instituttets faste opgaver. Gennem instituttets webside er der adgang til adskillige digitale ordbøger: den monolingvale *Íslensk nútímamálsorðabók* (Ordbog over moderne islandsk); de tosprogede ordbøger ISLEX (Úlfarsdóttir 2013, 2014), hvor kildesproget er islandsk og målsprogene er dansk, norsk, svensk, færøsk og finsk, som er samlet i én og samme database; samt ordbogen LEXIA mellem islandsk og henholdsvis fransk og tysk. Til trods for denne omfattende udgivelse af digitale ordbøger er der dog en åbenlys mangel, da der ikke findes en ordbog mellem islandsk og engelsk blandt instituttets udgivelser. På det islandske ordbogsmarked findes der selvfølgelig ældre ordbøger mellem islandsk og engelsk, men de imødekommer ikke længere nutidens krav om tilgængelighed, størrelse og løbende opdateringer. Man besluttede derfor at starte et nyt projekt, en islandsk-engelsk ordbog, og samtidig gøre et forsøg med at anvende nye metoder som man ikke har brugt tidligere på AMI.

I de seneste år og årtier har man i Island, ligesom i andre lande, oplevet en hastig udvikling inden for sprogteknologien, og AMI tager aktivt del deri. Forudsætningerne er at der foreligger gode sproglige ressourcer, og ordbogsmateriale er i vidt omfang blevet anvendt i sprogteknologi inden for forskellige sprog. Der er stadig større sammenfald mellem sprogteknologi og leksikografi, og i den forbindelse kan man bl.a. pege på den europæiske platform ELEXIS (European Lexicographic Infrastructure).

I takt med dette opstod den idé at anvende sprogteknologi ved udarbejdelsen af en ny islandsk-engelsk ordbog og hente ressourcerne blandt de sproglige data som AMI har opbygget i årenes løb, nemlig en ordbogsbase med 54.000 opslagsord som danner grundlag for de ovenfornævnte ordbøger. I denne base forefindes kildesproget, islandsk, med fuld ordbogsbeskrivelse, dvs. ordforklaringer, sprogbrugseksempler, faste udtryk osv. Projektet var desuden et interessant leksikografisk eksperiment, da de traditionelle metoder, som bekendt, er meget tidskrævende.

I denne artikel beskrives det islandsk-engelske eksperiment, både den tekniske side og selve ordbogsarbejdet. Artiklens første del omhandler de tekniske forudsætninger der ligger bag projektet, som for det første går ud på at danne ordpar mellem islandsk og engelsk (generering af engelske ækvivalenter), og for det andet at oversætte faste udtryk og sprogbrugseksempler med oversættelsesmaskiner. Herefter importerer hovedredaktøren de engelske ækvivalenter i en færdig islandsk kildesprogsdatabase. I sidste halvdel af artiklen drøftes den efterfølgende proces hvor redaktørerne tager over og foretager den endelige udvælgelse af engelske ækvivalenter og færdigredigerer artiklerne.

Artiklen er inddelt i følgende kapitler: I kapitel 2 gøres der rede for dannelsen af en ordliste mellem islandsk og engelsk, og kapitel 3 handler om maskinoversættelser af sprogbrugseksempler og faste udtryk. Kapitel 4 behandler dels det redaktionelle ordbogsarbejde efter at materialet med ordlisterne foreligger, dels projektets

vigtigste leksikografiske udfordringer. Kapitel 5 er en evaluering af dannelsen af en tosproget ordbog efter denne metode, og kapitel 6 indeholder konklusion.

2. Dannelsen af ordlister: metoder og data

Der blev udarbejdet en stor islandsk-engelsk ordliste som en del af Handlingsplan for sprogteknologi for islandsk (Nikulásdóttir, Guðnason & Steingrímsson 2017). Handlingsplanen strækker sig over årene 2018-2022. Handlingsplanen fremhæver behovet for at opbygge en infrastruktur for islandsk sprogteknologi, bl.a. ved dannelsen af forskellige sproglige databaser. Herunder falder den islandsk-engelske ordliste som giver muligheder for at blive brugt i projekter tilknyttet maskinoversættelser og ved opbygning af gode databaser til træning af oversættelsesmaskiner, f.eks. for at forbedre udvælgelsen af de sætningspar der bruges til træningen. Ordlisten kommer endvidere til gavn i software der bruges til at hente oplysninger i flersprogede databaser (Cross-language information retrieval) eller til sentimentanalyser. Den anvendte islandsk-engelske ordliste blev udarbejdet på AMI i 2021. Man brugte automatiske metoder til at generere en liste med forslag til ordpar, som siden blev gennemgået manuelt af en medarbejder for at godkende ordparrene som skulle opfylde én bestemt betingelse: at ordene kan anses som ækvivalenter i en bestemt kontekst.

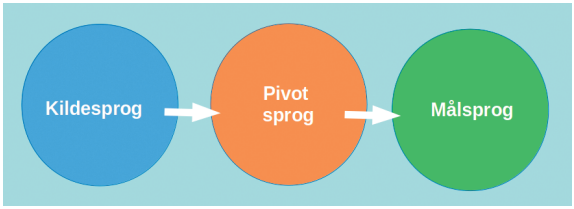
Denne ordliste blev grundlaget for de engelske ækvivalenter i den islandsk-engelske ordbog. Ved generering af ordlisten blev der anvendt fire forskellige metoder som bliver beskrevet i de efterfølgende afsnit: brug af pivot-sprog, parallelkorpusser og to slags maskinoversættelser.

2.1. Pivot-sprog

Den første metode til dannelsen af en islandsk-engelsk ordliste bestod i brugen af såkaldte pivot-sprog, som indebærer at de engelske ækvivalenter hentes igennem et tredje sprog som mellemled. Pivotsproget danner dermed en sproglig bro. Wikipedias definition af *pivot language* er følgende:

> A pivot language, sometimes also called a bridge language, is an artificial or natural language used as an intermediary language for translation between many different languages – to translate between any pair of languages A and B, one translates A to the pivot language P, then from P to B.

Da der var tale om engelsk som målsprog, var det nemmere end ellers, eftersom engelsk er et verdensomfattende sprog, og der foreligger en stor mængde sproglige data som knytter engelsk til andre sprog. AMI har adgang til sproglige data for forskellige målsprog som er blevet til gennem arbejdet med de tosprogede ordbøger. Målsprogene er de seks målsprog i ISLEX foruden de to målsprog i LEXIA. De fleste af disse sprog kunne anvendes ved genereringen af det engelske ordforråd.



Figur 1: Processen fra kildesprog (islandsk) via pivotsprog (nordiske sprog m.m.) til målsprog (engelsk).

At bruge et pivot-sprog er ikke en ny metode, men den er tilsyneladende ikke blevet anvendt inden for leksikografien i udstrakt grad. En af grundene til dette er selvfølgelig at metoden ikke var realiserbar indtil ca. år 2000 hvor ordbøger for alvor blev digitale. Verdens mest anvendte sprog, engelsk, rådede allerede over et udvalg af bilingvale ordbøger, og behovet var derfor evt. ikke så presserende. For sprog som ikke råder over så mange sproglige data, har pivot-metoden derimod været gavnlig, se bl.a. Varga & Yokoyama (2009) som brugte metoden mellem japansk og ungarsk med engelsk som pivot-sprog, og Aker et al. (2014) mellem engelsk og tysk, med fransk som pivot-sprog; se desuden Gamallo & Campos (2010).

I dette projekt anvendte man de bilingvale ordbøger med islandsk som kildesprog, og målsproget blev brugt som pivot-sprog. Siden blev ordbøger hvor pivot-sproget var kildesprog og engelsk var målsprog, parret sammen med de islandske opslagsord. På den måde blev det muligt at generere ækvivalenter til en stor del af kildesprogets opslagsord (islandsk). Her har man dog det problem at når det islandske ord, eller ordet på pivot-sproget, har mere end én betydning, får man forkerte ækvivalenter sammen med de korrekte. Denne metode alene bliver derfor aldrig særlig præcis. De brugte ordbøger var de islandsk-skandinaviske ordbøger i ISLEX samt de islandsk-fransk/tyske ordbøger i LEXIA. De ordbøger der blev brugt til at oversætte fra pivot-sprogene til engelsk, var *Apertium* (norsk/svensk/finsk/fransk/engelsk) og *dict.cc* (norsk/svensk/finsk/fransk/tysk/engelsk). *Apertium* er en open-source platform for maskinoversættelse. Den indeholder sproglige data, herunder ordbøger for et stort antal sprog. *Dict.cc* er en onlineordbog der er opbygget ved frivillig indsats. Den indeholder et stort antal sprogpar. Disse ordbøger er ikke blevet udarbejdet af professionelle leksikografer.

	Apertium		dict.cc	
	præcision	antal par	præcision	antal par
norsk	53 %	15.261	74 %	31.213
svensk	64 %	34.915	76 %	26.622
finsk	43 %	214.659	75 %	19.304
fransk	63 %	20.865	64 %	39.590
tysk			54 %	137.970

Tabel 1: Tabellen viser størrelsen og vurderet præcision af de lister med forslag til ordpar med engelsk der blev genereret igennem hvert sprog og hver ordbog.

Tabel 1 viser det antal forslag som hver metode gav og præcisionen, evalueret efter 500 tilfældigt valgte ordpar fra hver liste. Da det er tidskrævende at gennemgå listerne, ønsker man at projektets medarbejdere præsenteres for et materiale der allerede har en høj præcision. Ved kun at vælge ordpar der blev genereret med mange forskellige gennemgange af flere forskellige ordbogsresurser, var det muligt at opnå større præcision, men samtidig blev forslagene selvfølgelig færre. Ved f.eks. kun at vælge forslag som blev opnået ved at gennemgå svensk og fransk, fik man en liste med 11.274 forslag som blev evalueret til at være 97 % egnet. Andre kombinationer gav mindre præcision, men der blev dog fundet nogle kombinationer med over 90 % præcision. Alle forslag der blev genereret, blev gennemgået manuelt og accepteret eller afvist.

2.2. Oversættelsesmaskiner

Den anden metode til at fremkalde engelske ækvivalenter var at anvende oversættelsesmaskiner direkte på lemmaerne i ISLEX og oversætte ordene til engelsk. Alle de islandske opslagsord blev samlet til en liste som blev oversat til engelsk ved hjælp af to oversættelsesmaskiner som er tilgængelige på nettet, Google Translate

og Microsoft Translator. En gennemgang af 500 tilfældigt valgte ordpar fra hver oversættelsesmaskine gav 59 % præcision for Google Translate og 60 % præcision for Microsoft Translator.

2.3. Ord på pivot-sprog oversat til engelsk i oversættelsesmaskiner

Den tredje metode indebærer at man laver en liste med samtlige ækvivalenter af de islandske opslagsord i ISLEX og LEXIA-ordbøgerne, som beskrevet i afsnit 2.1, men i stedet for at slå pivotordene op i andre ordbøger oversætter vi dem med oversættelsesmaskiner. Her brugte vi fire forskellige oversættelsesmaskiner, Google Translate og Microsoft Translator (MS) ligesom før, men desuden en oversættelselsmodel fra OPUS-MT (Tiedeman & Thottingal 2020) samt M2M modellen (Fan et al. 2021). Microsoft Translator gav de bedste resultater, over 60 % præcision for alle pivotsprog, M2M gav derimod de ringeste resultater. Se tabel 2.

	Opus	**M2M**	**Google**	**MS**	**Samlet antal**
dansk	52 %		59 %	63 %	80.074
norsk			59 %	61 %	66.129
svensk	56 %	32 %	65 %	65 %	69.884
finsk	53 %	27 %	66 %	62 %	62.876
fransk	56 %	35 %	67 %	71 %	45.533

Tabel 2: Proportion af brugbare ordpar i 500 tilfældigt valgte par for hver oversættelsesmaskine og pivot-sprog.

2.4. Parallelkorpusser

Den fjerde metode til dannelsen af en islandsk-engelsk ordliste var at fange modsvarende ord i sætningspar i parallelkorpusser. Parallelkorpusser er tekster på to sprog, hvor den ene tekst er en

oversættelse af den anden. Teksterne er delt op i sætninger, og modsvarende sætninger sidestilles. Ved at undersøge et stort antal sætningspar er det muligt at finde modsvarende ord i sætningerne. På samme måde er det muligt at anvende tekster hvor der ikke er tale om oversættelser, men om tekster der forholdsvis specifikt handler om samme emne (comparable corpora), men her må man bruge andre metoder for at parre sætningerne, og der behøves mere materiale for at få et acceptabelt resultat (se f.eks. Steingrímsson et al. 2021).

For at opnå størst mulig præcision ved at finde modsvarende ord i sætningsparrene kører vi data igennem fem forskellige automatiske ordaligneringsværktøjer og vælger derefter de ordpar som de fleste af værktøjerne er enige om er de rigtige, og smider de resterende ud. Når man på den måde har dannet ordpar i alle sætninger i hver database, regner man points ud for hvert ordpar. Pointene bliver udregnet ud fra hvor tit ordparringsværktøjet danner ordpar proportionalt med hvor tit ordene forekommer i samtlige sætninger. Parrene bliver så accepteret, eller de bliver forkastet på grundlag af om scoren opnår et vist minimum som man finder frem til ved at gennemgå en lille del af forslagene og undersøge kvaliteten af ordparrene med visse mellemrum.

Der bliver anvendt seks tekstkorpusser, af tre forskellige slags: et parallelkorpus, ParIce (Barkarson & Steingrímsson 2019), tre sammenlignelige sprogkorpusser (comparable corpora) som allerede er blevet inddelt i sætningspar, WikiMatrix (Schwenk et al. 2021) og Paracrawl 7,1 og Paracrawl 8 (Bañón et al. 2020). Til slut bruges der to syntetiske korpusser (synthetic corpora) der bliver dannet via en oversættelsesmaskine der oversatte forskellige nyhedstekster henholdsvis fra islandsk til engelsk og fra engelsk til islandsk (Símonarson et al. 2020). Denne proces resulterede i sætningspar fra alle seks korpusser, hvor sætninger forfattet af personer blev oversat af en maskine, jf. tabel 3.

		Tillidsscore hvor over 50% af parrene er brugbare		
Korpus	**Samlet antal par**	**Godkendelse i %**	**Antal par**	**Forventet antal brugbare**
ParIce	346.723	51,6 %	45.646	25.553
Paracrawl 7.1	107.959	59,6 %	70.281	41.887
Paracrawl 8	342.444	62,6 %	93.850	58.750
WikiMatrix	15.781	77,2 %	6.944	5.360
Syntetisk data is-en	191.934	67,2 %	13.215	8.880
Syntetisk data en-is	229.661	60,2 %	132.381	79.693

Tabel 3: Antal ordpar, forhold mellem brugbare par og anslået antal brugbare par i forskellige korpusser.

2.5. Analyse af resultaterne fra de automatiske metoder

Der blev taget stikprøver fra resultaterne af samtlige automatiske metoder for at vurdere præcisionen af dem, som det fremgår af tabellerne foroven. For at reducere arbejdsindsatsen ved gennemgang af oversættelsesforslagene anvendtes kun de lister hvor formodet præcision var meget stor. Tilbage stod der et stort antal ordpar, og for at udnytte materialet mest muligt delte vi listen med forslag i to kategorier, på den ene side de forslag som blev genereret i arbejdet med korpusserne, og de forslag som blev genereret med maskinoversættelse eller ordbogsopslag gennem pivot-sprog på den anden side. Der blev siden udarbejdet en ny liste som kun indeholdt forslag som forekom i begge kategorier. På den måde fik man en ny liste med knap 30.000 par som blev anslået til 93,2 % præcision, se tabel 4.

		Også opnået med ordbøger igennem pivot-sprog eller maskinoversættelse		
Korpus	**Samlet antal par**	**Godkendelse i %**	**Antal par**	**Forventet antal brugbare**
ParIce	45.646	90,4 %	3.713	3.356
Paracrawl 7.1	70.281	95,8 %	18.836	18.045
Paracrawl 8	93.850	96,2 %	16.522	15.894
WikiMatrix	6.944	97,4 %	3.343	3.256
Syntetisk data is-en	13.215	97,3 %	4.986	4.851
Syntetisk data en-is	132.381	94,4 %	19.423	18.335

Tabel 4: Antal ordpar, forholdet mellem brugbare par og anslået antal brugbare par i forskellige korpusser, som også blev genereret med andre metoder.

3. Maskinoversættelse af sprogbrugseksempler og faste udtryk

I skrivende stund er systematisk arbejde med oversættelse af sprogbrugseksempler og definitioner ikke begyndt. For at fremskynde dette arbejde har vi oversat alle sprogbrugseksempler og definitioner med fire forskellige oversættelsesmaskiner, både direkte fra islandsk og igennem pivot-sprogene i ISLEX og LEXIA. Det første skridt bliver at gennemgå 1000 tilfældigt valgte islandske sætninger og undersøge om oversættelsesmaskinerne har leveret brugbare oversættelser for nogle af sætningerne. De brugbare sætninger bliver udvalgt, og det undersøges hvilke oversættelsesmaskiner og metoder der giver det bedste udfald. Resultaterne bruges siden til at inddele listerne med forslag i prioritetsorden således at de bedste oversættelser står øverst. På den måde kan redaktøren udvælge brugbare sætninger og/eller ændre de fremkomne forslag. Dette bliver gjort i håb om at det i stor udstrækning vil fremskynde det leksikografiske arbejde.

4. Leksikografisk arbejde

Efter den maskinelle bearbejdning som tidligere er beskrevet, er ordbogens redaktører i besiddelse af lange lister i alfabetisk orden som indeholder ordpar mellem islandsk og engelsk (sorteret efter det islandske ord). For at vælge de ord som skal med i ordbogen, er det næste skridt at gennemgå listerne og tynde ud i dem. I denne arbejdsgang bliver ca. 40 % af de engelske ord smidt ud, og tilbage står ca. 60 % af de engelske ækvivalentkandidater. Dette materiale bliver gjort klart til at blive indlæst i ordbogsbasen hvor de engelske ord indgår i ækvivalentfeltet.

handgerður adj		
1	HLUTAR	HAND-GERÐUR
2	BEYGING	→BEYGING
3	SKÝRING	búinn til eða unninn í höndunum
4	EN-jafn	handmade
5	DA-jafn	håndlavet

Figur 2: Fra den engelske ordbogsbase: lemmaet *handgerður* 'håndlavet' samt en engelsk ækvivalent som stammer fra ordlisten. Det danske ord under det engelske stammer fra ISLEX.

4.1. De islandsk-engelske ordpar

I bedste fald optræder de engelske ækvivalenter umiddelbart, oftest 1-4 ækvivalenter for hver betydning i en ordbogsartikel. I nogle tilfælde resulterer ordlisten dog i et stort antal ordpar. Det gælder bl.a. i de tilfælde hvor et ord har mange betydninger (polysemi, f.eks. *sterkur* 'stærk; robust' og *jörð* 'jord; landejendom').

Den oprindelige islandsk-engelske ordliste havde en mere omfattende rolle end den at danne det engelske målsprog i en tosproget ordbog. Til dette formål var listerne med ord alt for lange, og af den grund måtte de sorteres manuelt. Dette var enkelt når det drejede sig om rene fejloversættelser (som forekom nogle gange),

men i andre tilfælde var resultatet en vifte af ordpar hvor mange af kandidaterne var gode, jf. tabel 5.

ægilegur	
islandsk	*engelsk*
ægilegur	formidable fearsome atrocious gruesome terrible abominable horrific terrific horrid awful horrible appalling horrendous frightful dreadful

endrum og eins	
islandsk	*engelsk*
endrum og eins	at times every so often from time to time now and then occasionally once in a while sometimes

Tabel 5: To lemmaer med for mange engelske ækvivalenter som kræver en nøjagtig leksikografisk analyse og bearbejdning.

Som det fremgår af tabel 5, kan der blandt ordparrene opstå flere engelske ækvivalenter til det samme islandske ord, og disse ordlister kræver omhyggelig leksikografisk analyse. Ækvivalenterne er gerne (nær)synonymer i engelsk, og hvis de er flere end hvad der anses som passende i ordbogen, må der skæres ned på antallet. I andre tilfælde er der tale om islandske homonymer (f.eks. *kanna* subst. 'kande' og vb. 'undersøge', *ferja* subst. 'færge' og vb. 'færge') eller polysemer (f.eks. *slanga* 'slange; haveslange'), og så er det nødvendigt at indsætte ækvivalenterne i den rigtige ordbogsartikel eller i det rigtige nummererede afsnit i ordbogsartiklen.

En positiv ting var at når det drejede sig om tekniske begreber (termer), indeholdt listen tit et almindeligt engelsk ord samt ter-

men. Som eksempel kan nævnes de medicinske termer *rauðkorn* som fik ækvivalenterne *red blood cell* og *erythrocyte*; *þíamín* der fik ækvivalenterne *vitamin B1* og *thiamine*; *lærbein* der fik ækvivalenterne *thigh bone* og *femur*. Det redaktionelle arbejde går ud på at sætte en etiket på de tekniske begreber. Desuden skal der indsættes stilistiske markører på de engelske ækvivalenter hvor dette er nødvendigt, f.eks. *informal, vulgar, literary* og *dated.*

4.2. Britisk og amerikansk engelsk

Allerede i begyndelsen blev det besluttet at ordbogen skulle omfatte både britisk og amerikansk engelsk. Andre varianter af engelsk (f.eks. i Canada, Sydafrika og Australien) blev ikke taget i betragtning. Der er tit en betydelig forskel på britisk og amerikansk engelsk, f.eks. forskellig ortografi eller ordbrug. Eksempler på ord der har forskellig stavemåde er *colour* (Br), *color* (Am); *licence* (Br), *license* (Am); *jewellery* (Br), *jewelry* (Am). Eksempler på at der bruges forskellige ord er *pavement* (Br), *sidewalk* (Am); *anticlockwise* (Br), *counterclockwise* (Am); *(car)boot* (Br), *trunk* (Am).

I ordbogen bruges britisk som udgangspunkt mens amerikanske ord markeres specielt af redaktørerne.

4.3. Store verber

De store verber kan volde problemer eftersom et islandsk verbum tit får et stort antal engelske ækvivalenter. Det kommer ikke som en overraskelse da store verber foruden at være polyseme også forekommer i mange faste ordforbindelser i meget forskellige betydninger.[1]

1 Dette er sket i mange pivot-projekter selv om det drejer sig om ubeslægtede sprog, jf. Varga & Yokoyama 2009:869 (projekt som sammenkobler japansk og ungarsk).

Islandsk lemma	Engelsk ækvivalent
gefa	give yield grant give in donate accord confer allow quit sign up impart give up

Tabel 6: Verbet *gefa* 'give' samt nogle engelske ækvivalenter der blev genereret.

Tabel 6 viser det islandske verbum *gefa* 'give' og nogle ækvivalenter som er knyttet til det i listen med ordpar, men de engelske ord blev en del flere end vist her. Det er klart at de lange lister med forslag til ækvivalenter ikke er velegnede når man redigerer de store verber da brugen af dem i høj grad indgår i forskellige faste udtryk (f.eks. *gefa eftir, gefa upp, gefa út* osv.), noget som listerne med ækvivalenter ikke fanger så godt. For disse dele af ordbogen er det sandsynligvis bedre at anvende maskinoversættelser. Det må dog fremhæves at hvad angår "almindelige verber", dvs. langt de fleste verber på nær de ca. 60 største, er dette ikke noget problem. Det er især når verbet indgår i mange fraser at metoden med listerne ikke virker optimalt.

4.4. Ingen kandidater til ækvivalenter

Ved den manuelle redigering af ordbogsartiklerne må redaktørerne udvælge ækvivalenter fra de ordlister der er blevet indlæst i ordbogsbasen, og i nogle tilfælde må der også tilføjes ækvivalenter, da der somme tider behøves flere ord end dem som indlæses auto-

matisk. Desuden får nogle ord ingen ækvivalenter fra listerne, og de må derfor bearbejdes manuelt.

Ordbogen indeholder 54.000 opslagsord og 82 % af ordene fik én eller flere engelske ækvivalenter i den tekniske bearbejdning af de islandsk-engelske ordpar. Derimod fik 9.700 ord (18 %) ingen ækvivalenter. Dette kan skyldes flere ting:

1. Lemmaet er f.eks. en variant og indeholder kun en henvisning/et internt link til en anden ordbogsartikel (ca. 1.000 ord eller 2 % af ordforrådet).
2. Ordet anvendes kun i faste udtryk (ordforbindelser) og har derfor ikke en egentlig ækvivalent (ca. 900 ord eller 1,6 % af ordforrådet). Eksempel: *byssubrenndur* forekommer kun i det faste udtryk *hlaupa eins og byssubrenndur* 'løbe alt hvad remmer og tøj kan holde'.
3. Der eksisterer ikke en engelsk ækvivalent f.eks. for ordene *mannaferðir* 'menneskelig færden', *mófugl* 'mindre fugl af flere arter (fx hjejle) hvis biotop er græsbevoksede heder, moseområder e.l.'.
4. Ordet har en engelsk ækvivalent, men til trods for dette kom den ikke med på listen efter de anvendte metoder, formodentlig fordi de brugte ressourcer var mangelfulde.

Når et lemma ikke har en engelsk ækvivalent, må dets betydning forklares manuelt. Hvis det drejer sig om et lemma med kun et fast udtryk, anvendes maskinoversættelse for udtrykket, men i skrivende stund er man ikke kommet så langt. Faktum er at det stadig ikke er klart hvor mange af de 9.700 ord der skal bearbejdes manuelt, noget som vil vise sig på et senere stadie.

5. Evaluering af resultaterne

Selv om ordparrene alle er blevet evalueret som acceptable på første trin, er det ikke ensbetydende med at de alle kan anvendes i en ordbog. Nogle engelske ækvivalenter er meget specifikke, andre tilhører ældre sprog, er forældede eller sjældne. Desuden plejer man ikke at give mange ækvivalenter til samme betydning i en ordbog, og derfor må man kun vælge dem som passer bedst. Listerne med engelske ækvivalenter bearbejdes i to omgange. I første omgang gennemgår ordbogens redaktør listerne og skærer ned. På næste trin gennemgår medarbejderne forslagene til ækvivalenter og foretager det endelige valg af de engelske enheder der medtages i ordbogen.

Vi har undersøgt resultaterne af den første udvælgelse for tre bogstaver som er færdige, dvs. ord som begynder på bogstaverne L, M og N. Listen med forslag fra den oprindelige ordliste indeholdt 20.817 par for disse tre bogstaver. Af dem kom godt en tredjedel, dvs. 6.445 ord, ind på listen ved brug af én metode alene, mens de øvrige par kom med på listen ved brug af mere end én metode. Langt de fleste, eller over 85 % af ordparrene, blev genereret ved at bruge ordbøger via pivot-sprog, enten ved hjælp af denne metode alene eller kombineret med andre metoder.

		Maskin-oversættelse	Maskin-oversættelse via pivot-sprog	Ordbog via pivot-sprog	Korpus	**I alt**
Oprindelig liste	**Én kilde**	68	1.437	4.070	230	**6.445**
	Flere kilder	3.246	8.415	13.072	10.093	**14.372**
	I alt	3.314	9.852	17.782	10.323	**20.817**

Bearbejdet liste	**Én kilde**	68	670	2.277	80	**3.095**
	Flere kilder	2.931	6.644	8.009	5.770	**9.202**
	I alt	2.999	7.314	10.286	5.850	**12.297**
Beholdt ord i procenter	**Én kilde**	100 %	46,6 %	55,9 %	34,8 %	**48,0 %**
	Flere kilder	90,3 %	78,9 %	61,3 %	57,2 %	**64,0 %**
	I alt	90.5 %	74,2 %	57,8 %	56,7 %	**59,1 %**

Tabel 7: Antal par i den oprindelige liste med forslag og efter redaktionel gennemgang af bogstaverne L, M og N. Tabellen viser forslagenes antal og procentdel som opstår med kun én af de fire metoder, og siden med flere end én metode. Således er et forslag, som opstår både i oversættelsesmaskine og med hjælp af et korpus, talt på begge steder. Det samlede tal i den sidste kolonne er derfor ikke nødvendigvis summen af tallene i kolonnerne til venstre.

Efter en redaktionel gennemgang står 12.297 ordpar tilbage, eller knap 60 % af listen med forslag. Som forventet blev der skåret forholdsvis mere ned i de ordpar som blev genereret ved brug af én metode alene. På den anden side blev der skåret mindst ned i det ordforråd som blev genereret ved hjælp af maskinoversættelse eller maskinoversættelse mellem pivot-sprog. Forklaringen kan være at oversættelsesmaskinerne kun leverer én oversættelse af hvert ord, og når oversættelsen er korrekt, er det oftest den mest almindelige oversættelse. Derfor er sådanne ordpar også forholdsvis mere oplagte til at blive valgt i den type leksikografisk arbejde som beskrives her.

Under redaktørens gennemgang skulle der fjernes en del "støj". Hermed menes forkert stavede ord, ord med stort bogstav (foruden den korrekte skrivemåde), og forkerte ord efter redaktørens vurdering. Disse ord blev sigtet fra i første gennemgang. Desuden

må redaktørerne nogle gange tilføje de engelske ækvivalenter som de synes mangler i ordbogsartiklerne.

6. Konklusion

Projektet inddeles i nogle trin. Det første trin var at udarbejde lister med ordpar mellem islandsk og engelsk, som er en af de definerede opgaver i Handlingsplan for sprogteknologi for islandsk. For at producere listerne blev der anvendt fire forskellige metoder: brugen af pivot-sprog mellem islandsk og engelsk, to slags anvendelse af oversættelsesmaskiner, samt samkørsel af korpusser. Listerne med ordpar indeholdt langt flere islandske ord end der er opslagsord i ordbogen, og derfor blev de overflødige ord sigtet fra inden ordbogsredaktøren modtog ordlisterne. I alt blev der genereret engelske ækvivalenter for 82 % af ordforrådet.

Projektets anden fase er at filtrere listerne med ordpar for senere at bruge dem som stammen i ordbogen. Der skal skæres i ordlisterne med engelske ord indtil ca. 60 % står tilbage, som siden bliver indlæst i ordbogens database. Her kommer målsprogsredaktørerne ind i billedet, og de foretager den egentlige redaktion af ordbogsartiklerne. Dette arbejde foregår i skrivende stund.

En speciel fase i projektet er at anvende maskinoversættelse ved overførslen af sprogbrugseksempler og faste udtryk fra islandsk til engelsk. Denne fase er under udvikling, og de første forsøg giver forhåbninger om gode resultater.

Man kan gøre sig overvejelser over hvorvidt dette er en praktisk metode ved udarbejdelsen af en bilingval ordbog, og om metoden er tidsbesparende. Vi kan måle den tid der bliver anvendt og sammenligne med den tid det tog at udarbejde de bilingvale ordbøger i ISLEX, hvor målsprogsarbejdet udelukkende skete manuelt. Resultaterne viser os at denne nye metode er betragteligt hurtigere for målsprogsredaktørerne – men her tages de forudgå-

ende to gennemgange af de oprindelige islandsk-engelske ordpar dog ikke i betragtning. Projektets formål var imidlertid ikke alene at lave en islandsk-engelsk ordbog, men også at foretage en almen undersøgelse af maskinoversættelser mellem islandsk og engelsk.

Litteratur

Ordbøger

Apertium. A free, open-source machine translation platform. <apertium.org/> (marts 2022).

Dict.cc. <dict.cc/> (marts 2022).

ISLEX. Þórdís Úlfarsdóttir (hovedred.). Reykjavík: Árni Magnússon instituttet for islandske studier. <islex.dk/> (marts 2022).

Íslensk nútímamálsorðabók. Halldóra Jónsdóttir & Þórdís Úlfarsdóttir (red.). Reykjavík: Árni Magnússon instituttet for islandske studier. <islenskordabok.arnastofnun.is/> (marts 2022).

LEXIA. Þórdís Úlfarsdóttir (hovedred.). Reykjavík: Árni Magnússon instituttet for islandske studier. <lexia.arnastofnun.is/> (marts 2022).

Anden litteratur

Aker, Ahmet, Monica Paramita, Mārcis Pinnis & Robert Gaizauskus (2014): Bilingual dictionaries for all EU languages. I: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavík, Island, 483-489. <lrec-conf.org/proceedings/lrec2014/index.html>.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías,

Marek Strelec, Brian Thompson, William Waites, Dion Wiggins & Jaume Zaragoza (2020): ParaCrawl: Web-Scale Acquisition of Parallel Corpora. I: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4555-4567. <aclanthology.org/2020.acl-main.417>.

Barkarson, Starkaður & Steinþór Steingrímsson (2019): Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. I: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland, 140-145. <aclanthology.org/W19-6115>.

ELEXIS. European Lexicographic Infrastructure. <elex.is/> (marts 2022).

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli & Armand Joulin (2021): Beyond English-Centric Multilingual Machine Translation. I: *Journal of Machine Learning Research* 22(107), 1-48.

Gamallo, Pablo & José Campos (2010): Automatic Generation of Bilingual Dictionaries Using Intermediary Languages and Comparable Corpora. I: *Computational Linguistics and Intelligent Text Processing, 11th International Conference*. Iasi, Romania, 473-483.

Nikulásdóttir, Anna Björk, Jón Guðnason & Steinþór Steingrímsson (2017): *Language Technology for Icelandic 2018–2022*. Project Plan. Reykjavík: Icelandic Ministry of Education, Science and Culture.

Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong & Francisco Guzmán (2021): WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. I: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1351-1361. <aclanthology.org/2021.eacl-main.115>.

Símonarson, Haukur Barri, Vésteinn Snæbjarnarson & Vilhjálmur Þorsteinsson (2020): *En-Is Synthetic Parallel Corpus (20.09)*. CLARIN-IS <hdl.handle.net/ 20.500.12537/70>.

Steingrímsson, Steinþór, Pintu Lohar, Hrafn Loftsson & Andy Way (2021): Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. I: *Proceedings of the 14th Workshop on Building and Using Comparable Corpora*. <aclanthology.org/2021.bucc-1.3/>.

Tiedemann, Jörg & Santhosh Thottingal (2020): OPUS-MT – Building open translation services for the World. I: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal, 479-480. <aclanthology.org/2020.eamt-1.61>.

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

Úlfarsdóttir, Þórdís (2014). ISLEX – A Multilingual Web Dictionary. I: *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* Reykjavík, 2820–2825. <aclanthology.org/L14-1>.

Varga, István & Shoichi Yokoyama (2009): Bilingual dictionary generation for low-resourced language pairs. I: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, 862-870. <aclanthology.org/D09-1090>.

Þórdís Úlfarsdóttir
hovedredaktør
Árni Magnússon-instituttet for islandske studier
Laugavegur 13
IS-101 Reykjavík
thordis.ulfarsdottir@arnastofnun.is

Steinþór Steingrímsson
sprogteknolog
Árni Magnússon-instituttet for islandske studier
Laugavegur 13
IS-101 Reykjavík
steinthor.steingrimsson@arnastofnun.is

En historisk ordbogs digitale fremtid

Tarrin Wills

The Dictionary of Old Norse Prose (ONP) is an ongoing historical project covering an important medieval language and its literature. The dictionary is based on traditional methods but has amassed a wealth of resources as part of its long work. In order to stay relevant, ONP has been developed with a view to improve its editing methods, expand its user base and connect to other digital resources that have become available in more recent times. This paper presents a number of methods by which we achieve these ends.

1. Indledning

Det tager lang tid og kræver mange ressourcer at udarbejde store historiske dokumentationsordbøger. Man hører ofte om sådanne udfordringer, som for eksempel i denne opsummering af historien om *Middle English Dictionary* (MED) på projektets hjemmeside (<quod.lib.umich.edu/m/middle-english-dictionary/about>):

> The 70-year Middle English Dictionary project drew on support from (and probably exhausted the patience of) many agencies and individuals over that span, both internal and external, to an extent that can only be hinted at here.

Mange lignende tilfælde kendes fra andre ordbøger, eksempelvis *Oxford English Dictionary* (se f.eks. Mugglestone 2005:33-34). Da projekterne løber over en lang periode, er man nødt til løbende at opdatere de redaktionelle metoder når faget udvikler sig, og når der kommer nye standarder og flere eksterne ressourcer i form af nye opslagsværker, tekstudgaver o.l.

Denne artikel beskriver nogle tiltag der sigter mod at løse de udfordringer der er forbundet med langsigtede historiske ordbogsprojekter, især hvad angår tilgængelighed af ikke-redigeret materiale og redaktionsprocessens hastighed. De valgte løsninger gør brug af nye teknologier og eksterne ressourcer, og i mange tilfælde er resultaterne målbare.

2. Baggrund

Ordbog over det norrøne prosasprog (ONP) er et ordbogsprojekt ved Københavns Universitet, der blev grundlagt i 1939. ONP registrerer ordforrådet i oldnordiske prosatekster overleveret i norske og islandske håndskrifter fra ca. 1150 til slutningen af middelalderen. ONP bruger MED som model for sin redaktion, da MED på ordbogens begyndelsestidspunkt var den metodisk nyeste og bedste store historiske ordbog over et lignende gammelt sprog. I 80'erne begyndte projektet at bruge databaser til at samle og organisere dataene, og de trykte bind, der udkom 1989-2004, blev eksporteret direkte fra ONP's database.

Ifølge ONP's redaktionelle principper skal materialet afspejle originalkilden så nøjagtigt som muligt, hvilket betyder at citaterne er taget fra videnskabelige udgaver eller direkte fra håndskrifter. I de fleste tilfælde henviser ONP til både tekstudgaver og de håndskrifter som udgaverne bygger på (se f.eks. Battista og Jóhannsson 2014, Jóhannsson og Battista 2016).

I 1989 udkom ordbogens registerbind der fungerer som et selvstændigt opslagsværk over håndskriftskilder, videnskabelige tekstudgaver og relevant litteratur. Registerbindet blev efterfulgt af tre trykte bind med ordbogsartikler, som dækker alfabetet fra a- til em-. Efter 2004 blev udgivelsen af det trykte værk indstillet, og det blev besluttet at gøre ordbogsmaterialet tilgængeligt på nettet, i første omgang ved at scanne det materiale som endnu ikke var

udgivet. I 2010 blev onlineudgaven lanceret med både materiale fra de udgivne bind og andet materiale som primært bestod af lemmalisten samt tilhørende citater i form af scannede sedler fra ONP's arkivsamling. En ny onlineudgave blev lanceret i 2019, og denne giver adgang til flere materialer i ONP's database og forbinder ONP til en række eksterne ressourcer.

Det norrøne sprog og dets litteratur er vigtige for mange forskellige forskningsområder og fag. De norrøne tekster indeholder i de fleste tilfælde de tidligste belæg på en stor del af ordforrådet i de skandinaviske sprog, og teksterne er derfor meget centrale i sproghistorisk sammenhæng. Litteraturen omfatter mange betydningsfulde genrer og værker, f.eks. sagaerne og eddadigtningen. Kongesagaerne og andre tekster bruges som centrale kilder til Nordens historie i middelalderen, og de mange lovtekster er særligt interessante for lovhistorie og undersøgelser af sociale normer og værdier i middelalderen. Også Snorres Edda og den poetiske Edda er centrale kilder i vores forståelse af den førkristne religion og mytologi i Norden og andre steder. Der findes i dag mange gode oversættelser af de vigtigste værker, men historikere, religionshistorikere og andre forskere der ikke selv er sprogforskere, er stadig nødt til at kunne læse de oprindelige kilder for at kunne kontrollere deres tolkninger og for at kunne henvise til de oprindelige kilder.

Videreudviklingen af ONP har flere formål. Vi vil gøre de mange digitale ressourcer der allerede er blevet samlet, mere tilgængelige samt understøtte de forskellige forskningsprojekter der har brug for at tilgå og udnytte ordbogens oplysninger. Ordbogen kan også bruges som hjælp til læsning af norrøne tekster der ikke er blevet oversat. Vi vil gøre ordbogen mere kompatibel med nye leksikografiske metoder og med andre ressourcer, især korpuslingvistiske og leksikografiske ressourcer og digitale udgaver af norrøne tekster. Målet er at gøre redaktionsarbejdet mere effektivt, så ressourcerne kan bruges på at forbedre ordbogens indhold og gøre ordbogen lettere tilgængelig.

3. Forbedring af redaktionsprocessen

Allerede fra 1980'erne blev ONP redigeret i et databasesystem (DBase og senere Oracle), og den daværende redaktør Bent Chr. Jacobsen udviklede ONP's redigeringsprogram over mange år. Teknologien i systemet er avanceret nok til at forbinde ordbogens citater, artikler og registrene over udgaver og håndskrifter med hinanden, og alle dele kan redigeres gennem redigeringsprogrammet der kører i Windows-operativsystemet. Det tager dog mange måneder inden en ny redaktør kender systemet godt nok til at kunne skrive en ordbogsartikel, og det kræver mange års erfaring inden man behersker det fuldstændigt.

3.1. Redigering af ordbogen

Det gamle redigeringssystems form og arbejdsgange stammer fra 1980'erne hvor man var nødt til at redigere direkte i databasens tabeller. Lemmaer, definitioner og citater m.m. kan eksporteres til et kontrolformat, bogformat eller web. Så længe ordbogen har flere meget erfarne redaktører der arbejder effektivt i det gamle system, giver det ikke mening at lave det om. Men nye teknologier gør det nemmere at udvikle mere brugervenlige grænseflader, og derfor har ordbogen udviklet et nyt redigeringssystem der kører parallelt med det gamle som stadig er i brug. Det nye system bevarer de gamle strukturer, og det bruger det samme databasesystem som det gamle system.

Det nye system er integreret i ONP's website. Det giver mulighed for at redigere ordbogen direkte på websitet: Man kan tilføje definitioner, flytte citater rundt mellem forskellige afsnit og definitioner og se resultaterne med det samme og i det samme format som brugeren ser dem i. Det betyder at nye redaktører meget hurtigere kan lære at arbejde med redigeringssystemet. De to systemer kører parallelt, men der er stadig problemer med at implementere

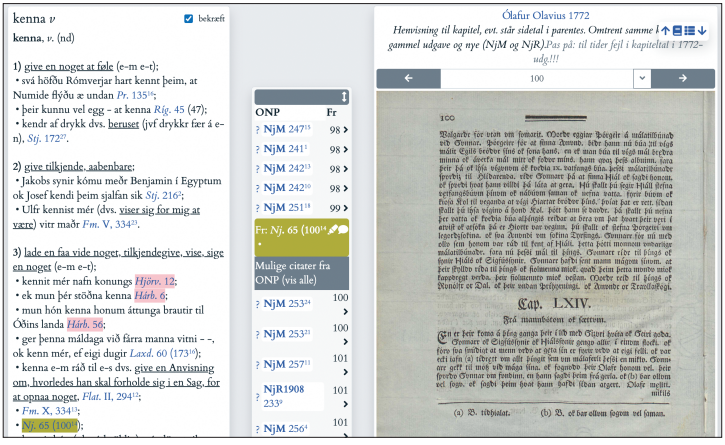
den nuværende datastruktur i to forskellige applikationer. Til gengæld er det meget nemmere i en webapplikation at oprette links til andre nyttige ressourcer for redaktionen, f.eks. til ordbøger og korpusser på nettet. I det lange løb skal strukturen laves om så den bedre afspejler de nyere digitale standarder for ordbøger (f.eks. Text Encoding Initiative (TEI) og OntoLex (Bosque-Gil & Gracia 2019)). De nye standarder bruger eksplicitte relationer mellem artiklens dele i selve data, i modsætning til ONP's datastruktur hvor de digitale relationer ofte er implicitte.

3.2. Supplering fra ældre ordbøger

En af de processer som har haft størst gavn af webapplikationens nye muligheder, er suppleringen med materiale hentet fra ældre ordbøger. ONP supplerer sine citater fra de mest centrale ældre ordbøger, især *Fritzner* og *ClV*. Processen er tidskrævende: Man tager hvert citat i den gamle ordbog og slår op i den tekstudgave som der henvises til i citatet. Den der supplerer, prøver så at finde det tilsvarende tekststykke i den udgave som ONP bruger. Hvis det lykkes, og citatet ikke allerede findes i databasen, tilføjer man det til databasen med de korrekte henvisninger til udgaven og ordet så det supplerede citat kan bruges i forbindelse med redigeringen af ordet. Gennemsnitligt er ca. 9 % af citaterne til et opslagsord i ONP kommet ind gennem denne suppleringsproces. Selvom ordbogen har adgang til flere digitale og fysiske ressourcer, vurderer vi dog at man stadig bruger ca. 20 % af den samlede redigeringstid på denne gamle måde at supplere på.

En ny digital redigeringsproces i webprogrammet gør arbejdet langt nemmere og hurtigere. Digitale udgaver af de vigtigste gamle ordbøger bliver behandlet automatisk og henvisningerne gøres klikbare. Henvisninger der ikke er relevante for ONP, f.eks. fordi de henviser til poetiske eller for unge tekster (jf. ordbogens register), fremhæves. Når man klikker på en henvisning, henter pro-

grammet citaterne fra artiklen der hører til det samme værk, samt henvisninger til citater der ligger i nærheden af den gamle ordbogs citat i den gamle udgave. Den gamle udgaves scannede side vises ved siden af så man hurtigt kan se citatets kontekst. Såfremt citatet allerede findes i artiklen, kan man markere det med et klik i et felt og dermed registrere overensstemmelsen. Findes citatet ikke i databasen, kan man klikke på et citat fra et andet ord i nærheden. Det åbner en side i den nye udgave, hvor alle citater på siden vises (se afsnit 5.1). Redaktøren kan bladre mellem siderne, finde citatet i den nye udgave og tilføje citattekst samt linjehenvisning. I tilfældet nedenfor (figur 1), fra *Fritzner*, forbinder systemet de nye oplysninger med den gamle ordbogs digitale tekst.



Figur 1: Suppleringssystemet med den gamle ordbogsartikel, citater fra ONP og billede af den udgave som den gamle ordbog henviser til.

Den nye suppleringsfunktion forbinder flere digitale ressourcer:

- gode digitale versioner af de gamle ordbøger
- et gammelt, men nyttigt digitalt register over de gamle ordbøgers litteratur
- billeder af de gamle udgaver, som nu alle er uden ophavsret, og som i de fleste tilfælde findes på nettet

- tusindvis af henvisninger fra supplerede citater i ONP til de gamle ordbøgers henvisninger, som tidligere blev indtastet manuelt
- visning i websitet, der samler alle citater der er blevet excerperet fra en side i originalværkets udgave, sammen med et billede af siden, og gør citaterne redigerbare for redaktørerne (jf. figur 3, afsnit 5.1).

Dette system gør processen mindst dobbelt så hurtig som den gamle, og det er simpelt nok til at en studentermedhjælper kan lære at bruge det. Det samler samtidig flere oplysninger om forbindelsen mellem ONP og de ældre ordbøger og gør det nemmere at finde citaterne herefter.

4. Tilgængelighed og formidling

I de trykte ordbogsbind (a- til em-, ca. 20 % af det samlede antal citater) bruges både engelsk og dansk som målsprog, men siden 2010 er langt de fleste definitioner kun blevet skrevet på dansk. I skrivende stund udgør det allerede redigerede materiale 55 % af alle citater, og heraf mangler en stor del altså engelske definitioner. Derfor er der store huller i ordbogens semantiske dækning af ordforrådet, især hvad angår den semantiske analyse med engelsk som målsprog. Trods de manglende engelske definitioner vil ONP sikre at ordbogen er så brugbar som mulig så man kan anvende den som primært udgangspunkt og ikke behøver at konsultere ældre ordbøger.

ONP bruger to metoder til at tilføje de manglende semantiske analyser midlertidigt: supplering fra ældre ordbøger på engelsk og dansk/norsk, og automatisk oversættelse udført af Google Oversæt.

4.1. Supplering af uredigerede ord

ONP's digitale ordliste indeholder detaljerede henvisninger til 36 andre historiske ordbøger, glossarer og ordlister over det norrøne sprog. I mange tilfælde er disse ordbøger blevet digitaliseret, især de vigtigste (*Fritzner* og *ClV*; se ovenfor i afsnit 3.2), hvilket åbner muligheden for at sammenkæde dem digitalt med ONP's artikler.

For at linke et ord i ONP med de ældre ordbøger bruger webprogrammet henvisninger fra databasen til at rekonstruere det tilsvarende ord i de ældre ordbøger ved at konvertere ortografien. Henvisningerne til de ældre ordbøger i ONP er imidlertid udviklet til den trykte udgave af ordbogen. De indeholder ofte kun dele af opslagsordet i den ordbog der bliver henvist til, for at vise de ortografiske forskelle mellem ONP's opslagsord og opslagsordet i den ældre ordbog. Det er derfor vanskeligt at rekonstruere opslagsordet alene ud fra disse data således at man kan søge efter ordet i en digital version af den ældre ordbog. Programmet forsøger at rekonstruere opslagsordet alligevel, hvilket lykkes i langt de fleste tilfælde, og derfor kan webprogrammet vise de relevante ordbogsartikler fra de andre ordbøger på den samme webside som ONP's artikel. I de tilfælde hvor det ikke lykkes, kan redaktørerne hente de rigtige artikler i andre digitale ordbøger og manuelt linke dem til ONP's artikel. Brugere kan derfor få adgang til indholdet i de ordbøger som der ikke kan linkes til automatisk.

Hvad angår artikler/ord i *Fritzner*, *ClV*, *Hertzb* og ordbøgerne fra *málið.is* (en fælles portal til flere islandske ordbøger), kan ONP vise digitale artikler af høj kvalitet fra disse ordbøger. Det samme system der bruges af ONP til at supplere fra disse ordbøger (se afsnit 2.2), bruges også til at lade en ekstern bruger læse de tekster som de gamle ordbøger henviser til. Ved *Lexicon Poeticum* (*LP*) viser websiden et scannet billede af ordbogens side.

For dem der helst vil læse definitioner på engelsk, har *ClV* engelsk som målsprog. Herudover trækker webprogrammet på en

række andre ressourcer til at supplere det engelske. Det henter via en Application Programming Interface (API) oplysninger fra The Skaldic Project (Skaldeprojektet) som er et udgivelsesprojekt der lemmatiserer alle ord med henvisning til ONP. En API gør det muligt at forbinde et eksternt program (en webapplikation for eksempel) til et andet program. I dette tilfælde kan man søge på et lemma og derefter hente en detaljeret konkordans af ord med det samme lemma i korpusset. I The Skaldic Project oversætter man teksterne til engelsk ord for ord så vidt det er muligt, og der findes nu oversættelser af 150.000 ord med 11.000 lemmaer. Det betyder at ONP for mange artiklers vedkommende kan hente alle citater fra The Skaldic Project's hjemmeside sammen med alle oversættelser af det enkelte lemma. Disse oversættelser supplerer ONP's danske eller manglende engelske definitioner.

Med disse eksterne ressourcer kan ONP supplere sit materiale med de manglende oplysninger om ordets betydninger og kontekst uden at brugeren er nødt til at forlade sitet. Det supplerende materiale giver oplysninger på både engelsk, dansk, norsk og islandsk. For mange brugere er engelsk dog det eneste relevante målsprog. I disse tilfælde bruger vi automatisk oversættelse.

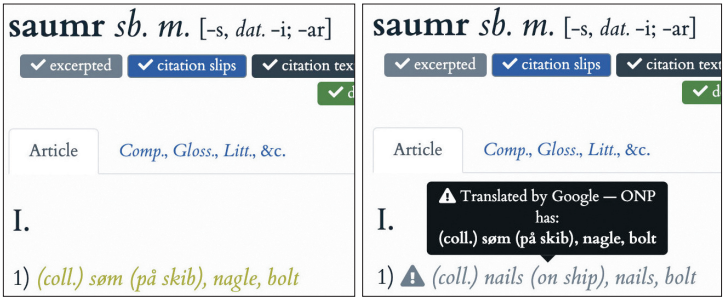
4.2. Engelsk

Ifølge Google Analytics har kun 13 % af ONP's brugere svensk, dansk eller norsk som standardsprog i deres browser. 20 % af brugerne besøger ONP fra Danmark, Norge og Sverige, og 12 % fra Island hvor de fleste sandsynligvis kan læse dansk. Det betyder at højst en tredjedel af ordbogens brugere kan antages at kunne læse dansk (se Wills, Battista & Jóhannsson 2021) selvom 60 % af det semantisk redigerede materiale kun har dansk som målsprog, ligesom flere af de eksterne ressourcer.

Som en midlertidig løsning bruger ONP Google Oversæt til at oversætte danske og norske definitioner flere steder. Oversættelsen

sker når brugeren klikker på et stykke norsk eller dansk tekst som er blevet fremhævet. Det betyder at søgemaskiner og webarkiver ikke ser den automatisk oversatte tekst, hvilket ellers kunne antyde at det er ONP's oversættelse. Oversættelsen forsynes med tydelig kildeangivelse, og brugeren kan se den oprindelige tekst (jf figur 2). I en undersøgelse (Wills 2021) vurderede vi at omkring 90 % af de brugeranmodede oversættelser var tilstrækkelige til at kunne hjælpe brugeren.

Denne funktion bruges især til ONP's definitioner på dansk, men den er ydermere blevet anvendt på de linkede eksterne ressourcer der har digital tekst på dansk og norsk, dvs. *Fritzner*, *Blöndal* og *Hertzb*. Når man henter disse artikler fra eksterne ressourcer, gør webapplikationen teksten klikbar så brugeren kan anmode om en oversættelse fra Google Oversæt.



Figur 2: Brugeranmodet automatisk oversættelse af en dansk definition.

5. Forbindelse med andre ressourcer

Når ordbogen kan forbindes med andre ressourcer, er det også muligt at bruge den til andre formål. ONP er ikke kun en lingvistisk ressource: Mange bruger ordbogen til at kunne læse tekster der ikke er blevet oversat. Andre bruger den til at undersøge sociale og kulturelle fænomener ved at slå ord op der bruges i forbin-

delse med sådanne fænomener så de kan finde forskellige citater og kilder til disse fænomener.

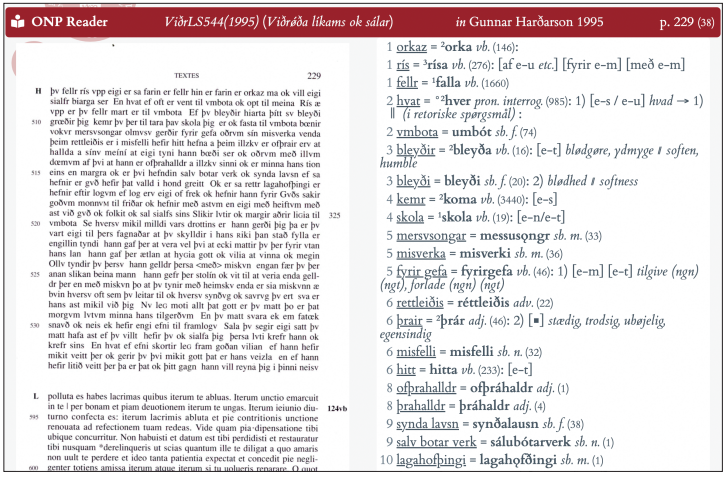
ONP har sammenkædet mange typer af data i løbet af de sidste årtier som en del af redaktionsarbejdet. Nogle af disse henvisninger refererer til traditionelle ressourcer: tidligere ordbøger, håndskrifter, udgaver og paralleller (kilder). Andre henviser til nye, bl.a. begrebsordbøger (tesaurusser) og andre ressourcer der indeholder semantiske oplysninger om ordene og korpusset.

5.1. Tovejshenvisninger til udgaver

En stor del af ONP's brugere anvender ordbogen til at slå ord op når de læser en norrøn tekst. For dem der kan lidt norrønt, men stadig har brug for hjælp med sværere ord og betydninger, er ordbogen meget nyttig. På websitets citatsider hentes mange oplysninger om citatet: bl.a. tekstudgaven, håndskriftet, citatsedlen osv. I de fleste tilfælde (99 %) er den scannede udgave uden ophavsret eller dækket af ONP's aftale med CopyDan så siden kan vises sammen med citatet.

Da databasen har henvisninger til specifikke sider og linjer i alle anvendte udgaver, og desuden har billeder af dem, kan dataene bruges den modsatte vej: Programmet kan hente en side fra en udgave og dertil alle citater der er registreret i ordbogen på den pågældende side. Denne funktion kaldes ONP Reader og er tilgængelig gennem den enkelte citatside, samt værkregistret og bibliografien (jf. figur 3).

Gennemsnitligt har ONP excerperet omkring 7 % af alle ord i hele korpusset af norrøne tekster, og 4 % af alle ord i hele korpusset er forsynet med en definition i ONP. Man kan forvente at de redaktører der i sin tid foretog excerperingen, især valgte eksempler på mindre hyppige ord og betydninger, samt dem der vurderedes som gode eksempler. Disse excerperede ord og deres betydninger er derfor dem som læseren plejer at have mest brug for når de



Figur 3: ONP Reader med billede af udgaven og ONP's citater fra udgavens side med eventuelle definitioner og andre oplysninger <onp.ku.dk/r10680-229> (april 2022).

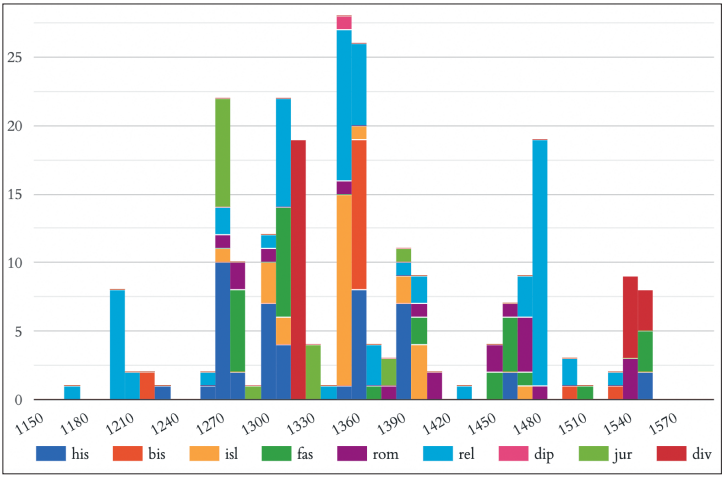
læser en ukendt tekst. Således har gennemsnitligt 4 % af alle ord på en side en definition i ONP på dansk og desuden enten ONP's egen engelske definition eller en automatisk oversættelse af den danske. I de øvrige tilfælde findes der stadig relevante oplysninger for brugeren: opslagsordet, links til andre ordbøger, eventuelle paralleller og grammatiske oplysninger.

5.2. Henvisninger til traditionelle materialer

Oplysningerne om udgaverne er et eksempel på at ordbogen kan bruges fra andre indgangsvinkler end ordforrådet alene. Gennem webapplikationen kan man også finde en tekst i en udgave og finde alle citater i ONP der er hentet fra tekstudgaven, og herved få et overblik over tekstens ordforråd.

Ordbogen er designet ud fra det princip at alle citater i ordbogen kan føres tilbage til en fysisk kilde – et håndskrift. ONP har samlet oplysninger om næsten 5.000 håndskrifter og diplomers

datering samt deres indhold. Man kan derved bruge ONP som et katalog over den norrøne prosalitteraturs kilder, og under hvert håndskrift opstilles alle citater der stammer fra det pågældende håndskrift, hvilket kan bruges til en undersøgelse af den fysiske kildes sprogbrug. For hvert lemma kan man få et overblik over ordets diakroniske forekomst i håndskrifterne, se figur 4.



Figur 4: Citater under *drekka* ('drikke') efter håndskriftsdato og genre <onp.ku.dk/015159> (april 2022).

Fra citatsiderne linkes der også til to andre ressourcer der har billeder af håndskrifterne: Handrit.is, som har digitale billeder af mange norrøne håndskrifter, og NorS Sprogsamlinger, som offentliggør scannede fotografier af håndskrifter i Den Arnamagnæanske Samlings fotosamlinger. Brugeren kan derfor få adgang til billeder af det oprindelige fysiske kildemateriale gennem websitet.

Som en del af redaktionsprocessen har ordbogen sammenlignet citaterne med udenlandske kilder og andre relevante paralleller til ordbogens kildeværker. I 35.000 citater findes der henvisninger til disse kilder som også er linket til registrene. Man kan så finde parallellerne under hvert værk i registrene, og webapplika-

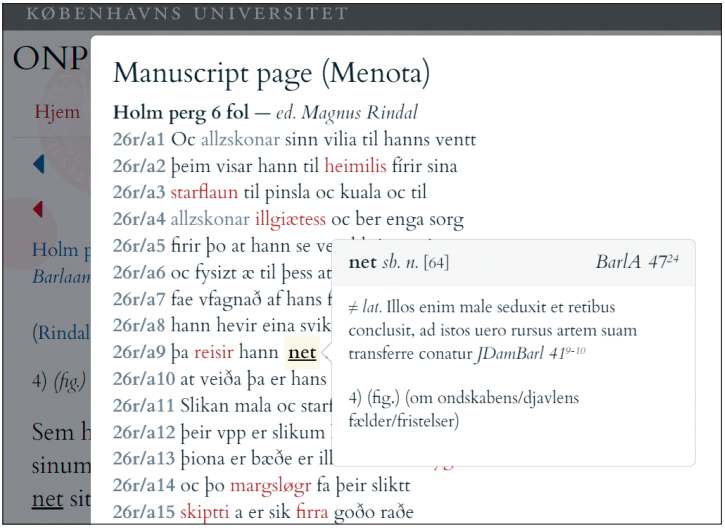
tionen kan generere en liste over alle citater hvor den parallelle tekst er blevet registreret med tilsvarende ord. Parallellerne findes også under ordbogsartiklerne selv, men den konsekvente måde hvorpå de er blevet registreret, betyder at ordforrådet også kan undersøges med de udenlandske kilder som udgangspunkt.

5.3. Forbindelse til digitale korpusser

Selvom ONP er en traditionel, excerperingsbaseret ordbog, har den flere fælles referencepunkter med nyere digitale korpusser. ONP's lemmaliste bruges som standarden for lemmatisering af digitale tekster, f.eks. i The Skaldic Project og Menota (se Haugen 2019, §11.2). De to projekter har stor nytte af ONP fordi der er tale om håndskriftsbaserede udgaver af norrøne tekster. The Skaldic Project er i gang med at redigere en stor del af det norrøne poesikorpus fra håndskrifterne, og det inkluderer også en oversættelse til engelsk med links på ordniveau mellem kildeteksten og oversættelsen. Menota arkiverer og offentliggør TEI-baserede udgaver af enkelte norrøne (og andre nordiske) håndskrifter. Disse udgaver har derudover ofte oplysninger om ordenes morfologi.

I The Skaldic Project's tilfælde har ONP givet adgang til sin digitale lemmaliste, og projektet har brugt den til at lemmatisere korpusset. De to projekter kan linkes sammen entydigt med hensyn til homografer og ortografi. ONP dækker ikke poetiske ord og tekster, men citaterne i ONP kan suppleres med ord fra The Skaldic Project's korpus (jf. ovenfor i afsnit 3.1).

Menotas tekstarkiv indeholder især prosalitteratur. Udgaverne bruger som regel ONP's ordliste til at identificere deres lemmaer, men uden præcisering af homografer. Menotas tekster har et andet fælles referencepunkt med ONP, da udgaverne er baseret direkte på enkelte håndskrifter. I princippet kan ord fra Menota linkes ikke alene med ONP's lemmaliste, men også med citaterne selv. ONP anvender nogle af Menota-teksterne så lemmaerne



Figur 5: Side fra Menotas korpus, med links til ONP's citater, definitioner m.m.

kan linkes utvetydigt til ONP's ordliste, og korpusset kan supplere ONP's citater med alle eksempler fra enkelte tekster (jf. figur 5). Yderligere automatisk og manuel bearbejdning kan linke selve citaterne til Menotas ord i fuld kontekst (se Wills, Jóhannsson & Battista 2018). I disse tilfælde er det muligt at generere en interaktiv tekst hvor ONP's oplysninger om de excerperede citater kan hentes direkte ved at klikke på ordene i Menotas udgave.

5.4. Digital adgang til ONP's data

ONP's webapplikation bruger flere eksterne ressourcer der hentes via en Application Programming Interface (API), som nævnt ovenfor (se afsnit 3.1). ONP gør også allerede nu sit eget materiale tilgængeligt for eksterne applikationer på en lignende måde, og den bruger således de relevante standarder.

ELEXIS-projektet arbejder på forskellige områder på at bygge en digital leksikografisk infrastruktur. Projektet har udviklet en standard-API for digital adgang til ordbøger (jf. ELEXIS Protocol for accessing dictionaries (1.1)), som specificerer hvordan eksterne programmer kan hente oplysninger om en ordbog, søge i lemmalisten og hente artikler. ELEXIS' API er blevet implementeret i ONP. Et eksternt program kan derfor søge i ONP's ordliste, hente hele ordlisten, finde opslagsord ved hjælp af en søgestreng (for eksempel efter første eller sidste del af ord) og hente digitale versioner af enkelte ordbogsartikler.

ELEXIS' API-specifikationer giver mulighed for at bruge to standarder til ordbogsartiklerne: Ontolex-Lemon (Bosque-Gil & Gracia 2019) i to forskellige digitale formater, eller TEI. Ontolex er ikke ideelt til historiske ordbøger da det er udarbejdet for moderne ordbøger og ikke specificerer kodning af henvisninger til tekster. Vi implementerer det alligevel da det kan bruges mere konsekvent end TEI.

I princippet kan API'en bruges til at integrere ONP's materiale i andre portaler og applikationer. ONP er i gang med at samarbejde med *málið.is* om at inkludere ONP i søgeresultaterne i deres portal.

6. Resultater

På trods af at ONP ikke er færdigredigeret, fungerer den allerede nu som et vigtigt hjælpemiddel for undersøgelser af det norrøne prosasprog og dets kilder. ONP og de digitale ressourcer som er tilgængelige på ONP's hjemmeside, er resultatet af 40 års arbejde med at supplere og effektivisere redigeringsprocessen ved til stadighed at tilføje nye data og gøre dem tilgængelige digitalt. Disse data indeholder righoldige oplysninger om det norrøne sprog, de fysiske og tekstuelle kilder, udenlandske paralleller og ældre ord-

bøger. Den seneste, aktuelle proces har gjort disse ressourcer, både ældre og nyere, tilgængelige for nye brugergrupper der nu kan drage nytte af dem i forskellige anvendelsessituationer. Samtidig bidrager ressourcerne fortsat til at forbedre redigeringsprocessen: Som nævnt ovenfor har de nye funktioner i webapplikationen gjort dele af redaktionsarbejdet væsentligt hurtigere.

ONP trækker også på forskellige eksterne digitale materialer og kilder for at supplere og berige ONP's materiale, som stadig er under udarbejdelse. Disse eksterne ressourcer består af digitale korpusser, digitaliserede ordbøger og andre leksikalske hjælpemidler. Korpusserne bidrager med konkordanser og i visse tilfælde oversættelser til ordene, og de leksikalske ressourcer samt ordbøger bidrager med artikler som fortsat mangler i ONP, målsprog/definitioner (især engelsk), nyere kildemateriale og lingvistiske oplysninger (etymologi og bøjningsoplysninger). Disse processer og ressourcer betyder at brugere altid kan tage udgangspunkt i ONP's hjemmeside for at undersøge et norrønt ord og dets brug, også i de tilfælde hvor ordet endnu ikke er blevet redigeret.

Som et resultat af denne udvikling kan vi konstatere at ONP's webbrugerbase ifølge Google Analytics er vokset fra ca. 1.000 pr. måned (2017) til over 5.000 pr. måned (2022). Brugerne besøger flere sider og bruger flere funktioner end tidligere. Dog bruges de mange muligheder der ligger dybere inde i webapplikationen, ikke i samme udstrækning. Vi arbejder fortsat på at gøre dem mere synlige og tilgængelige for vores brugere.

De metoder der beskrives og fremlægges i denne artikel, viser hvordan en digital ordbog kan sammenkoble funktioner fra forskellige ressourcer for at supplere sit eget materiale. Ordbogen – eller den hjemmeside hvor ordbogen publiceres – kan derfor altid fungere som udgangspunkt i en undersøgelse af sprogets, i dette tilfælde det norrøne prosasprogs, ordforråd.

Litteratur

Ordbøger

Blöndal = Blöndal, Sigfús (1920-1924): *Islandsk-dansk Ordbog. Íslensk-dönsk orðabók.* Reykjavík: Verslun Þórarins B. Þorlákssonar / Aschehoug.

ClV = Cleasby, Richard, Gudbrand Vigfusson [Guðbrandur Vigfússon] and W. A. Craigie (1957): *An Icelandic-English Dictionary*. 2. udg. Oxford: Clarendon.

Fritzner = Fritzner, Johan (1883-96): *Ordbog over det gamle norske sprog*. Kristiania (Oslo): Den norske forlagsforening.

Hertzb = Hertzberg, Ebbe (1895): Glossarium. I: Storm, Gustav og Hertzberg, Ebbe (udg.): *Norges gamle Love indtil 1387*. Bind 5. Kristiania.

LP = Finnur Jónsson (1931): *Lexicon poeticum antiquæ linguæ septentrionalis: Ordbog over det norsk-islandske skjaldesprog oprindelig forfattet af Sveinbjörn Egilsson*. 2. udg. Copenhagen: Møller.

málið.is = <malid.is> (april 2022).

MED = Robert E. Lewis, et al. (1952-2001): *Middle English Dictionary*. Ann Arbor: University of Michigan Press. Onlineudgave: Frances McSparran, et al. (2000-2018): *Middle English Compendium*. Ann Arbor: University of Michigan Library. <quod.lib.umich.edu/m/middle-english-dictionary> (april 2022).

ONP = Degnbol, Helle, Bent Chr. Jacobsen, Eva Rode, Christopher Sanders, Þorbjörg Helgadóttir, James E. Knirk, Aldís Sigurðardóttir, Alex Speed Kjeldsen, Ellert Þór Jóhannsson, Pernille Ellyton, Johnny Lindholm & Tarrin Wills (1989-): *Ordbog over det norrøne prosasprog // A Dictionary of Old Norse Prose*. København: Den Arnamagnæanske Kommission. <onp.ku.dk> (april 2022).

Anden litteratur

Battista, Simonetta & Ellert Þór Jóhannsson (2014): Ordbog over det norrøne prosasprog 2004-2014: fra trykt udgave til netversion. I: *LEDA-nyt* 57, 6-23.

Bosque-Gil, Julia & Jorge Gracia (udg.) (2019): *The OntoLex Lemon Lexicography Module*. <w3.org/2019/09/lexicog/> (april 2022).

ELEXIS: European Lexicographic Infrastructure. <elex.is> (april 2022).

Haugen, Odd Einar (2019): Linguistic Annotation. I: Odd Einar Haugen (udg.): *The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources*. Version 3.0. Bergen: Medieval Nordic Text Archive. <menota.org/HB3_ch11.xml>.

Jóhannsson, Ellert Þór & Simonetta Battista (2016): Editing and Presenting Complex Source Material in an Online Dictionary: The Case of ONP. I: Margalitadze, Tinatin et al. (udg.): *Proceedings of the XVII EURALEX International Congress*. Tbilisi: Tbilisi University Press, 117-128.

Mugglestone, Lynda (2005): *Lost for Words: The Hidden History of the Oxford English Dictionary*. Yale: Yale University Press.

TEI Consortium (2022): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 4.4. TEI Consortium. <tei-c.org/Guidelines/P5/> (april 2022).

Wills, Tarrin, Ellert Þór Jóhannsson & Simonetta Battista (2018): Linking Corpus Data to an Excerpt-based Historical Dictionary. I: Čibej, Jaka et. al. (udg.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 979-988.

Wills, Tarrin, Simonetta Battista & Ellert Þór Jóhannsson (2021): Making an ongoing historical dictionary accessible to an English-speaking audience. Euralex XIX, Greece. <eura-

lex2020.gr/wp-content/uploads/2021/10/EURALEX2020_BookOfAbstracts-Preview.pdf> (april 2022).

Anvendte korpusser

Handrit.is = Handrit.org / Handrit.is. <handrit.is> (april 2022).

Menota = Medieval Nordic Text Archive: Public Catalogue. <clarino.uib.no/menota/catalogue> (april 2022).

NorS Sprogsamlinger = Institut for Nordiske Studier og Sprogvidenskabs Digitale Sprogsamlinger. <sprogsamlinger.ku.dk> (august 2022).

The Skaldic Project = Skaldic Poetry of the Scandinavian Middle Ages. <skaldic.org> (april 2022).

Tarrin Wills
Ordbogsredaktør og -leder
Ordbog over det norrøne prosasprog
Københavns Universitet
Njalsgade 136
DK-2300 København S
tarrin@hum.ku.dk

RECENSIONER

Svensk ordbok – anden og reviderede udgave

Lars Trap-Jensen

Svensk ordbok utgiven av Svenska Akademien. Udarbejdet ved Institutionen för svenska språket, Göteborgs universitet. Gratis tilgængelig som app til iOS og Android og på Svenska Akademiens ordbogsportal svenska.se. Udkom maj 2021.

1. Indledning

Svensk ordbok (SO) redigeres på grundlag af en leksikalsk database der vedligeholdes af Institutionen för svenska språket ved Göteborgs universitet. Første udgave af ordbogen udkom som et trykt værk i to bind i 2009. Den underliggende database (*Lexikalisk databas*) går dog tilbage til det pionerarbejde der har foregået ved Göteborgs universitet siden 1970'erne, og som også var udgangspunkt for arbejdet med SO's forgængere, *Svensk ordbok* (SOB) udarbejdet af Språkdata ved Göteborgs universitet og Esselte Studium AB (1986) og *Nationalencyklopedins ordbok* (NEO) udarbejdet af Språkdata ved Göteborgs universitet og Bra Böcker AB (1995-96). Den første digitale version af SO kom i 2015 da en app blev lanceret, og siden 2017 har SO også været tilgængelig som netordbog via Svenska Akademiens ordbogsportal svenska.se. Både appen og netordbogen var indholdsmæssigt stort set identisk med den trykte udgave, dog med den væsentlige undtagelse at lydfiler i form af indtalt udtale af opslagsordene blev introduceret som en del af både app og netordbog.

SO har flere gange tidligere været anmeldt i *LexicoNordica*. Første udgave (SO1) blev anmeldt i *LexicoNordica* 17 af Kristina Nikula (Nikula 2010), appen af Susanna Karlsson i *LexicoNordica*

23 (Karlsson 2016), og en samlet anmeldelse af svenska.se ved Harry Lönnroth kan ses i *LexicoNordica* 25 (Lönnroth 2018).

Den nye udgave af SO så dagens lys i maj 2021 (SO2). Brugerne kunne dermed stifte bekendtskab med nyt ordbogsindhold for første gang siden første udgave fra 2009. Denne anmeldelse koncentrerer sig primært om de ændringer der er foretaget mellem første og anden udgave, og inddrager kun andre aspekter i det omfang de ikke har været behandlet i de nævnte tre foregående anmeldelser. Alle undersøgelser og opslag er foretaget i maj 2022.

2. Lemmabestand

Netordbogen på svenska.se indeholder kun i beskedent omfang omtekster der forklarer ordbogens principper, brugergrupper og indhold. Til gengæld får man en ganske udførlig forklaring på instituttets hjemmeside (gu.se/svenska-spraket/svensk-ordbok), og det meste af samme tekst genfindes i omteksten i appen (under menupunktet "Om Svensk ordbok"). Heraf fremgår det at anden udgave indeholder "närmare 65 000" opslagsord, mens første udgave indeholdt mellem 64.500 og 65.000[1]. De 12 år der er gået mellem første og anden udgave, er tilsyneladende ikke primært blevet brugt til at øge artikelbestanden. Det betyder dog ikke at der ikke er sket noget.

På den ene side er der blevet tilføjet nye ord der er kommet ind i sproget: *boost/boosta*, *chia* (med sammensætningerne *chiabröd*, *chiafrö*, *chiapudding*), *crowdfunding*, *flockimmunitet*, *flygskam*, *hen*, *klicker/clicker*, *kryptovaluta*, *mikroplast*, *popup*, *supermåne*, *zikavirus* for at nævne nogle stykker. Omvendt er der også fjernet ord fra ordbogen hvis redaktionen har fundet at de er sjældne i den moderne sprogbrug. Eksempler er *benrangelsmannen*, *datorpost*,

1 Begge tal angives i forord og indledning, s. VII henholdsvis s. IX (jf. Nikula 2010:351).

frökenspel, *förtroendechocka*, *nervdroppar*, *telegrafnyckel*, *tjuvspråk*, *trasgrann*, *vårdslösa* og *äggpunkterare* (ifølge Nyhedsmeddelelse, Göteborgs universitet (2021)).

Hvor stor en udskiftning af ordbestanden der er foretaget, kan man få et indtryk af ved at se på et alfabetisk udvalg. Jeg har foretaget en sammenligning af de 200 første ord i bogstavet L i SO1 og SO2. Kun hovedopslagsord er talt med i denne opgørelse. Lemmalisten indeholder derudover et antal underopslagsord der vises som orddannelseseksempler, typisk sammensætninger nævnt under simpleksordet uden at have egen artikel. De ord som efter min optælling adskiller sig i SO1 og SO2, er anført i tabel 1.

I SO1, ikke i SO2	I SO2, ikke i SO1
laboratris, *lafs*, *lafsa*, *lafsig*, *lagerbärsblad*, *lagerkörs*, *landdriven*	*labbe*, *ladda upp*, *laddare*, *laddhybrid*, *laddram*, *laddstolpe*, *lagning*, *lagringsminne*, *lagspelare*, *lajv*, *lajva*, *lallare*, *laminär*

Tabel 1: Forskel i lemmabestand; sammenligning af de 200 første ord i bogstav L.

Det fremgår at 7 ord er fjernet, 13 nye er tilføjet, mens de øvrige opslagsord i det alfabetiske udsnit er sammenfaldende efter min optælling. I runde tal vil det sige at omkring 95 % af ordbestanden er uforandret, mens 5 % er udskiftet i det undersøgte udsnit. Om udsnittet er repræsentativt for hele lemmabestanden, er naturligvis behæftet med ganske stor statistisk usikkerhed når det kun drejer sig om 200 ord. Men hvis man leger med tanken at det er tilfældet, ville det svare til at omkring 3.250 ord er taget ud og omtrent samme antal føjet til, hvorved den samlede ordbestand er forblevet nogenlunde uændret.

At der er fjernet omtrent lige så mange som der er tilføjet, vidner om at redaktionen tager det alvorligt at SO skal afspejle det aktuelle sprog. Hvordan redaktionen definerer aktuelt sprog, har jeg ikke kunnet læse i omteksten, men et vist redaktionelt skøn må

antages at være på spil. Ordforrådet i SAOL14 spiller en vigtig rolle i opdateringen på lemmaniveau ifølge Sköldberg & Hannesdóttir (2017:334-338), dog suppleret med lemmakandidater fra et bredere udvalg af korpustekster (op.cit.:338). For ældre eller nu historisk sprogbrug må brugeren konsultere SAOB. Ikke mindst i dag hvor alle Akademiets ordbøger er tilgængelige fra samme portal, giver det god mening at have sådan en arbejdsdeling.

Ændringer i lemmabestanden andrager ikke noget vældig stort antal set over en periode på 12 år, og som konsekvens heraf har jeg søgt forgæves efter en del oplagte ord der er kommet til i sproget i nyere tid. Man finder således kun få ord der vedrører den nylige pandemi og slet ingen hovedopslagsord med *corona-* eller *covid-*. Ligeledes har jeg søgt forgæves efter en del hovedopslagsord inden for nogle områder der erfaringsmæssigt leverer mange nyord: teknik (fx *kroppsscanning*, *robotdammsugare*, *smartmobil*, *smarttelefon*, *videolänk*, *videomöte*, *videosamtal*) og madlavning (fx *edamameböna*, *halloumi*, *hoisinsås*, *kimchi*, *poke*, *ramen*, *tagine*, *tempura*, *teriyaki*). Med til billedet hører dog at SO2 har et stort antal søgbare underopslagsord, og selvom disse altså ikke har egne artikler, får man undertiden en kort forklaring. Eksempelvis findes et enkelt af de ovennævnte ord som underopslagsord, nemlig i artiklen *smart*: "äv. bildligt, särskilt om datastyrda processer SAMMANSÄTTN./AVLEDN.: *smarttelefon*".

3. Betydningsbeskrivelse

Der er ikke foretaget radikale ændringer af de overordnede principper i SO2. Redaktionen har bevaret de centrale elementer i artiklernes opbygning og definitionspraksis, og hvorfor også ændre på det? Principperne blev grundlagt med SOB i 1986 og er for størstepartens vedkommende overtaget i de følgende udgaver. Det hedder således i anmeldelsen af NEO: "Däremot verkar inte defi-

nitionerna och språkexemplen ha ändrats i särskilt stor utsträckning för de uppslagsord som är gemensamma" (Cantell & Martola 1996:221).

Nikula (2010:371) havde overordnet ros til betydningsforklaringerne i SO1 ("Betydelsesbeskrivningarna är så vitt jag kunnat se korrekta och tillräckliga för reception af text"), og også i anmeldelsen af forgængeren, NEO, blev dette fremhævet.[2] Jeg slutter mig til rækken af tilfredse brugere: SO's definitioner fungerer i langt de fleste tilfælde godt. Stilen med en central beskrivelse skrevet med ordinær skrift og efterstillede tillægsoplysninger i mindre skriftgrad er bevaret fra de tidligere trykte udgaver. Det er en stil som jeg ikke kender fra mange andre ordbøger, men den er med til at give SO sit eget præg, og det er en stil som brugeren godt kan vænne sig til og endda lære at sætte pris på.

Selvom der altså overordnet er ros til betydningsbeskrivelserne, er der dog også enkeltstående kritikpunkter. Eksempelvis nævner Nikula (2010:358) at forklaringen af ordet *svin* allerede ved udgivelsen af SOB i 1986 blev kritiseret for at indeholde for vanskelige ord ('typ av partåigt hovdjur med lång, kraftig kropp, förlängt nosparti och korta ben'). Alligevel optræder definitionen igen i SO1, og jeg kan konstatere at det samme er tilfældet i SO2. Det samme gælder en tilsvarende kritik af ordet *byggnad* i anmeldelsen af NEO: "Frågan är om inte uppslagsordet är förståeligare än definitionen som skall förklara det" (Cantell & Martola 1996: 211). Også denne definition genfindes imidlertid uændret i både SO1 og SO2. Heller ikke Nikulas kritik af inkonsistent valg af træk til beskrivelse af ordfeltet 'sygdom' har ført til ændringer i SO2 på de punkter som blev fremført (Nikula 2010:359f.).

2 "Över lag är definitionerna i NEO klargörande och bra" (Cantell & Martola 1996:211).

4. Indholdsrevision

En ordbog der grunder sig på arbejde fra 1970'erne, kan ikke undgå at rumme definitioner, brugseksempler og redaktionel tekst som set med dagens øjne virker utidssvarende. Redaktionen af SO2 har derfor naturligt nok været optaget af at opdatere indholdet i SO2. Et prioriteret område har været at undgå at formidle værdier og udsagn som virker diskriminerende på grupper i samfundet (jf. Petersson & Sköldberg 2020). Dette arbejde omfatter opdatering af artikelbestanden, udformning af betydningsbeskrivelse, valg af sprogbrugsmarkører, valg af sprogbrugseksempler såvel som linkning mellem artikler og betydninger (Petersson & Sköldberg 2020:102).

I flere sammenhænge (se Petersson & Sköldberg 2020:98, Sköldberg 2018) har det været nævnt at redaktionen har fokus på ord der falder inden for de syv områder som anføres i den svenske diskriminationslov fra 2020: køn, kønsoverskridende identitet eller udtryk, etnisk tilhørsforhold, religion eller anden trosopfattelse, funktionsnedsættelse, seksuel orientering og alder.

Jeg har derfor undersøgt et tilfældigt antal ord der falder inden for disse områder for at se hvad og hvor meget der er revideret i de pågældende artikler. Ordene har jeg udvalgt fra Den Danske Ordbogs interne liste over potentielt kontroversielle ord (udarbejdet af min kollega Sanni Nimb) og uden på forhånd at have undersøgt hvordan de var behandlet i de svenske ordbøger. Resultatet er sammenfattet i tabel 2.

Opslagsord	SO1	SO2
bimbo	flicka eller ung kvinna som väcker (medial) uppmärksamhet genom utmanande stil och som är (el. låtsas vara) intellektuellt relativt outvecklad	<vardagligt; nedsättande> (ung) kvinna som väcker (medial) uppmärksamhet genom utmanande stil och som är (el. låtsas vara) intellektuellt relativt outvecklad
bushman	(vanl. plur.) medlem af sanfolket <åld.>	<ålderdomligt> (vanligen plur.) medlem av sanfolket
dvärg	onormalt liten människa vars kropslängd väsentligt understiger den för åldern normala, men som vanl. har normalstort huvud	<nedsättande; något ålderdomligt> onormalt liten människa vars kroppslängd väsentligt understiger den för åldern normala men som vanligen har normalstort huvud
dövstum	döv och stum vanl. (förr) om persons om inte kan tala på grund av medfödd el. tidigt uppkommen dövhet <föråldrad (och) missvisande) beteckning>	<föråldrad (och missvisande) beteckning> döv och stum vanligen (förr) om person som inte kan tala på grund av medfödd el. tidigt uppkommen dövhet
efterbliven	som inte utvecklats på normalt sätt om person, särsk. intelligensmässigt <ngt nedsätt.; åld>	<nedsättande; ålderdomligt> som inte utvecklats på normalt sätt om person, särsk. intelligensmässigt
fläskberg	äv. mycket fet person	äv. <vardagligt> mycket fet person
fruntimmer	kvinna ofta ngt nedsätt. el. skämts.; särsk. i mäns språk>	<ofta något nedsättande> kvinna
horbock	man som ägnar sig åt erotiska utsvävningar <ngt åld.; nedsätt.>	<något ålderdomligt; nedsättande> man som ägnar sig åt erotiska utsvävningar
krympling	person vars kropp är vanställd och som därigenom har nedsatt rörelseförmåga särsk. om person som har mist ngn av sina lemmar <ngt åld.>	<ålderdomligt> person vars kropp är vanställd och som därigenom har nedsatt rörelseförmåga särsk. om person som har mist någon av sina lemmar
mulatt	(manlig) avkomling av en svart och en vit person el. av två mulatter	<kan uppattas som nedsättande> avkomling av en person med mörk hy och en person med ljus hy el. av två mulatter

naturfolk	folk som lever i nära kontakt med och i direkt beroende av naturen på ett kulturstadium som inte omfattar högre jordbruksteknik el. industri <kan uppfattas som nedsätt.>	<kan uppfattas som nedsättande> folk som lever i nära kontakt med och i direkt beroende av naturen på ett kulturstadium som inte omfattar högre jordbruksteknik el. industri
slampa	sexuellt vidlyftig kvinna <vard., starkt neds.>	<vardagligt, starkt nedsättande> kvinna som anses vara (för) sexuellt vidlyftig
svagsint	(ofta substantiverat) som saknar grundläggande själslig förmåga <åld. i fackmässiga sammanhang>	<ålderdomligt i fackmässiga sammanhang> (ofta substantiverat) som saknar grundläggande själslig förmåga
ungtupp	ofta bildligt (självsäker) ung man gärna med anspråk på makt el. inflytande	<ofta bildligt> (självsäker) ung man gärna med anspråk på makt el. inflytande
vamp	sexuellt utmanande kvinna särsk. (förr) som rolltyp i filmer <vard.>	<vardagligt> sexuellt utmanande kvinna särsk. (förr) som rolltyp i filmer

Tabel 2: Definitioner og sprogbrugsoplysninger for 15 (potentielt) kontroversielle ord.

Umiddelbart iøjnefaldende er sprogbrugsmarkeringernes placering og de udskrevne forkortelser. Men da det vedrører elementernes præsentation snarere end deres indhold, er det ikke det der kommenteres her (se nærmere herom i 6.1). Indholdsmæssigt er 8 af de 15 artikler blevet revideret, altså omtrent halvdelen[3]. I alle tilfælde går ændringen i retning af en skærpet advarsel, enten ved at en sprogbrugsmarkering er tilføjet: <nedsättande> (i *bimbo, dvärg*, *mulatt*), <ålderdomligt> (i *dvärg*) og <vardagligt> (i *bimbo*, *fläskberg*), ved at én af flere markeringer er fjernet eller ændret (i *fruntimmer*, *krympling*), eller ved at definitionen er omformuleret (*bimbo, slampa*).

3 En fin service er at man på svenska.se kan se tidligere udgaver af ordbøgerne. Man kommer til dem ved at følge linket SAOLhist i nederste højre hjørne (et noget misvisende navn eftersom man altså også får adgang til SO1).

Af de øvrige artikler, hvori ingen revisioner er foretaget, er blot én helt uden sprogbrugsmarkering (*ungtupp*); de øvrige havde i forvejen en markering som er bevaret. Dette stemmer godt overens med Nikulas iagttagelse vedrørende eksempler: "Ur genussynvinkel kan konstateras att en del av den gamla surdegen rensats bort" (Nikula 2010:364). I SO1 var allerede meget gjort, og med SO2 fortsætter arbejdet for at rette op på skævheder og advare om kontroversiel sprogbrug. Det ses også af at antallet af advarende sprogbrugsmarkører er jævnt og kumulativt stigende fra SOB til SO2: I SOB var 7 af de 15 artikler helt uden markering, i NEO 6 (desuden var hverken *bimbo* eller *fläskberg* opslagsord i SOB og NEO), i SO1 5 og i SO2 altså bare 1.

Indtrykket er – i det omfang man kan få det ved nedslag i et begrænset antal mere eller mindre tilfældige ord – at redaktionen har arbejdet ganske grundigt med ordfeltet "kontroversielle ord" og foretaget revisioner hvor det var påkrævet, i forlængelse af den indsats der allerede var gjort i tidligere udgaver.

5. Konstruktionsoplysninger

Redaktionen har taget et andet kritikpunkt til sig. Nikula (2010:368) siger om valensoplysningerne:

> Här frågar man sig om det inte vore bättre att omedelbart efter betydelsen också ange valensen eftersom det då skulle vara möjligt att se valensuppgifterna konkretiserade i de efterföljande exemplen.

Og netop det er sket, jf. Blensenius (2019:211). I SO2 står oplysning om valens mellem sammensætninger/afledninger og eksemplerne. Man har endda fulgt Nikulas anbefaling og udskiftet eksemplet der illustrerer angivelsen *läsa ngn*, så der nu står "läsa Strindberg"

frem for det lidt sværere afkodelige "vår mest läste författare".

En anden ændring er at SO2 tilbyder hjælp til afkodning af komplekse valensangivelser i tilfælde hvor flere optionelle aktanter er opregnet i samme mønster. Det er gjort på den måde at brugeren med et klik kan udfolde den komplekse angivelse og få de enkelte muligheder vist for sig. Et eksempel er vist i figur 1.

Figur 1: Valensoplysning, sammenfoldet og udfoldet.

Det sammenfoldede mønster fylder ikke meget, og den bruger der har slået op af helt andre grunde, kan let læse hen over valensangivelsen, mens den interesserede bruger får klar besked når mønstret udfoldes med et klik. Det er en elegant løsning og en klar forbedring i forhold til SO1.

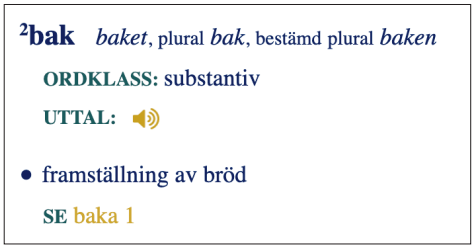
Man kan samtidig notere sig at den kondenserede stil fra den trykte SO1 med brug af tilde for opslagsordet og forkortelserne *ngn*, *ngt* osv. er ændret til et mere læsevenligt layout med aktanterne i kapitæler og opslagsordet skrevet fuldt ud. Også det hilses velkomment.

6. Fra papir til skærm

At ændre publiceringsform fra trykt bog til elektronisk udgave er ingen triviel sag. Papirformatets komprimerede og pladsøkonomiske løsninger kræver omlægning ved overgang til digital publicering, ligesom der skal gøres en indsats hvis det digitale formats muligheder for linkning mellem artikler og oplysninger skal udnyttes optimalt. I det følgende ser jeg på hvor godt det er lykkedes for to udvalgte områder: forkortelser og søgning efter faste udtryk.

6.1 Forkortelser

Karlsson (2016:255) savner i SO-appen muligheden for, i lighed med SAOL13-appen, at se bøjningsoplysninger skrevet helt ud i et fuldstændigt paradigme. Det kan hun glæde sig over er sket i SO2. Faktisk er ikke kun bøjningsoplysningerne udfoldet; det samme gælder oplysning om ordklasse samt de kommentarer der knytter sig til bøjningerne. Et eksempel er givet i figur 2.



Figur 2: ordklasse og bøjningsoplysninger.

Det er en klar forbedring, og man må håbe at redaktionen går videre med indsatsen. Der er ikke rigtig gode argumenter for at bevare forkortelser i digital tekst ud over de helt etablerede almensproglige forkortelser som *dvs.*, *m.m.*, *osv.* og *t.ex.*, og her mangler der stadig at blive gjort en del i SO2. I ordlisten til højre i netordbogen optræder stadig *subst.*, *adj.* og *adv.* i forkortet form,

mens de øvrige ordklasser er skrevet helt ud. Det er næppe pladshensyn der er baggrunden – eksempelvis fylder *infinitivmärke* og *subjunktion* mere, men disse er foldet ud. Når ordklasseangivelser optræder andre steder i redaktionel tekst, er de nogle gange forkortede, fx i kommentarer som "(med prep. **med**)", "(i förbindelse med subst.)", i andre tilfælde er de skrevet ud, fx "(i förbindelse med adjektiv)".

Oplysninger om stilleje, anført i spidse parenteser før ordforklaringen, er nu ligeledes skrevet fuldt ud, vistnok overalt. Det betyder at betegnelser som *högt.*, *neds.*, *skämts.*, *vard.* og *åld.* er blevet erstattet af *högtidligt*, *nedsättande*, *skämtsamt*, *vardagligt* og *ålderdomligt*, og det er glædeligt. Jeg opfatter udfoldningen af forkortelser som en igangværende opgave. De anførte stilmarkeringer optræder stadig på forkortelseslisten (tilgængelig i appen), og jeg antager derfor at de er i brug i anden redaktionel tekst. Det samme gælder en række øvrige forkortelser som *avledn.*, *äv.*, *bildn.*, *el.*, *fornnord.*, *förkort.*, *lat.*, *pass.*, *plur.*, *refl.*, *sammansätn.*, *särsk.*, *spec.* og *övers.* som jeg alle er stødt på i redaktionel tekst. Det er blot at håbe på at redaktionen fortsætter det gode arbejde så flere forkortelser foldes ud.

6.2 Søgning efter faste udtryk

Placeringen af faste udtryk er et tema som har optaget leksikografien gennem mange år (se fx Atkins & Varantola 1998, Bogaards 1998, Lorentzen 1996): Under hvilket opslagsord placeres det faste udtryk, og hvor forventer brugeren at finde det? I en digital ordbog burde problemet egentlig være løst. Dels gør det ikke længere eksisterende pladsproblem at udtrykket kan opføres under et hvilket som helst af ordene i udtrykket, dels kan brugeren blot søge på flere ord og komme direkte til udtrykket.

I princippet er det ganske vist sådan, men det kræver at den underliggende database er organiseret på en måde så søgninger og

visninger fungerer hensigtsmæssigt. Det gør de desværre ikke altid i netordbogen på svenska.se.

Nikula (2010:365) nævner *gå som smort* som eksempel på et udtryk det er svært at finde i papirordbogen SO1. I SO2 er det ikke blevet meget nemmere. Hvis man indtaster udtrykket i søgefeltet, bliver man skuffet: "Sökningen på **gå som smort** i SO gav inga svar". Man må derfor forsøge sig frem med ét af ordene. En fordel ved den elektroniske søgning er dog at en søgning på *smort* leder brugeren direkte til *smörja*. Men her bliver brugeren igen skuffet: Udtrykket er ikke forklaret dér. Der er blot – som et levn fra papirordbogen – en henvisning til *gå.* Nået frem til *gå* må brugeren derefter åbne de enkelte betydninger ved at klikke på "VISA MER +" – disse er som udgangspunkt sammenfoldede – og scrolle sig forbi 90 (sic!) faste udtryk før det ønskede udtryk endelig findes.

Noget tilsvarende gør sig gældende med den inkonsistens Nikula (2010:360f.) nævner med hensyn til placeringen som hovedlemma/sublemma: *till fullo* behandles på samme måde som partikelverber og anføres som hovedlemma på alfabetisk plads, mens *för övrigt* må findes som et sublemma under adjektivet *övrig*. De brugere der ikke kan gennemskue princippet, får heller ikke i SO2 nogen hjælp hvis de har slået *till fullo* op under adjektivet *full*. Og det samme er tilfældet hvis de forsøger sig med at skrive *för övrigt* i søgefeltet. Det er ikke så godt.

Så går det noget bedre ved søgning i appen. Her har man mulighed for at slå udtrykket op ved at skrive ét eller flere ord i søgefeltet og derefter få matchende udtryk som resultat. Det er stadig sådan at *gå som smort* kun findes under *gå* (og ikke under *smörja*, hvor der i lighed med netordbogen blot er en henvisning), men i det mindste kommer man hurtigt til det søgte udtryk. Og man får også brugbare resultater ved søgning på *fullo* og *för övrigt*.

Teknisk og datamæssigt er det altså muligt at tilbyde en acceptabel søgevej. Når det alligevel ikke er gjort i netordbogen, kan det skyldes søgetekniske forhold (måske et krav om at bruge samme sø-

gemåde i alle portalens tre ordbøger?), eller det kan være en simpel fejl. Uanset hvad årsagen er, håber jeg der findes en løsning, for det nedsætter desværre netordbogens anvendelighed betragteligt. Selv bruger jeg derfor konsekvent appen til at søge efter faste udtryk.

7. Stilbokse

For mange kommer det nok som en overraskelse at SO2 har valgt at fjerne de ca. 400 stilbokse ("stilrutor") der ellers blev lanceret i SO1 som "en nyhet i svenska ordböcker" (SO1:VII). I disse stilbokse blev især sprogrigtighedsspørgsmål taget op, men de kunne også behandle emner som etymologi eller encyklopædisk information. Måske har redaktionen fundet det for vanskeligt at navigere mellem den mere normerende og vejledende rolle der anvendes i stilboksene (af typen "Uttalet [asepte´ra] är inte ovanligt men måste anses felaktigt") og den mere deskriptive linje som ellers kendetegner ordbogen. Som ordbogskollega kan jeg have forståelse for redaktionens dilemma, men som bruger af ordbogen vil jeg savne stilboksene. De behandlede netop den slags sprogspørgsmål som mange brugere er optaget af. Ud over deres oplysende karakter bidrog de også til ordbogens underholdningsværdi og stimulerede til lystlæsning – på samme måde som man har glæde af de mere end 1000 sentenser der er drysset rundhåndet ud blandt ordbogsartiklerne. Forhåbentlig er stilboksene ikke gemt længere væk end at de kan blive taget til nåde igen ved en senere lejlighed. De kan jo bruges til andet end sprogrøgt.

8. Sammenfatning

Svensk ordbok indskriver sig i en videnskabelig leksikografisk tradition der går mindst 50 år tilbage. Arbejdet med *Lexikalisk data-*

bas er den røde tråd som går igennem det leksikografiske arbejde der har foregået konstant siden 1970'erne ved Göteborgs universitet, og det har nu resulteret i fire ordbogsværker der hver især har bygget oven på sine forgængere. Udgangspunktet var godt – SOB fra 1986 er en udmærket ordbog – og de følgende ordbøger har kunnet tage det bedste fra den med videre og forbedre den på de punkter der måtte ønskes. Det er derfor ikke mærkeligt at Nikula (2010) allerede i titlen på sin anmeldelse kalder SO1 for "en guldgruva för språkintresserade". Fra SO1 til SO2 er der foretaget yderligere forbedringer: Layoutet er forbedret (til skærmen) med mange udfoldede forkortelser; der er fundet en elegant løsning på de ellers svære konstruktionsoplysninger, og deres nuværende placering er en støtte for brugeren; en stor og nyttig indsats er gjort for at opdatere artikelbestanden så artikler og citater ikke videreformidler alt for mange stereotype og ekskluderende udsagn. Når det forholder sig sådan, er det indlysende at SO2 stadig er en guldgrube. Der er dog også ting der kan forbedres: Lemmabestanden er ikke helt opdateret; der er stadig for mange uudfoldede forkortelser i redaktionel tekst, og ikke mindst fungerer søgning efter flerordsforbindelser dårligt, især i netordbogen på svenska.se. Så ja, *Svensk ordbok* er fortsat en guldgrube, men ikke alt i gruben er guld af højeste karat.

Litteratur

Ordbøger

NEO = *Nationalencyklopedins ordbok* 1995-1996. Språkdata, Göteborg, och Bokförlaget Bra Böcker AB, Höganäs.

SOB = *Svensk ordbok* 1986. Stockholm: Esselte Studium.

SO1 = *Svensk ordbok utgiven av Svenska Akademien*, Band 1 (A-L), band 2 (M-Ö), utarbetad vid Redaktionen för Svenska Akade-

miens samtidsordböcker, Lexikaliska institutet, Institutionen för svenska språket vid Göteborgs universitet. Förlag: Norstedts Akademiska förlag (i distribution). Stockholm 2009.

SO2 = *Svensk ordbok utgiven av Svenska Akademien*, utarbetad vid Institutionen för svenska språket vid Göteborgs universitet, app till iOS och Android og på Svenska Akademiens ordbogsportal svenska.se. Göteborg 2021.

Anden litteratur

Atkins, Beryl S. & Krista Varantola (1998): Language learners using dictionaries: the final report on the EURALEX/AILA research project on dictionary use. I: Beryl S. Atkins (ed.): *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Lexicographica. Series Maior 88. Tübingen: Niemeyer, 21-81.

Blensenius, Kristian (2019): Revision av konstruktionsuppgifter i *Svensk ordbok utgiven av Svenska Akademien*. I: *LexicoNordica* 26, 203-223.

Bogaards, Paul (1998): Scanning long entries in learner's dictionaries. I: Thiery Fontenelle, Philippe Hiligsmann, Archibald Michiels, André Moulin & Siegfried Theissen (eds.): *Actes EURALEX '98: Communications soumises a EURALEX '98 (Huitième Congrès International de Lexicographie) à Liège, Belgique*. Liège: Université de Liège, Départements d'anglais et de néerlandais, 555-563.

Cantell, Ilse & Nina Martola (1996): *Nationalencyklopedins ordbok*. I: *LexicoNordica* 3, 209-221.

Göteborgs universitet (2021): Forskargrupp bakom ny upplaga av Svensk ordbok [nyhed på universitetets hjemmeside]. <gu.se/nyheter/forskargrupp-bakom-ny-upplaga-av-svensk-ordbok> (maj 2022).

Institutionen för svenska, flerspråkighet och språkteknologi: *Svensk ordbok* [omtekst og brugervejledning]. <gu.se/svenska-spraket/svensk-ordbok> (maj 2022).

Karlsson, Susanna (2016): *Svensk ordbok* som app – något att glädjas över. I: *LexicoNordica* 23, 247-259.

Lorentzen, Henrik (1996): Lemmatization of Multi-Word Lexical Units: In which Entry? I: Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström, Catarina Röjder Papmehl (eds.): *EURALEX '96. Proceedings I-II*. Göteborg 1996: Göteborg University, Department of Swedish, 415-421.

Lönnroth, Harry (2018): Portalen svenska.se – en ny digital samlingsplats för språkresurser från Svenska Akademien. I: *LexicoNordica* 25, 281-292.

Nikula, Kristina (2010): Svensk ordbok – en guldgruva för språkintresserade. I: *LexicoNordica* 17, 351-375.

Nyhedsmeddelelse, Göteborgs universitet. Göteborg 2021, <gu.se/nyheter/forskargrupp-bakom-ny-upplaga-av-svensk-ordbok> (maj 2022).

Sköldberg, Emma & Anna Helga Hannesdóttir (2017): Svenska ord – men vilka? Om uppslagsorden i *Svensk ordbok utgiven av Svenska Akademien*. I: Emma Sköldberg, Maia Andréasson, Henrietta Adamsson Eryd, Filippa Lindahl, Sven Lindström, Julia Prentice & Malin Sandberg (red.): *Svenskans beskrivning* 35. Göteborg: Göteborgs universitet, 329-340.

Sköldberg, Emma (2018): Samhällsförändringar ska speglas i Svensk ordbok. I: *Språkbruk* 4/2018, 11-15. <sprakbruk.fi/-/samhallsforandringar-ska-speglas-i-svensk-ordbok> (maj 2022).

Lars Trap-Jensen
Ordbogsredaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
ltj@dsl.dk

Ordklok – mer om ordens levande fränder än deras urgamla anor

Bo-A. Wendt

Staffan Fridell: *Ordklok. Svenska ords släktskap och ursprung*. Stockholm: Kaunitz-Olsson 2021. 360 s.

1. Inledning

Att det utkommer en ny ordbok i tryck är inte längre någon alldaglig händelse, och att det utkommer en ny etymologisk ordbok har, inom det nordiska språkområdet och för den delen överhuvudtaget, alltid varit en sällsynthet. Så det är därför verkligen ingen liten händelse att det för svenska nu har ägt rum både det ena och det andra av detta – nämligen i form av ordboken *Ordklok. Svenska ords släktskap och ursprung* (2021) av Uppsalaprofessorn Staffan Fridell.

Etymologiska ordböcker har som sagt aldrig duggat tätt. För svenska kom den senaste 2008: Birgitta Ernbys *Norstedts etymologiska ordbok*. Detta är en enbandsordbok som är mer populärt hållen än det enda större verket i den svenska genren: Hellquists *Svensk etymologisk ordbok* i två digra band, i sin första upplaga utgiven 1922, i sin andra (efter författarens död) 1939. Ett supplement till detta monumentalverk har varit undervägs ända sedan 1970-talet, men ännu återstår flera delar av alfabetet att skriva och inget av det skrivna har gjorts allmänt tillgängligt. I den behändigare och populärare subgenren fanns före Ernbys ordbok framförallt Elias Wesséns *Våra ord*, först utgiven 1961 och sedan i en ny upplaga med flera nytryck, i sin tur en omarbetad upplaga av Levander & Wessén 1932. Inte alls lika spridd är väl Einar Odhners etymologiska ordlista med kortfattade etymologier från 1952 (och

senare). Med en mycket mer berättande framställning och ett urval av ord som lämpar sig för en sådan finns Gösta Bergmans *Ord med historia* från 1966 och senare. Också ett par andra Bergman har lämnat bidrag inom denna sidofåra: Bo Bergman med *Ordens ursprung* från 2007 och Olle Bergman i en serie tematiska småskrifter, inledd med *Krigiska ord* från 2011.

För de närmast besläktade grannspråken är förhållandena överlag snarlika. För danska finns Niels Åge Nielsens gedigna *Dansk Etymologisk Ordbog. Ordenes Historie* från 1966 (och senare) och *Politikens Etymologisk Ordbog* (av Jan Katlev) från 2000. För isländska finns Ásgeir Blöndal Magnússons *Íslensk orðsifjabók* från 1989. Och för norska – som det av dessa språk som ståtar med det allra lödigaste och mest uppdaterade bidraget i genren – föreligger Harald Bjorvands & Fredrik Otto Lindemans *Våre arveord* från 2000 (och senare) med fylligare och mer explicit resonerande etymologier men i gengäld enbart, som titeln anger, för språkets centrala arvord; därutöver finns Hjalmar Falks och Alf Torps ordböcker från tidigt 1900-tal, med endast den tyskspråkiga av dem i en senare upplaga från 1960. För fornvästnordiska är det senaste bidraget Jan de Vries' ordbok från 1961 (med en andra upplaga året därpå).

Fridells ordbok är alls ingen ersättare för den ännu osupplementerade Hellquist och än mindre en svensk motsvarighet till Bjorvand & Lindeman. Istället är den närmast en avlösare till Ernby och Wessén, men dess uttalat populariserande inriktning är ännu tydligare och ordurvalet avsevärt mer insnävat än i dessa ändå rätt traditionella föregångare. *Ordklok* föreligger i smakfullt gult tygband med ett omslag som visar en abborre simmande genom ett stort Ö; *abborre* är ordbokens första ord, medan dess sista, *övrig*, inte lika naturligt lånat sig till bildsättning. Som undertiteln anger är det de svenska ordens inbördes släktskap som är ett av ordbokens huvudärenden, kanhända det enskilt viktigaste, och detta sätter också sin prägel på både ordurval och artikelstruktur, som det skall framgå av följande granskning.

Denna granskning är uppdelad i tre huvudavdelningar: ordurval (avsnitt 2), åtkomststruktur (avsnitt 3) och – som den utförligaste – artikelinnehåll (avsnitt 4). Sist ett sammanfattande omdöme (avsnitt 5). Ordbokens artiklar omtalas, efter dess egen grafiska sed, med versalinledd fetstil, i förekommande fall plus homografsiffra.

2. Ordurval

I ordbokens förord (s. 3) heter det:

> Bokens fokus på släktskap och sammanhang innebär att endast ord med germanskt ursprung har tagits med, närmare bestämt arvord samt lånord från andra germanska språk, där de tyska är helt dominerande.

Att detta fokus *innebär* just detta utelämnande av andra lånord än de germanska följer väl inte av någon helt vattentät logik. Det finns ju ett stort antal utomgermanska, främst romanska eller grekiska, lånord som också ytterst har släktskap med arvord, om än likheten i ljudbild genast blir mindre genomskinlig. Att det finns sådan utomgermansk släktskap som man också kan bli ordklok på visas för övrigt av att författaren i några, i och för sig få, fall ändå inte kunnat låta bli att nämna den, såsom i artikeln **Flat** släktskapen med *platt*, i **Fä** den med *pekuniär*, i **Varm** den med *termos* och *termometer* och i **Örn** den med *ornitologi*. I en del artiklar nämns utomgermanska ordfränder utan (explicit) koppling till lånord: i **Ejder** och **Åda** latinets *avis*, i **Mjöd** ryskans *medved* 'björn' (< 'honungsätare') och i **Tjur** latinets *taurus*. Riktigt gamla lånord från annat än germanska systerspråk får till och med vara med i egna artiklar, framförallt troliga eller säkra keltiska lån som i **Alp**, **Järn**, **Rik** och **Rike** samt **Ämbete**, men också **Kar** som kan "vara lånat

från ett semitiskt språk". Däremot inte de (lika?) gamla latinska lånorden *köpa*, *vin* och *öre*.

På det hela taget ter det sig emellertid ytterst rimligt att koncentrera framställningen på de germanska arvorden, inte bara för att släktskapsförhållandena är ojämförligt tydligast här utan också för att det ju är tal om en populärvetenskaplig framställning, där de "viktigaste svenska orden, det grundläggande ordförrådet, svenska språkets kärna" (förordet, s. 3) torde vara det som flest användare vill veta något om och där aha-upplevelserna av att det och detta är släkt med vartannat ger mest och bäst verkan. Givet detta blir man emellanåt en aning överraskad av att i ordboken finna ord med egna artiklar som inte på länge har tillhört detta grundläggande ordförråd. Man kan här ana författarens egen bakgrund som dialektolog och ortnamnsforskare. Det förra gäller artiklar som **Andor**, **Blägde**, **Drita**, **Fly 1** ('myr') och **Tåka**, det senare sådana som **Al 2** 'helig plats' (liksom det liktydiga **Vi**), **Ryd**, **Älmt** (med en utläggning om Älmtaryd och IKEA) och **Ör**, där den nu huvudsakliga förekomsten som ortnamnsled uttryckligen anges.

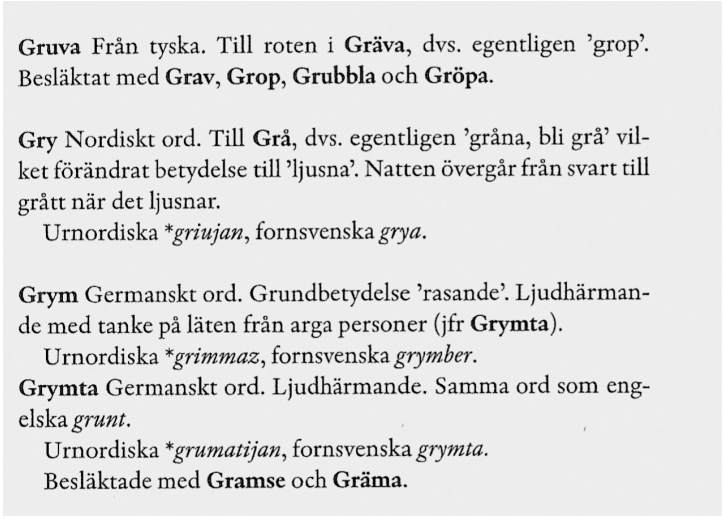
Å andra sidan saknas ur det grundläggande ordförrådet många av de rena formorden, till exempel av personliga pronomen alla utom *jag* och *ni*, bland basala prepositioner *i*, *med*, *under*, *ur* och *åt* (men inte *av*, *från*, *på*, *vid*, *över* och de grammatikaliserade *bland*, *mot*, *hos*, *till*), av konjunktioner *men* och *så* (men inte *eller*, *för*, *och* och *ty*), av adverb bland många andra *in* och *ut*, av räkneord alla utom *elva*, *tolv* och *hundra* och bland interjektioner *ja* men inte *nej*. Inget av de saknade formorden är väl särskilt spännande ur det sidledes släktskapets synvinkel, men kognater i andra germanska språk finns det – kanske likväl bedömda som alltför uppenbara för att förtjäna sitt utrymme?

3. Åtkomststruktur

Oaktat de synpunkter som framkommit i avsnitt 2 är det så att det allra mesta av språkets lexikala kärna (åtminstone av innehållsord) förvisso finns med, och dessa ord redovisas överlag på sin egen alfabetiska plats. I några fall finns det hänvisningar på alfabetisk plats till en annan artikel, där ordet det hänvisas ifrån nämns i förbigående samman med artikelns eget ord, till exempel **Backa** > **Bak**, **Bedrövad**[1] > **Drav**, **Marskalk** > **Märr** och **Skalk 1** samt **Örngott** > **Var 1** (nämnt som synonym och förklarat där). Annars avklaras alltså de olika med varandra besläktade orden på var sina egna alfabetiska platser, själva grundprincipen för framställningen, också om det innebär upprepningar på ömse håll. Sådan upprepning är förstås något som är till gagn för den som går till ordboken för att slå upp ett enda enskilt ord; oavsett vilket detta är i ett komplex av med varandra besläktade ord får man samma fullödiga besked om vilka andra ord som ingår i komplexet. Endast när ordfränder råkar hamna alldeles efter varandra i alfabetsräckan (och artiklarna då sammanhålls utan radmellanrum sinsemellan) ges uppräkningen av andra ordfränder samlat och sist efter de grafiskt sammanhållna artiklarna. Se figur 1 för exempel på den grafiska strukturen.

Beskedet om ordfränderna ges i form av en strikt alfabetiskt ordnad uppräkning. Särskilt långa sådana räckor hittar man till exempel i **Hall** och dess 19 fränder eller **Stå** och dess 21 fränder (också här nogsamt upprepade i alla de olika artiklarna). Om hur dessa fränder inbördes förhåller sig till varandra får man däremot inget veta av dessa uppräkningar. En ansats till en annorlunda lös-

1 Svenskans många lånade (eller inhemskt bildade) *be*-ord lyser nästan helt med sin frånvaro i ordurvalet. Utöver *bedrövad* finns bara *beläte* med, också som en hänvisning (till **Bild**); *begära* nämns i **Gärna** men har ingen hänvisning på alfabetisk plats. Det finns ytterligare ord som bara nämns inne i en annan artikel och väl med samma rätt kunde ha förtjänat hänvisningar, till exempel *skulor* (i **Skölja**), *varsam* (i **Varna**) och *ost* (i Öst).



Gruva Från tyska. Till roten i **Gräva**, dvs. egentligen 'grop'. Besläktat med **Grav**, **Grop**, **Grubbla** och **Gröpa**.

Gry Nordiskt ord. Till **Grå**, dvs. egentligen 'gråna, bli grå' vilket förändrat betydelse till 'ljusna'. Natten övergår från svart till grått när det ljusnar.
Urnordiska **griujan*, fornsvenska *grya*.

Grym Germanskt ord. Grundbetydelse 'rasande'. Ljudhärmande med tanke på läten från arga personer (jfr **Grymta**).
Urnordiska **grimmaz*, fornsvenska *grymber*.
Grymta Germanskt ord. Ljudhärmande. Samma ord som engelska *grunt*.
Urnordiska **grumatijan*, fornsvenska *grymta*.
Besläktade med **Gramse** och **Gräma**.

Figur 1: Utdrag från s. 115 i *Ordklok*.

ning återfinns emellertid i artikeln **Barn**, som förtjänar att citeras *in extenso*:

> **Barn** Germanskt ord. Till **Bära**, alltså egentligen 'det kvinnan bär på (under havandeskapet)'. Nära anknutet är ordet **Börd**, egentligen 'bärande (av barn under havandeskapet)' sedan utvecklat till 'födsel' och vidare till 'anor, härstamning'. Till **Bära** har även bildats orden **Bår** 'något man bär med' och **Börda** 'något man bär'. Till samma rot hör också orden **Böra**, **Bördig** och **Börja**.
> Urnordiska **barna*, fornsvenska *barn*.

Här utgör alltså uppslagsordet en ingång till en utredning av rotens formella och semantiska förgreningar åt olika (om än ej alla olika) håll. Det här upplägget ger, åtminstone mig, klar mersmak, och jag inspirerades därav att leka med tanken på hur det skulle ha kunnat vidareutvecklas till ett helt annat sätt att organisera

åtkomststrukturen och därmed framställningen: att alla de besläktade orden (utom ett) enbart fått hänvisningar på sin alfabetiska plats till en och samma artikel som har det semantiskt mest primära ordet som ingång (i detta fallet **Bära**) och där sedan alla de andra också får sin beskrivning i en utredning i stil med, men alltså fullständigare än den i **Barn**. (Det kan nämnas att artiklarna **Bår**, **Börd** och **Börda** inte är hänvisningsartiklar utan upprepar de för vardera ordet specifika egentliga betydelserna.) En sådan struktur skulle förvisso ha försvårat för den som nu lätt kan hitta upplysningar om just ”sitt” ord utan att behöva gräva fram dem i en längre artikel, men det väsentligt större utrymme för mer resonerande och parallelliserande framställning som nu enbart finns i just artikeln **Barn** hade kunnat göra de olika ordens inbördes släktskap tydligare – och, tror jag, mer läslockande. Men känslan av traditionellt ordnad ordbok hade rimligen försvagats, om lemmalistan med ens skulle bestå av mest hänvisningar.

4. Artikelinnehåll

Artiklarna i *Ordklok* anger inledningsvis om ordet är svenskt, nordiskt eller germanskt och sedan huruvida det är lånat genom att i så fall meddela från vilket språk det lånats (för det allra mesta tyska, som anges utan åtskillnad mellan låg- och högtyska). Därnäst anges vilken grundbetydelse ordets rot har haft (eller till vilket mer primärt ord det är bildat), och – uttömmande – vilka andra ord i ordboken som är besläktade med det. Allra sist anges för arvord vanligtvis urnordisk (vanligen rekonstruerad) och fornsvensk (vanligen belagd) form. Motsatt traditionella etymologiska ordböcker är *Ordklok* inte särskilt upptagen av parallellord i andra språk (vare sig samtida eller äldre), förutom att artiklarna ofta anger eventuella nuengelska kognater. Tanken torde här vara den att engelskan är det ojämförligt bäst (och kanske det enda)

kända främmande språket bland ordbokens läsare, och det är helt visst en riktig tanke att de engelska parallellerna såsom välkända bidrar till de aha-upplevelser som förordet (s. 7) utlovar, på ett sätt som tyska paralleller inte (längre) skulle göra.[2] Det engelska parallellspåret stannar inte enbart vid dessa kognater, utan i många fall anförs också kort etymologi för helt obesläktade semantiska ekvivalenter till artikelns svenska ord, till exempel:

> **Anka** […] Engelska *duck* är bildat till verbet *duck* som är besläktat med **Dyka** (se även **Ducka**). *Duck* betyder alltså egentligen 'den som dyker'.

> **Lax** […] Engelska *salmon* från franska (ytterst från latin).

På **Äga** anförs engelska *own* av samma semantiska skäl – för att sedan återkomma i **Ägna** som dess formella kognat.

Detta återkommande utrymme för engelska paralleller (med sidospår) är ett av ordbokens särpräglande drag. Ett annat, än mer centralt har jag redan nämnt flera gånger: angivelsen av alla ord i ordboken som är besläktade med varandra. Släktskap är förstås viktigt i all etymologisk framställning, men i traditionella etymologier anges den vanligtvis som enskilda länkar i ibland långa kedjor: ord A kan till exempel anges vara avlett av ord B, som man sedan får gå vidare till för att läsa att det i sin tur kanske står i avljudsförhållande till ord C, utifrån vilket senare man sedan får nysta vidare. I och med att *Ordklok* istället ger alla de samhöriga orden i varje till komplexet hörande artikel får man med ens syn på ordfränder långt i sidled, inte bara förälder och/eller syskon utan också kusiner, nästkusiner och än avlägsnare släktförbindelser. Fastän en mer kvalificerad användare av ordboken nog kan vara latent medveten om dessa, kan den tydliga exponeringen av dem helt visst ofta fungera som ögonöppnare.

2 I artikeln **Hult** anförs dock tyskans *holz* i brist på engelsk kognat.

En tredje specialitet i *Ordklok* är alltså att urnordisk form ges för det övervägande flertalet arvord. Också fornsvensk form ges som sagt, men det är ju en normalitet i svenska etymologier, om än den rekonstruerade fornsvenska form som *Ordklok* ger när medeltida belägg saknas (till exempel **kuker* på **Kuk**) inte är det (vanligen en form från svenska dialekter istället). Om man ser till den breda allmänhet som utpekas som ordbokens målgrupp, ter sig kanhända valet att ange urnordiska, och därtill vanligtvis rekonstruerad urnordiska, som något iögonfallande. Och inslaget kan nog också vara en aning förbryllande för en och annan läsare. Så kan formerna på gamla svaga maskuliner som **Spjäla** och **Vana** ge intryck av en obeslutsamt vinglig formhistoria: från urnordiska **wana* till fornsvenska *vani* till nutidens *vana*, som en troskyldig läsare skulle kunna tolka som en urnordisk revansch. Urnordiskan ges dessutom med ursprungligt *-z* istället för *-ʀ*, som nog gör formerna än mer främmande; detta används också i former som inte är rekonstruerade, till exempel *winiz* på **Vän 1**. Ordbokens urnordiska är ändå en finess, eftersom detta därmed gör en i grunden populariserad etymologisk redovisning relevant också för språkvetare. Som klassiskt skolad nordist, dessutom specifikt ljudhistoriker och ortnamnsforskare, är författaren fena på urnordiska ljudstrukturer och stammar, och man kan ana att han funnit ett särskilt nöje i att knåpa med dessa rekonstruerade former. Också om dessa alltså torde ha fallit naturligt ut under författarens vana hand, är insatsen att ha lagt ett så omfattande och detaljerat urnordiskt pussel imponerande nog – och som sagt till stor glädje och nytta för den inte lika klassiskt skolade nordisten utan samma fingerfärdighet i rekonstruktionens formsträngа konst. Man kan i detta sammanhang för övrigt notera att etiketterna ”svenskt ord” respektive ”nordiskt ord” inte är ett bud på i vilket språk ordet bör vara bildat utan bara på var det faktiskt är belagt; också de angivet svenska orden **Hynda**, **Klösa** och **Mylla** får ansatt varsin urnordisk form (utan det *i*-omljud

som gör det sannolikt att de verkligen uppstått i samnordisk tid).

Överlag har alltså artiklarna en rätt ensartad och stram innehållsstruktur, men i några få fall ger författaren utrymme åt något längre kulturhistoriska utvikningar, till exempel:

> **And 1** […] I danskan heter *Kalle Anka* i stället *Anders And*.

> **Hjon** […] Genom sammansättningen *fattighjon* 'någon som är inhyst i fattigstuga' kom ordet att ges negativa associationer och närmast att bli ett skällsord. Det gick så långt att till och med mansnamnet *Jon* – som ännu vid 1800-talets början var ett av de vanligaste namnen i Sverige – började minska kraftigt i popularitet och nästan försvann, eftersom det uttalades likadant som *hjon*.

> **Vase** […] Kungaätten *Vasa* har troligen sitt namn efter släktgodset Vasa i Uppland, och till ättens vapensköld valdes en sädeskärve, en vase (i svealändsk dialekt *vasa*, jfr en *bulla* 'en bulle') eftersom ordet råkade sammanfalla med godsets namn.

Ibland förekommer tillika mer lättsamma utvikningar, som på **Tappa** ("Dagens Nyheter hade en gång rubriken 'S A S tappar tunga kunder'").

Artiklarna anger normalt en grundbetydelse (vanligen för roten), och betydelseutvecklingen därur ges gärna mer resonemangsvis, till exempel:

> **Bry** […] Till **Brud** med grundbetydelse 'älska med bruden på bröllopsnatten', vilket utvecklats till 'sexuellt antasta', vidare till allmännare 'plåga, genera' och slutligen till 'bekymra sig för, bry sig om'.

> **Gemen** […] Från grundbetydelsen 'gemensam' har en lång kedja av betydelser utvecklats efter varandra: 'allmän', 'vanlig', 'enkel', 'simpel', 'dålig', 'elak'. Det som är gemensamt är allmänt, det som är allmänt är vanligt, det som är vanligt är enkelt, det som är enkelt är simpelt, det som är simpelt är dåligt eller elakt.
>
> **Skugga** […] Till en rot som betyder 'se, skåda', dvs. 'syn, något man ser' som har utvecklats till 'skugga' och (i flera germanska språk även) 'spegel'.

Semantiska reflexioner helt vid sidan av det etymologiska förekommer understundom också, som på **Hyfsa**: "*Hyfsad* har kommit att betyda 'ganska bra', men används numera rätt ofta som en underdrift i betydelsen 'jättebra'".

Genomgående gäller emellertid att ordens formsida inte ägnas någon uttrycklig utredning. Till exempel sägs **Arbete** och **Arvode** vara samma ord och läsaren lämnas att själv inse hur *b* och *v* kunnat uppstå ur (det frikativa *b* som skrivs) *b* i den urnordiska rekonstruktionen; på **Bjärt** nämns engelskans *bright* som samma ord utan någon formförklaring; på **Trana** och **Tugga** lämnas läsaren att själv klura på hur urnordiskt *k-* blivit till svenskt *t-*. På **Akta** sägs förvisso att en "enkel tumregel är att de flesta ord som har konsonantkombinationen *kt* är lånade från tyska", men en sådan allmän synpunkt blir ju (utom för den som råkar vilja slå upp just *akta*) bara tillgänglig för sträckläsare. Det hade nog varit välgörande med några fler tumregler för återkommande ljudhistoriska samband, men då placerade i inledningen. Nu finns det i förordet (s. 5 f.) enbart en liten uttalsnyckel till de urnordiska och fornsvenska formerna.

Själva den etymologiska kärnan i artiklarna är, förutom angivelsen av en grundbetydelse, släktskapen med andra svenska ord. Detta anges i slutet av artikeln, före de fornspråkliga formerna,

och företrädesvis med formeln ”Besläktat med”, om än enstaka alternativa men av allt att döma helt likvärdiga uttryckssätt också kan förekomma, då inne i texten, till exempel ”Se även **Brudgum** och **Bry**” (på **Bröllop**) och ”Jfr **Ihjäl**” (på **Helvete**). Normalt listas som redan sagts de besläktade orden rakt upp och ned, i bokstavsordning, men någon enstaka gång förekommer en gruppering – förvisso sakligt välmotiverad men inte så uppenbart mer här än på månget annat ställe – som på **Hjärna**: ”Nära besläktat med **Hjässa** samt avlägsnare med **Hjort**, **Hjortron**, **Horn**, **Hörn** och **Ren 2**”, eller **Spad**: ”Besläktat med **Spä** och **Späd** liksom med **Spann**, **Spindel**, **Spinna**, **Spång** och **Spänstig**”. I fråga om närhet anges förstås avledning för sig, som en inledande del av etymologin, åtskilt från räckan med andra besläktade ord, till exempel på **Jäst**: ”Till **Jäsa**”. Det gängse uttryckssättet är här just detta, med *till*, och möjligen är det något som också kunde ha krävt några förklarande ord i inledningen, eller med andra ord något om hur ordbildning normalt går till.

Inte så sällan modifieras sannolikheten (”sannolikt”/”troligen”/”kanske” med mera) för en inledningsvis angiven förbindelse med ett annat ord, men denna mån av tveksamhet i släktskapen hindrar inte att detta senare ords fränder sedan räknas upp efter slutformeln ”Besläktat med” men nu utan den inledande reservationen. Detta kan väl vara en godtagbar förenkling inom ramen för det kortfattade artikelformatet, men jag hade gärna istället sett en generell sådan tillämpning av upprepad reservation som undantagsvis förekommer och inte inbjuder till missförstånd, till exempel:

> **Dregla** […] Kanske till samma rot med betydelsen ’rinna, droppa’ som i **Droppe** och **Drypa**. I så fall besläktat även med **Drabba**, **Dryfta**, **Dråp**, **Dråplig**, **Dråsa**, **Drälla**, **Dräpa** och **Träffa**.

På **Mård** heter det att man ”föreslagit samband med **Mord** (’det mordiska djuret’) men det är mycket osäkert”, och här nämns överhuvudtaget inte **Mård** på **Mord**. Samma återhållsamhet råder inte på **Fläsk** och andra ordfränder till **Flik**; dessa anför inte bara släktskapen med **Flik** utan också, utan reservation, till **Flicka**, fastän detta senare ord bara är ”[k]anske besläktat med **Flik**”.

Däremot påtalas det återkommande, på vällovligt pedagogiskt vis, när ord man kanhända skulle kunna tro är släkt inte alls är det, som till exempel:

> **Meja** […] *Dagsmeja* hör inte hit utan innehåller ett ord som betyder ’kraft, styrka’ (besläktat med **Makt**) med syftning på solens påverkan på snön på dagen under våren.

När besläktade ord i svenska (eller engelska) saknas, blir etymologin i regel mycket kort och stannar vid en upplysning om grundbetydelsen, om att ordet är ljudhärmande eller om att ursprunget är oklart eller liknande, till exempel:

> **Lön** Germanskt ord. Grundbetydelse ’pris, belöning’.
> Urnordiska **launo*, fornsvenska *løn*.
>
> **Gnida** Germanskt ord. Ljudhärmande.
> Urnordiska **gnidan*, fornsvenska *gnidha*.
>
> **Värre/Värst** Germanska ord. Oklar etymologi.

Emellanåt lämnas alltså etymologin öppen, antingen helt och hållet som i sist anförda exempel eller med endast ett ”kanske” som i ett par tidigare exempel. I några fall ges alternativ, till exempel på **Torg** (slaviskt ord lånat i nordiska eller tvärtom). Vad gäller ett klassiskt tvisteämne som **Viking**, där alternativa förslag verkligen inte saknas, har författaren emellertid tagit ställning. Det är han

ju ingalunda ensam om, men det blir missvisande för den häri oinsatta när det inget sägs om att andra auktoriteter skulle anföra helt andra förklaringar som de troliga:

> Troligen till **Vika** i en betydelse ’fara bort, resa ut’ som är belagd i fornisländska. Vikingarna var alltså de som reste ut på havet på vikingafärd, inte de som stannade hemma.
>
> Urnordiska **wikingaz*, fornsvenska **vikinger*.

På andra ord återfinns formeln ”Oklar etymologi”, och den hade varit tacknämlig också här. Just här och just med den föredragna etymologin blir dessutom den rekonstruerade urnordiska formen något problematisk; med den grundbetydelsen bör väl ordet ha bildats först under vikingatiden, efter urnordisk tid. Ett annat i den nordiska kulturhistorien centralt ord med omtvistad etymologi, *jul*, har däremot lämnats oförklarat: ”Betydde före kristnandet ’midvinterfest’. Oklar etymologi”. Författaren har sålunda inte nappat på Bjorvands och Lindemans (2019, s. 606 ff.) förslag att det hör samman med *jaga* (och ursprungligen avsett kappridning som ett inslag i midvintersederna).

För de angivet ljudhärmande orden meddelas det i flera fall att de samtidigt är besläktade med andra ord, och man blir då en aning förbryllad; så sägs till exempel vara fallet med *knäcka*, besläktat med *knacka* och *knaka* och med en grundbetydelse ’få något att knäckas’. Detta senare låter som en kausativbildning, och i så fall är det ju bara avledningsbasen som är ljudhärmande? Visst kan väl också avledningen då ytterst sägas vara ljudhärmande, men en mindre sparsmakad och mer klargörande formulering hade ändå varit bra.

För lånorden anges det långivande språket efter ett ”Från”, men någon åtskillnad mellan låg- och högtyska görs som sagt inte, så lånets ålder får läsaren föga besked om. För de tyska lånorden gör detta kanske inte så mycket ändå, eftersom tyskans långivande

kraft kan ses som ett i stort sett sammanhållet, låt vara mångahundraårigt men nu helt avklingat, skede av tyskt kulturinflytande. Mer påfallande blir det att riktigt gamla fornengelska och nära på 1000 år yngre, nästan nutida engelska lån faller under samma etikett ”Från engelska”, *skrud* likaväl som *tips*. Nu ger sig förstås ordboken inte ut för att vara annat än etymologisk, inte i vidare mening ordhistorisk, så det är nu inte preciserade tidfästningar av lånen som jag efterlyser. Däremot hade det kunnat vara diskret upplysande, inom ramen för det redan befintliga formatet, om de medeltida lånorden förslagsvis fått fornsvensk form.

Att många av de tyska lånorden har nordiska arvord bland sina ordfränder ligger ju snarast i sakens natur och är återigen en sådan insikt som man nog ”egentligen” redan har men som den uttömmande uppräkningen gör uppenbar. Också i annat tonas tyskheten ned. Betydelselån anges sålunda högst diskret om alls, som på **Handla**, som upptas som rent arvord och där betydelsen ’köpa’ bara anges som den sista i betydelseutvecklingen, eller på **Hov**, där det bara nämns inom parentes. Att **Strumpa** framställs vara inhemskt är väl däremot bara en miss; någon urnordisk eller fornsvensk form anförs heller inte. I några fall blir det en aning förvirrande när det som sägs vara ett arvord anges höra samman med ett ord som i sin tur sägs vara lånat från tyska, exempelvis arvordet **Slippa** föregivet bildat till lånordet **Slipa** eller, omvänt, lånordet **Släp** föregivet bildat till arvordet **Släppa**. Också i dessa fall hade de kortfattade formlerna varit betjänta av något mångordigare utvecklingar.

5. Sammanfattande omdöme

En ny etymologisk ordbok på svenska är som helhet, som inledningsvis sades, en stor händelse. Sedan består en ordbok av ett sådant myller av enskildheter att det nästan ofrånkomligen inbju-

der till mer eller mindre kritiska synpunkter på denna detaljnivå. Det ligger därför nära till hands att en ordboksgranskning som denna kommer att upptas av sådana och därmed lätt kan synas väl mycket präglad av just enskildheter, på bekostnad av mer genomgripande synpunkter. Jag vill ändå hävda att några av de synpunkter som framkommit ovan också synliggjort mer principiella sidor av *Ordklok*:s framställning där man kunde ha önskat sig denna på ett delvis annat sätt. Ibland kunde den ha omprioriterat ordurvalet (gärna färre dialektala ord till förmån för andra, mer centrala ord), ibland varit tydligare (såsom i gärna upprepad tveksamhet kring ordfränderna), ibland varit upplysande kompletterad (såsom gärna med fornsvensk form hos gamla lånord). Sådant hade ännu bättre tjänat dess uttalade syfte att vara lättbegriplig för en så bred allmänhet som möjligt. Jag tror att detta också gäller för det ojämförligt mest genomgripande uppslaget, att samla alla de besläktade orden i en sammanhållen artikel kring det semantiskt mest basala ordet med hänvisningar dit från de andra orden – ett mitt ändringsförslag som författaren i viss mån får skylla sig själv för, eftersom han givit ett så smakligt prov på denna ansats i artikeln **Barn** (se citatet ovan i avsnitt 3).

Emellertid må icke dessa mer principiella invändningar skymma det förhållande att *Ordklok* precis som den är utgör en mycket välfungerande populärt framställd etymologisk ordbok, framförallt genom sitt konsekvent genomförda, i sig enkla men verkningsfullt ögonöppnande grepp att ange alla besläktade ord på varje ifrågakommande ord, att med andra ord satsa på en beskrivning mer å sido än till rygga i ätten än vad etymologier traditionellt brukar göra. Det är högst sannolikt att detta bäst fångar det som folk flest är etymologiskt hågade av, och det är vackert så. Till detta kommer så att ordboken dels genom detta samma grepp som bekvämt synliggör något som bara stegvis kan grävas fram ur andra etymologier (om man nu inte går till de allindoeuropeiska som man återfinner hos Pokorny (1959 och senare)),

dels genom sin nästan lika konsekventa angivelse av rekonstruerad urnordisk form gör sig relevant också för mer kvalificerade användare – utöver att den för dessa som för alla andra nu utgör det allra färskaste svenska budet på en uppdaterad etymologisk syntes.

Litteratur

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

Bergman, Bo (2007): *Ordens ursprung. Etymologisk ordbok över 2 200 ord och uttryck*. Stockholm: Wahlström & Widstrand.

Bergman, Gösta ([1966/1981,] 1990): *Ord med historia*. Samlingsvolym av *Ord med historia*, femte, avsevärt utökade uppl. (1977) och *Nya ord med historia* (1981). Stockholm: Prisma.

Bergman, Olle (2011): *Krigiska ord*. Lund: Historiska media.

Bjorvand, Harald & Fredrik Otto Lindeman ([2000,] 2019): *Våre arveord. Etymologisk ordbok*. Tredje utg. Oslo: Novus.

Ernby, Birgitta (2008): *Norstedts etymologiska ordbok*. Stockholm: Norstedts Akademiska Förlag.

Falk, Hjalmar & Alf Torp (1903–1906): *Etymologisk ordbog over det norske og danske sprog*. Kristiania: H. Aschehoug & Co.

Falk, H. S. & Alf Torp ([1910–1911,] 1960): *Norwegisch-Dänisches etymologisches Wörterbuch*. Zweite Aufl. Oslo & Bergen: Universitetsforlaget.

Hellquist, Elof ([1922,] 1948): *Svensk etymologisk ordbok*. Tredje uppl. Lund: C.W.K. Gleerup.

Levander, Lars & Elias Wessén (1932): *Våra ord. Deras uttal och ursprung. Populär etymologisk ordbok*. Stockholm.

Nielsen, Niels Åge ([1966,] 1989): *Dansk Etymologisk Ordbog. Ordenes Historie*. 4. udg. København: Gyldendal.

Odhner, Einar ([1952,] 1967): *Etymologisk ordlista*. Andra omarbetade uppl. Stockholm: Liber.

Pokorny, Julius ([1959,] 2002): *Indogermanisches etymologisches Wörterbuch*. 4. Aufl. Tübingen & Basel: Francke.

Politikens Etymologisk Ordbog. Af Jan Katlev. København: Politiken 2000.

Torp, Alf (1919): *Nynorsk etymologisk ordbok*. Kristiania: H. Aschehoug & Co.

Wessén, Elias ([1961,] 1973): *Våra ord. Deras uttal och ursprung. Kortfattad etymologisk ordbok*. Andra, tillökade uppl. Stockholm: Nämnden för svensk språkvård.

de Vries, Jan ([1961,] 1962): *Altnordisches etymologisches Wörterbuch. Mit Literaturnachweisen strittiger Etymologien sowie deutschem und altnordischem Wörterverzeichnis*. Zweite verbesserte Aufl. Leiden: E. J. Brill.

Bo-A. Wendt
Huvudredaktör för SAOB, docent i nordiska språk
Svenska Akademiens ordboksredaktion
Dalbyvägen 3
S-224 60 Lund
Bo.Wendt@svenskaakademien.se

MEDDELANDEN

Rapport fra styret for Nordisk forening for leksikografi

Hanne Lauvstad

Vitenskapelig

Nordisk forening for leksikografi (NFL) arrangerer symposier hvert år og konferanser annethvert år. Foredragene fra symposiet utgis i årsskriftet *LexicoNordica* (LN). Symposiene samler en forholdsvis liten gruppe deltagere, mange spesielt invitert, til å holde foredrag om et bestemt tema. Foredragene og de faglige diskusjonene gir mulighet for fordypning i leksikografifaglige emner. Hittil har symposiene vært lagt til begynnelsen av året, i lokalene til Fondet for dansk-norsk samarbeid, enten på Lysebu i Oslo eller på Schæffergården utenfor København. I 2022 er Schæffergården solgt, og Fondet har dermed mindre kapasitet til å huse symposiene. De vil derfor måtte arrangeres også andre steder enn på Lysebu i fremtiden.

Det 29. *LexicoNordica*-symposiet 2022 hadde tittelen «Nordisk leksikografi – nu og i fremtiden». Det ble arrangert som en hybridkonferanse (fysisk og digitalt) 10.–12. februar 2022 på Höllviksnäs kurssted i Skåne. Bidragene publiseres i dette nummeret av *LexicoNordica*.

Konferansene henvender seg til en større gruppe, og målet er å samle forskere og andre leksikografiinteresserte fra hele Norden. På konferansene behandles fastsatte temaer, men det er også plass til rapporter fra pågående ordboksprosjekter og andre emner som de påmeldte foredragsholderne ønsker å ta opp. Det er også mulighet for posterpresentasjoner. De nordiske landene er etter tur vertskap for konferansene, som arrangeres av ett eller flere fagmil-

jøer. En rekke bidrag fra konferansene publiseres i NFLs skriftserie *Nordiske studier i leksikografi*.

NFL-konferansen som skulle ha vært arrangert i Lund i 2021, måtte utsettes ett år pga. covid-19-pandemien med smittevernbegrensninger og reiserestriksjoner mellom landene. 27.–29. april 2022 var det endelig klart for Den 16. konferansen om leksikografi i Norden. Temaet var «Lexikografiska utmaningar». Arrangementet var et samarbeidsprosjekt mellom Svenska Akademiens ordboksredaksjon i Lund og Institutionen för svenska språket ved Göteborgs universitet. Kvelden før konferansen ble det holdt en mottagelse i lokalene til ordbordboksredaksjonen ved Svenska Akademiens Ordbok (SAOB). Styremøte for NFL ble holdt samme sted før mottagelsen. Konferansen ble en forsinket feiring av 30-årsjubileet for NFL-konferansene, idet den første konferansen ble holdt ved Universitetet i Oslo i 1991. Etter konferansen har man kunnet sende inn sitt innlegg til fagfellevurdering, og antatte bidrag vil bli publisert i skriftserien *Nordiske studier i leksikografi*, i samarbeid med *Meijerbergs arkiv för svensk ordforskning*.

Siden Lund-konferansen var forskjøvet, blir det bare ett år mellom den og neste konferanse. 24.–26. mai 2023 arrangerer Universitetet i Bergen (UiB), i samarbeid med redaksjonen for Det Norske Akademis ordbok (NAOB), den 17. konferansen om leksikografi i Norden. Planleggingen av konferansen er i god gjenge. Hovedtemaet er «Det nye og det gamle i ordbøkene». I det ligger alt som gjelder språklig opphav og endring, f.eks. etymologi, lån, nyord, ord som er gått ut av bruk og semantisk og grammatisk variasjon. Etter innspill er det åpnet for tematisk organisering av foredrag i sesjoner: én om den leksikografiske arven etter Ivar Aasen og en annen om nyord i Norden og hvordan disse blir behandlet fra de blir registrert til de er på plass i ordbøkene. Det er invitert tre plenumsforelesere: direktør Åse Wetås, Språkrådet, hovedredaktør Bo Wendt, SAOB og professor/direktør Alexander

Geyken, *Digitales Wörterbuch der deutschen Sprache*, Berlin. Stedet for konferansen blir Hotel Terminus i Bergen sentrum.

Referat fra NFLs generalforsamling og informasjon om kommende konferanser og annet kan leses på NFLs hjemmeside: <nordisk-leksikografi.com/> og på NFLs Facebook-side (søk på «Nordiska Föreningen för Lexikografi – NFL»). Samtlige numre av *LexicoNordica* og *Nordiske studier i leksikografi* blir publisert digitalt på <tidsskrift.dk> i samarbeid med Dansk Sprognævn. En viktig ressurs for leksikografer, en digital utgave av standardverket Nordisk leksikografisk ordbok (NLO, 1997) foreligger i pdf-format på NFLs hjemmeside: <nordisk-leksikografi.com/publikationer-1.html>.

Administrasjon

NFL-styret har ikke lenger noen ekstern administrator. Oppgavene fordeles isteden mellom styremedlemmene og LNs hovedredaktører. Systemet for medlemsbetaling er forenklet og digitalisert, slik at hvert medlem betaler direkte via NFLs hjemmeside, og medlemmene er registrert i et felles medlemsregister. NFL har egne bankkonti (en bruks- og en sparekonto). Man kan kontakte styret direkte via e-postadressen: nordisk.lexikografi@gmail.com. Neste generalforsamling vil finne sted ved konferansen i Bergen. Da vil det være valg til ledige styreverv.

Økonomi

NFL har fortsatt en solid økonomi, med midler til dekning av utforutsette utgifter. Men foreningen er fortsatt avhengig av eksterne økonomiske bidrag til mer kostnadskrevende tiltak, som symposiene og konferansene. NFL har fått bevilget støtte fra Nordplus

Nordens språk, både til symposiet og til konferansen 2023. Språkrådet (i Norge) har bevilget midler til publikasjon av konferanserapporten. I tillegg bidrar vertsinstitusjonene med arbeidstid og eventuelt andre ressurser.

Takk

NFL takker for alle eksterne økonomiske bidrag til symposier, konferanser og publisering av konferanserapporter, for uten den økonomiske støtten hadde det ikke vært mulig å gjennomføre denne typen virksomhet.

NFLs styre 2021–2023

Hanne Lauvstad
hovedredaktør, dr.art.
Det Norske Akademis ordbok
Grensen 3
NO-0159 Oslo
hanne.lauvstad@naob.no

REDAKTIONSANVISNINGAR

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, fransk eller tysk kan også optages hvis særlige forhold taler for det.

2. **Bidrag** sendes til det medlem af redaktionskomitéen som repræsenterer bidragyderens land:

   - Danmark: Anna Braasch, Institut for Nordiske Studier og Sprogvidenskab (NorS)/Center for Sprogteknologi, Københavns Universitet, Emil Holms Kanal 2, Bygning 22, 3., DK-2300 København S. <braasch@hum.ku.dk>.
   - Norge: Kjetil Gundersen, Erika Nissens gate 7, NO-0480 Oslo. <kjetil.gundersen@sprakradet.no>.
   - Sverige: Lennart Larsson, Stormgatan 19, SE-754 31 Uppsala. <lennart.larsson@nordiska.uu.se>.
   - Finland: Harry Lönnroth, Institutionen för språk- och kommunikationsstudier, PB 35, FI-40014 Jyväskylä universitet. <harry.h.lonnroth@jyu.fi>.
   - Island: Ásta Svavarsdóttir, Árni Magnússon-instituttet for islandske studier, Laugavegur 13, IS-101 Reykjavík. <asta.svavarsdottir@arnastofnun.is / asta@hi.is>.

   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. **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 2009", "Gudiksen, Hovmark & Monka 2009". For kilder med fire eller flere forfattere anføres kun det første efternavn efterfulgt af "et al.", fx "Gudiksen et al. 2009". Forfatternavnene adskilles af komma, undtagen de to sidste navne, som adskilles af "&" (tilsvarende hvis der kun er to forfatternavne).

   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)". Der anvendes så vidt muligt

kun permanente internetadresser. 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 bør naturligvis 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 omvendt 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).

NRG = *Norsk referansegrammatikk*, se Faarlund et al. (1997).

Zgusta, Ladislav (1971): *Manual of lexicography*. The Hague: Mouton.

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.