NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Data og repræsentativitet i ordbogsarbejdet

Forfatter: Henrik Hovmark

Kilde: Nordiska Studier i Lexikografi 11, 2012, s. 296-308
Rapport från Konferens om lexikografi i Norden, Lund 24.-27. maj 2011

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive

Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Data og repræsentativitet i ordbogsarbejdet

Henrik Hovmark

Nowadays, lexicography is closely associated with information technology and language technology tools. However, until fairly recently lexicography was based on traditional printing and collections of paper slips. Taking a description of the working processes in lexicographic work based on collections of paper slips, I discuss the relations between empirical data, work flow and final product in lexicography in general. I focus on the question of representativeness: When is a collection of paper slips representative? How and to what extent is it possible to make generalizations based on data which is highly qualitative? I point to the fact that any kind of data, text corpora as well as collections of paper slips, is the outcome of a selection and should be scrutinized critically by lexicographers.

Jeg arbejder til daglig på Ømålsordbogen (ØMO), en ordbog over dialekterne på Sjælland, Lolland-Falster, Fyn og omliggende øer i perioden 1750-1945. Indsamlingen til ØMO startede i 1909 (jf. Gudiksen & Hovmark 2009), og kildegrundlaget er en seddelsamling på ca. 3 mio. sedler. Vi benytter også et mindre korpus af udskrevne båndoptagelser på ca. 1,3 mio. løbende ord som et yderst vigtigt supplement (jf. Gudiksen & Hovmark 2008), men seddelsamlingen er det primære udgangspunkt.[1]

Når jeg og mine kolleger i dag fortæller om hvordan vi henter sedler frem fra seddelkasserne, forsøger at tyde karakteristiske håndskrifter, og sorterer sedlerne i bunker på skrivebordet, er reaktionen ofte at det må være et spændende og fascinerende arbejde. Men undertiden kan man også notere en vis undren: Hvorfor arbejder man ikke med et tekstkorpus? Og er håndskrevne sedler, i visse tilfælde kun en håndfuld eller færre, virkelig tilstrækkelige til at udarbejde en hel ordbogsartikel? Reaktionen er interessant og vil danne udgangspunkt for den følgende artikel. Først og fremmest gør reaktionen opmærksom på et vigtigt aspekt ved ordbogsarbejdet, nemlig omgangen med kilder og det empiriske

[1] Jeg har tidligere arbejdet som redaktør ved Den Danske Ordbog (DDO), der jo bygger på et tekstkorpus på ca. 40 mio. ord, jf. DDO, bind 1: 54.

grundlag for de ordbøger der produceres. Men den vidner også om hvor gennemgribende en forandring hele ordbogsfeltet har gennemgået siden it-teknologien for alvor holdt sit indtog i 1980'erne og begyndte at ændre datagrundlag og arbejdsmetoder væsentligt. Til trods for at seddelsamlinger – og den dataindsamlingsmetode der ligger bag, nemlig excerpering – indtil for få år siden var enerådende inden for leksikografien og stadig danner grundlag for en række ordbogsprojekter, er der mange leksikografer der i dag ikke har kendskab til seddelsamlingernes virkelighed: deres tilblivelse og indhold, og arbejdet med dem. Måske kender man dem kun fra Simon Winchester bog om tilblivelsen af Oxford English Dictionary: *The Meaning of Everything* (Winchester 2003).

Jeg vil i denne artikel åbne døren ind til seddelsamlingerne og støve nogle af sedlerne af. Formålet er imidlertid ikke primært at fortælle en historie om gamle dage. Det er derimod min hensigt at bruge eksemplerne fra seddelsamlingerne til at kaste et blik på et mere generelt leksikografisk spørgsmål, nemlig spørgsmålet om data og repræsentativitet i ordbogsarbejdet, et spørgsmål som ordbogsredaktører må forholde sig til uanset om de data der arbejdes med foreligger i form af en seddelsamling eller et tekstkorpus – eller evt. begge dele, sådan som det er tilfældet for fx ØMO. Jeg vil hente mine eksempler fra ØMO's seddelsamling, og på den måde kan artiklen forhåbentlig samtidig tjene til at give et indtryk af seddelsamlinger og arbejdet med dem.[2]

1. Seddelsamlinger og korpus som data

Der er ikke nogen tvivl om at seddelsamlinger i en vis forstand er forældede, og at fremkomsten af tekstkorpora er et uhyre stort fremskridt i ordbogsarbejdet. Man fornemmer en næsten beruset begejstring for de nye muligheder i it-teknologien i den første udgave af Svenséns *Handbok i lexikografi* fra 1987:

> … datorn [har] revolutionerat själva det lexikografiska arbetet: den låter oss göra saker med materialet som vi tidigare inte ens kunnat föreställa oss som möjliga, och den tvingar oss att på ett mycket välgörande sätt ifrågasätta och revidera mycket av det vi ansett som en gång för alla givet. (Svensén 1987: 243)

Begejstringen bliver ikke mindre i den reviderede og udvidede udgave fra 2004, hvor Svensén skriver følgende om tekstkorpora og brugen af dem:

---

[2] Denne artikel bygger på mit foredrag ved leksikografikonferencen i Lund, men af pladsmæssige hensyn vil fokus ligge på den del af foredraget der handlede om data og repræsentativitet. Diskussionen af termen korpus, herunder om en seddelsamling kan kaldes et korpus, berøres kun indirekte, og eksempelmaterialet er reduceret.

> Det stora utbudet av texter i elektronisk form och möjligheten att lagra och analysera stora textmängder på ett ekonomiskt sätt har revolutionerat inte bara lingvistiken utan också lexikografin. Stora elektroniska korpusar har givit lexikografer möjlighet at fatta beslut baserade på språket i verkligt bruk snarare än på intuition och användning av sekundärkällor. (Svensén 2004: 58)

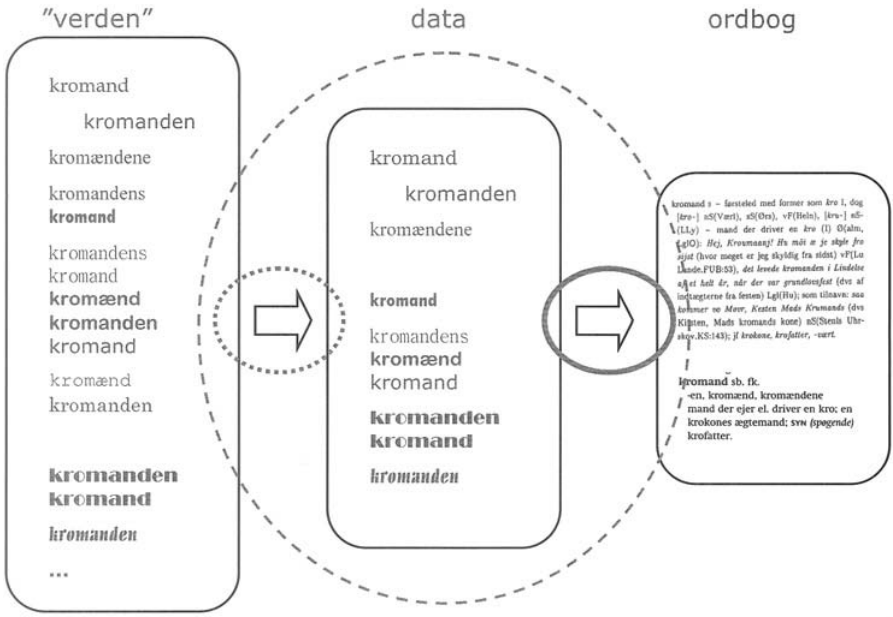
For ubefæstede sjæle kan Svenséns beskrivelse forstås på den måde at seddelsamlinger er baseret på intuition eller sekundærkilder. Det er næppe meningen – Svensén er selv kritisk over for en unuanceret tilgang til (brugen af) tekstkorpora i både positiv og negativ retning, jf. afsnit 3.1.3.8 (Korpusskeptiker och korpusfundamentalister) i Svenséns bog, og følgende opsummering:

> Att någonting *inte* existerar i språket kan man aldrig få bekräftelse på, inte ens om man baserar sin undersökning på de mest omfångsrika och mångsidigt sammansatta korpusar och har de mest sofistikerade analysverktyg till sitt förfogande. Och det är fortfarande människor som skriver ordböcker, inte korpusarna. (Svensén 2004: 73; Svenséns fremhævelse)

Men der er ikke nogen tvivl om at det er fristende og nærliggende at anse seddelsamlinger for at være et ringe(re) kildegrundlag end tekstkorpora, bl.a. fordi korpora rummer mulighed for kvantitative undersøgelser og giver et bedre indblik i fx syntaktiske forhold og mere upåfaldende ord og betydninger der ikke altid er blevet indstreget når man har excerperet til en seddelsamling (jf. Svensén 2004: 57, Gudiksen & Hovmark 2008: 176). Jeg vil imidlertid gerne argumentere for det frugtbare i ikke at fokusere for snævert på datagrundlagets form (seddelsamling eller tekstkorpus). Det er i alle tilfælde vigtigt at have en kritisk tilgang til det specifikke datagrundlag man arbejder med, dets tilblivelse og sammensætning, og brugen af det i det daglige arbejde, set i forhold til den generelle problematik: data og repræsentativitet.

Man kan som et banalt eksempel tage ordet *kromand* der optræder i både ØMO og DDO (jf. Figur 1). ØMO's beskrivelse bygger på en seddelsamling, og DDO's på et tekstkorpus. Men der er i begge tilfælde tale om at produktet, dvs. ordbogsartiklerne, bygger på et udsnit af virkeligheden, ikke den fulde virkelighed. Ordet *kromand* findes i en mængde former og tilfælde i virkeligheden eller "verden", og må antages at følge bestemte, genkendelige mønstre. Det som ordbogsredaktørerne har bygget på, er imidlertid "kun" en delmængde af en vis beskaffenhed af denne virkelighed – en delmængde som jeg vil kalde data. Data kan være en seddelsamling eller et tekstkorpus – eller en enkelt ordbogsforfatters intuitive viden om og fornemmelse for sproget. Samlingen af data kan være mere eller mindre statisk eller dynamisk, lukket eller åben. I løbet af redigeringsprocessen bliver data omsat til et bestemt produkt i en karakteristisk, ofte

stærkt konventionaliseret form, nemlig en ordbogsartikel (fx *kromand*). I såvel udvælgelsen af data som i behandlingen af data i forbindelse med udarbejdelsen af ordbogsartiklen, skal der foretages vurderinger og træffes valg. Redaktøren må være opmærksom på 1) om data er repræsentative for den store virkelighed, og 2) opbygge nogle kritiske arbejdsrutiner i omgangen med data og omsætningen af det til en ordbogsartikel. Det vil aldrig være hele den sproglige verden man indfanger i et korpus, og der træffes nogle beslutninger, menneskelige beslutninger som Svensén siger (2004: 73; jf. citatet ovenfor), både i udarbejdelsen af korpus og undervejs fra korpus til ordbogsartikel. Det er disse to processer jeg vil se nærmere på i det følgende, primært med udgangspunkt i det helt konkrete arbejde med en seddelsamling, dels i forbindelse med indsamling af data (excerpering), dels i forbindelse med vurderingen af dataenes repræsentativitet.



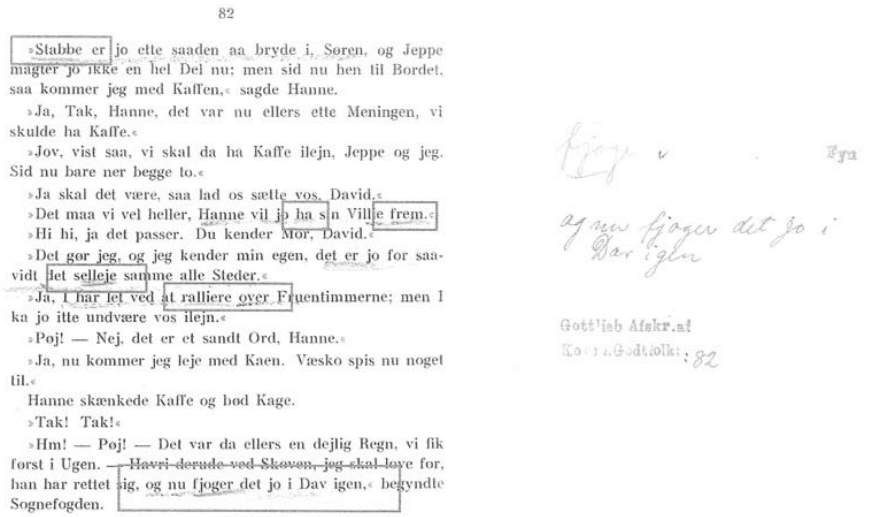
Figur 1: Ordet *kromand*, fra "verden" over data til ordbogsartikel (i ØMO og DDO).

2. Udvælgelse af data i en seddelsamling: excerpering

Hvad er det der står på sedlerne i en seddelsamling? Der står naturligvis bestemte uddrag af virkeligheden eller "verden" (jf. Figur 1), men hvilke? Og hvordan er de blevet indsamlet eller udvalgt? Den særlige udvælgelsesproces som ligger til grund for seddelsamlinger, kaldes traditionelt for excerpering. Ter-

merne excerpt og excerpering dækker over lidt forskellige ting afhængigt af hvilken historisk kontekst og tekst der er tale om, men i ordbogssammenhæng er følgende definition et godt udgangspunkt: "A citation is a short extract from a text which provides evidence for a word, phrase, usage, or meaning in authentic use" (Atkins & Rundell 2008: 48). Definitionen peger på at et excerpt til en ordbog skal give et autentisk belæg, ikke kun på et ord, men på en eller anden dimension af sproget. Definitionen åbner op for en lidt bredere opfattelse af hvad et excerpt rent praktisk er, som vil vise sig nyttig når vi dykker ned i ØMO's seddelkasser. Der står nemlig mange forskellige typer oplysninger på sedlerne i ØMO's seddelsamling (og tilsvarende samlinger).

Figur 2 giver et eksempel på et klassisk excerpt. Et antal ord eller ordforbindelser i en trykt tekst er indstreget, og der er efterfølgende udskrevet en seddel på hvert belæg. ØMO er en talesprogsordbog, så det er naturligt nok replikkerne fra den pågældende tekst (en roman af Morten Korch) der har haft redaktørens interesse.



82

»Stabbe er jo ette saaden aa bryde i, Søren, og Jeppe magter jo ikke en hel Del nu; men sid nu hen til Bordet, saa kommer jeg med Kaffen,« sagde Hanne.

»Ja, Tak, Hanne, det var nu ellers ette Meningen, vi skulde ha Kaffe.«

»Jov, vist saa, vi skal da ha Kaffe ilejn, Jeppe og jeg. Sid nu bare ner begge to.«

»Ja skal det være, saa lad os sætte vos, David.«

»Det maa vi vel heller, Hanne vil jo ha sin Ville frem.«

»Hi hi, ja det passer. Du kender Mor, David.«

»Det gør jeg, og jeg kender min egen, det er jo for saavidt let selleje samme alle Steder.«

»Ja, I har let ved at ralliere over Fruentimmerne; men I ka jo itte undvære vos ilejn.«

»Poj! — Nej, det er et sandt Ord, Hanne.«

»Ja, nu kommer jeg leje med Kaen. Væsko spis nu noget til.«

Hanne skænkede Kaffe og bød Kage.

»Tak! Tak!«

»Hm! — Poj! — Det var da ellers en dejlig Regn, vi fik først i Ugen. — Havri derude ved Skoven, jeg skal love for, han har rettet sig, og nu fjoger det jo i Dav igen,« begyndte Sognefogden.

fjoger

Dyn

og nu fjoger det jo i Dav igen

Gott'jeb Afskr.af

Kor[illegible]Godtjolk: 82

Figur 2: Excerpering af ordet *fjoge* 'småregne' fra en roman af Morten Korch.

Men eksemplerne fra ØMO viser også at sedlerne i denne samling ikke kun rummer ukommenterede sprogbrugseksempler (svarende til en slags konkordanslinjer ved søgninger i et tekstkorpus), men også kan rumme input af mere kommentaragtigt tilsnit til andre dele af en ordbogsartikel, fx definitionen eller de særlige encyklopædiske, kulturhistoriske afsnit som er et særkende for ØMO (jf. "evidence .. for use, or meaning", Atkins & Rundell 2008: 48). Det gælder fx et belæg på ordet *guldgalon*, udskrevet fra en bog fra 1924 om fester og højti-

der. Den omgivende tekst er også medtaget: "de ældres Huer var med Guldgalon og sorte Baand", men ikke som eksempel på en sproglig kontekst for ordet *guldgalon*. Den omgivende tekst er taget med fordi den bidrager til betydningsbeskrivelsen af ordet. På sedlen er der en henvisning til kilden, hvor redaktøren kan læse mere om især den kontekstuelle brug af guldgaloner.

Der kan altså optræde mange typer sproglige data på en ordbogsseddel, nøjagtig lige så mange som man har ønsket at beskrive i ordbogen, og som man derfor har haft brug for at nedfælde oplysninger om på sedlerne. I talesprogsordbogen ØMO's samlinger indgår der således mange sedler hvor det alene er udtalen af et ord der anføres. Én enkelt seddel kan naturligvis også rumme oplysninger om flere sproglige elementer på samme tid (fx udtale, syntaks og betydning). Det gælder fx følgende seddelbelæg på ordet *fjoge* (Figur 3), hvor ordet er lydskrevet i citaterne, og hvor citaterne desuden er medtaget i den redigerede artikel.

fjoge III v – [*fja·w·*, sjæld *fjaw·*] øF(alm, jf FØF:-273), sF(VSk), m acc 1 T(Bjb), [*fjö·w·*] vF(Br, Søs, Ejlby), [*fjä·w·*] vF(Lu). Bøjn: ede-bøjn. – småregne, støvregne, *fjog(e)regne* (jf *fnuge* 2) øF(alm), vF(Lu, Br), sF, T(alm), Æ(Bregn): *det begyndte* [*a fjaw·*] *lige ret før* øF(Kert), [*de fja·wɪ ilid*] øF(Revn), [*de· æ någ bljw räj·nvæ:ɪ / ja· de fjä·wɪ ən bidə*] vF(Lu); jf *fjog* II ‖ hertil fjog(e)-regn, -regne, -vejr.

Figur 3: Talesprogseksempler på ordet *fjoge* 'småregne', medtaget som citater i ØMO.

Undertiden kan de sprogprøver man møder på sedlerne have et kunstigt præg eller være resultatet af en forholdsvis stram redaktionel styring (jf. Hovmark 2004), men begge citater er i dette tilfælde ganske spontane og autentiske – glimt af ordet *fjoge* i typisk løbende tale.

Når man har et tekstkorpus, har man hentet større, sammenhængende stykker af den sproglige virkelighed ind i ordbogsværkstedet, hvor alle deres dele konstant kan hentes frem og studeres i deres kontekst. Når man har en seddelsamling, har man ikke på samme måde de sammenhængende stykker tekst og

konteksten til stede – man har udvalgt det man har fundet interessant og relevant. Dette forhold bliver særligt påtrængende når der er tale om en talesprogsordbog hvor data i høj grad er indsamlet før båndoptagerens tid, for her står den samlede tekst ikke længere til rådighed når først den er blevet ytret. Det betyder til gengæld at man er nødt til at gå direkte til primærkilden, "verden", hvis der er noget man vil slå op. Man har ikke "fastfrosset" større dele af konteksten, men man har mulighed for at gå tilbage til de informanter man har kontakt til, og som på en måde udgør ordbogens levende korpus. Det er netop hvad man ofte gjorde på ØMO (og stadig gør i begrænset omfang): Som arbejdet med ordbogen og samlingerne skred frem, stødte man uundgåeligt på huller i materialet, ubelyste spørgsmål osv., hvorefter man henvendte sig til relevante informanter for at komplettere og supplere datagrundlaget for ordbogen (jf. Bévort 1965, Gudiksen & Hovmark 2009). De to seddelbelæg på ordet *fjoge* ovenfor (Figur 3) stammer netop fra sådanne kompletteringsrunder (jf. markeringerne "KOMPL. KLADDE 4686" og "KOMPL. LISTE D 385"), udført af redaktørerne "K.M." (Kristen Møller) i 1966 og "PA" (Poul Andersen, den mangeårige leder af ØMO) i 1937. I ØMO's tilfælde har redaktørerne været vant til at skulle bevæge sig ud over det forhåndenværende seddelkorpus når man ønskede at undersøge noget, og man er som redaktør på ordbogen stadig nødt til at være opmærksom på sammenhængen mellem "verden" og det aktuelle datagrundlag.

3. Arbejdet med data i en seddelsamling

Når man udarbejder artikler på basis af en seddelsamling, er der ofte kun ganske få belæg, undertiden endda kun et enkelt, til rådighed for redaktøren. Det er dog ikke pr. definition ensbetydende med at der altid er flere belæg på et ord i et tekstkorpus, det kommer helt naturligt an på samlingen og korpusset. Fx findes der 39 sedler på *kromand* i ØMO's samlinger, men faktisk kun 24 belæg på samme ord i Korpus90 der rummer størsteparten af det oprindelige korpus til DDO (jf. ordnet.dk/korpusdk/). Man må også være opmærksom på at flere belæg kan stamme fra samme kilde – det gælder både seddelsamling og tekstkorpus (der er fx kun 13 kilder til ordet *kromand* i Korpus90). Men det er rigtigt at datagrundlaget for artikler og betydninger i ØMO og tilsvarende ordbøger kan forekomme temmelig spinkelt. Men netop derfor kan seddelsamlinger og arbejdet med dem være et udmærket udgangspunkt for en diskussion af data og repræsentativitet, herunder de konkrete kildekritiske overvejelser det kan være relevant at gøre sig som ordbogsredaktør når man arbejder med sine data.

Lad mig som eksempel tage skældsordet *krapillik*. Dette belæg er blevet underkastet den samme kildekritiske vurdering som alle andre belæg i ØMO.

Ender vurderingen med et ja, vil belægget komme med i ordbogen – det gælder også hvis sedlen står alene og altså er det eneste belæg på et ord. Det er med andre ord ikke antallet af belæg, men kvaliteten af et belæg der er afgørende. Og kvaliteten vurderes på basis af velkendte kildekritiske faktorer: Hvem er meddeleren? Hvordan har indsamlingssituationen været? Og hvem er optegneren?

En vurdering af meddeleren stod i centrum helt fra projektets start, dvs. under indsamlingsarbejdet. Knappe resurser gjorde det nødvendigt at finde de bedste meddelere, og man førte kartotek over meddelerne og deres kvaliteter – eller mangel på samme (jf. Gudiksen & Hovmark 2009, Køster 2009). Meddelerne blev vurderet på to skalaer: om de var gode (1. eller 2. klasses) sprogligt henholdsvis sagligt (dvs. hvad angår den kulturhistoriske viden om den ældre bondekultur). En vurdering af meddeleren er også i dag en naturlig del af arbejdet. Et belæg fra en dårlig meddeler kommer ikke med. Hvad angår meddeleren til ordet *krapillik*, Lars Hansen, beskrives han som en "meget interesseret Meddeler" og "I-Klasses" hvad angår det saglige. Der nævnes ikke noget om det sproglige, men der står at meddeleren har været bofast i det pågældende sogn hele livet bortset fra de senere år, hvilket er et plus i karakterbogen i ØMO-sammenhæng, for det kan tyde på et stabilt og forholdsvis upåvirket dialektalt sprog. Karakteristikken er ikke overvældende, men der er heller ikke noget der taler imod at acceptere belægget på *krapillik*. Derudover tilhører *krapillik* en bestemt orddannelsestype, dannet med suffikset *-ik(e)* der bruges i afledninger med emotiv betydning, i dette tilfælde en række af synonymer brugt som nedsættende betegnelse for en gammel udslidt hest. Ordet *krapillik* står derfor ikke helt alene alligevel.

Ud over meddeleren vurderes også indsamlingssituationen og optegneren. Indsamlingssituationen kan have præget de sproglige data i mere eller mindre autentisk retning (jf. Atkins & Rundell 2008: 48), dvs. gjort data, fx sprogeksempler, mindre autentiske eller "naturlige" (jf. det såkaldte "observer's paradox" inden for fx talesprogsforskningen og etnografiske videnskaber, dvs. at forskerens tilstedeværelse påvirker informanten og gør dennes optræden og sprogbrug mindre spontan og naturlig; jf. fx Hovmark 2004). Og optegnervaner og -holdninger kan også spille en rolle, fx i forbindelse med en udtaleangivelse (udtaleangivelserne i ØMO er normaliserede efter strukturalistiske principper, og materialet kan tolkes lidt forskelligt; jf. fx Andersen 1958: 13-16, Gudiksen 2001: 43f.).

Den kildekritiske vurdering henholder sig til en række objektive faktorer, men i sidste instans kommer man ikke uden om et vist skøn: Skal man tro på hvad der står på en seddel, eller skal man ikke? Svaret vil til en vis grad afhænge af den enkelte redaktør og dennes erfaringer og vurderinger. Det kan fx være en vurdering af om et ord er centralt eller perifert i forhold til begreber som folke-

ligt sprog vs. rigsmålspræget sprog. Under alle omstændigheder kræves der et indgående og detaljeret kendskab til stoffet og kilderne når man arbejder på ØMO, dvs. ikke kun sproglige forhold, men også mere interne forhold omkring indsamlingen, meddelerne og optegnerne. Og det er vigtigt at have løbende, fælles diskussioner af typiske problemstillinger på redaktionen for at etablere faste rammer for omfanget af skøn.

Spørgsmålet om data og repræsentativitet trænger sig også på i forbindelse med angivelsen af et ords eller en betydnings udbredelse. ØMO dækker et stort geografisk område med mange forskellige dialekter. I alt er der afgrænset 14 hovedområder (fx Sjælland, Falster og Fyn) og 15 underområder (fx Sydvestsjælland, Nordfalster og Østfyn) (jf. ØMO Tillægsbind: 37). Det er en obligatorisk del af arbejdet at stedfæste alle de data der præsenteres i ordbogen. De fleste optegnelser er blevet stedfæstet til et bestemt sogn, og ofte vil man opremse de pågældende sogne i ordbogsartiklen. Men hvis man vurderer at materialet rækker til det, vil man foretage en generalisering af udbredelsen til et eller flere hovedområder – eller evt. til hele ømålsområdet. Det sidste er fx tilfældet ved ordet *kromand*. Her er der anført: "Ø(alm)". Det betyder at ordet vurderes til at være almindeligt i hele ømålsområdet. Mere forsigtige markører er "vist alm", der står for 'vist almindeligt', og "spor opt", der ofte bruges og betyder 'sporadisk optegnet og muligvis almindeligt'. Endelig kan man bruge markøren "spor" 'sporadisk og ikke almindeligt', men det kræver at redaktøren har belæg for at ordet er mindre almindeligt end et andet (eller flere andre) ord, der da efter reglerne skal nævnes eksplicit i artiklen (jf. ØMO Tillægsbind: 22).

Hvad skal der fx til for at et ord som *kromand* kan opnå stemplet Ø(alm)? Det er klart at antallet af belæg, dvs. et kvantitativt element, spiller en vis rolle her. Ved *kromand* var der 39 belæg, et ganske hæderligt antal i ØMO-sammenhæng. Ganske vist er der kun anført udtaleoplysninger på 22 af de 39 sedler, men da der kun er én betydning, er dette ikke noget problem. Men naturligvis spiller fordelingen på de forskellige områder også en rolle. Også her består *kromand* testen: materialet er pænt fordelt på hovedområderne: 14 belæg fra Sjælland m.v., 5 belæg fra Møn og Lolland-Falster, og 20 belæg fra Fyn, Tåsinge, Langeland og Ærø m.v. (man forsøgte hurtigt at tilrettelægge indsamlingsarbejdet så der blev indsamlet data fra alle hovedområder, jf. fx Gudiksen & Hovmark 2009). Endelig spiller meddelerne også en rolle, men her er der flere kendte, solide meddelere repræsenteret. I vurderingen indgår imidlertid også et lidt større skøn. Ordet *kromand* er almindeligt i rigsmålet, og der foreligger ikke andre væsentlige alternativer til ordet, bortset fra *krofatter* der bruges uformelt på Fyn og *krovært* der er langt svagere belagt og kun opnår betegnelsen "spor opt". At drage den konklusion at *kromand* er almindeligt udbredt i alle ømål er derfor ikke voldsomt dristigt, heller ikke selvom der fx mangler belæg fra et par

af de mindre hovedområder. Ud fra kendskabet til seddelsamlingens tilblivelse og karakteristika er der ikke nogen tvivl om at data er tilstrækkelig repræsentativt, og at *kromand* er almindeligt i ømålene.

Som nævnt udelades belæg fra dårlige meddelere, og der bliver ikke oprettet en artikel hvis et enkeltstående belæg fra en sådan meddeler står alene. Et eksempel på en dårlig meddeler er den såkaldte Valentiner, der på kartotekskortet kort og kontant karakteriseres som "uden synderlig interesse". Men hvorfra ved vi at Valentiner er dårlig? Det slutter vi os bl.a. til ud fra et basalt kendskab til hans baggrund: han var født på egnen syd for København, mellem Køge og Roskilde, men han havde i sit voksne liv sejlet som kaptajn og noteres i kartoteket som bosiddende i det indre København. Men vi ved det også ud fra erfaringer med de oplysninger han har bidraget med. At en person flytter til et andet sted, er ikke pr. definition ensbetydende med at den pågældende mister følingen med barndommens sprog. Men de oplysninger som Valentiner kom med i 1940'erne forekom allerede dengang at være spredte og tilfældige, og der blev sået tvivl om deres kvalitet. Og når man konfererede med andre meddelere fra det levende korpus, viste tvivlen sig ofte at være velbegrundet. Valentiner havde fx leveret et enkeltstående belæg på ordet *langlars* 'doven person'. Tre andre belæg fra samme område er imidlertid såkaldte "minus-sedler". På spørgsmålet om hvorvidt en doven person kan "kaldes for en Langlars?", svares der henholdsvis: "ukendt", "Det har jeg aldrig hørt" og "en doven Person kaldes for en Dovenlars". Resultatet er at vi i dag normalt ser bort fra belæg fra Valentiners hånd, med mindre de er underbygget af andre kilder.

4. Afslutning og perspektivering

Jeg har i denne artikel diskuteret spørgsmålet om data og repræsentativitet på basis af eksempler fra ØMO's seddelsamling. Seddelsamlinger er forældede som datatype, men det er problematikken omkring data og repræsentativitet ikke: Uanset om data foreligger i form af en seddelsamling, et tekstkorpus, en kombination eller noget helt fjerde, vil der være tale om et udsnit af virkeligheden som skal vurderes kritisk som kildegrundlag. I 1987 nævner Svensén at den nye it-teknologi kan give et frisk pust til vante forestillinger og traditionelle arbejdsrutiner inden for ordbogsarbejdet (jf. afsnit 1 ovenfor, Svensén 1987: 243). Eksemplerne fra ØMO's seddelsamling kan måske bruges som et muligt afsæt for generelle overvejelser omkring data og repræsentativitet, også i den it-styrede ordbogsvirkelighed baseret på tekstkorpora.

Eksemplerne fra ØMO har fx peget på at en kritisk omgang med data omfatter to relationer eller overgange: relationen mellem virkeligheden og data,

og mellem data og ordbogsprodukt (artikel). Netop fordi seddelsamlinger rummer relativt små og selektivt udvalgte udsnit af virkeligheden, vil man ofte efterspørge en større kontekst. I ØMO's tilfælde er dette særlig påtrængende fordi data her sjældent foreligger i trykt form, men eksisterer som en flygtig virkelighed der kun kan gribes i farten – og i de første mange år indfangede man ikke engang denne virkelighed med båndoptagere, men kun med pen og papir. At foretage en kildekritisk vurdering af de glimt af virkeligheden som har været til rådighed i seddelsamlingen har derfor været og er stadig en indarbejdet rutine, ligesom det har været en nødvendig del af indsamlingsarbejdet og arbejdet med seddelsamlingen jævnligt at gå tilbage til virkeligheden og spørge den til råds. Denne opmærksomhed kan utvivlsomt stadig være nyttig i vore dage hvor man med de store tekstkorpora i højere grad kan have oplevelsen af at have om ikke den fulde, så dog den tilstrækkelige og nødvendige virkelighed liggende på computeren (jf. Svensén 2004: 72 med henvisninger). I hvor høj grad er tekstkorpus repræsentativt? Hvilke sprogtyper og genrer er repræsenteret, og i hvilket forhold?

Overvejelser af denne art er naturligvis ikke ukendte når man arbejder med et tekstkorpus (jf. fx Asmussen (2006) om KorpusDK). Et andet forhold der kommer til syne når man ser på arbejdet med seddelsamlinger er måske mere interessant at overveje. Det gælder det meget indgående arbejde med den enkelte kilde, først og fremmest dens ophav, men også forholdene omkring indsamlingen og indsamleren. Denne fornemmelse for den enkelte kilde opnås ikke på samme måde ved arbejdet med tekstkorpora. Den store mængde data med mulighed for undersøgelse af kontekstuelle forhold, frekvens osv., ofte på basis af statistiske værktøjer og komplekse søgninger (jf. Atkins & Rundell 2008: 103ff., DDO, bind 1: 54-59), er den store og revolutionerende styrke ved tekstkorpora – men kan det undertiden også være en svaghed? Det resultat man præsenteres for er anonymt, og man har ikke altid behov for at komme tættere på den enkelte kilde. Det er i visse sammenhænge en fordel fordi det sætter fokus på generelle og typiske træk i sproget, og mindsker risikoen for at særtræk og måske ligefrem idiosynkratisk sprogbrug kommer til at fylde for meget i ordbogen. Men det kan også skubbe den enkelte kilde og fornemmelsen for den i baggrunden i den daglige arbejdsrutine.

En af de store udfordringer i arbejdet med ØMO's seddelsamling er vurderingen af de meget kvalitativt udvalgte datas repræsentativitet i redigeringen af materialet til en artikel, eksemplificeret ved muligheden for at generalisere udbredelsesangivelser. Her har man en stor fordel når man har et tekstkorpus og statistiske værktøjer til rådighed der kan give mere velfunderede svar på den slags spørgsmål. Når blot man husker at data i en seddelsamling ikke er fuldstændig tilfældige – i ØMO's tilfælde blev data fx indsamlet systematisk ud fra

kriterier om repræsentativitet (geografisk fordeling, meddelerkarakteristik), hvilket giver mulighed for at nå frem til et svar der også er kvalificeret. Og når man husker at de resultater som maskinen spytter ud ikke uden videre kan eller bør tages som udtryk for en endegyldig eller eneste sandhed (man kan jo fx have formuleret sin forespørgsel på en uhensigtsmæssig måde, ligesom de forskellige korpusværktøjer kan anvende forskellige statistiske metoder; jf. Hovmark 2009). De er yderligere input til at give en så fyldestgørende og præcis fremstilling af den sproglige virkelighed som muligt, og de kan med fordel gøres til genstand for kildekritisk opmærksomhed – i lighed med oplysningerne i en seddelsamling.

Kilder og litteratur

Andersen, Poul, 1958: Fonemsystemet i Østfynsk. København.

Atkins, B.T. Sue & Michael Rundell, 2008. The Oxford Guide to Practical Lexicography. Oxford.

Bévort, Inger, 1965: Indsamlingen af materiale til Ømålsordbogen. I: Danske Folkemål 19. S. 239-250.

DDO = Den Danske Ordbog 1-6, 2003-2005. København.

Gudiksen, Asgerd, 2001: -vorn's lydlige form. I: Danske Talesprog 2. S. 27-76.

Gudiksen, Asgerd & Henrik Hovmark, 2008: Båndoptagelser som kilde til Ømålsordbogen. I: Svavarsdóttir, Ásta m.fl. (red.): Nordiske Studier i Leksikografi 9 (Skrifter udgivet af Nordisk Forening for Leksikografi 10). S. 171-180.

Gudiksen, Asgerd & Henrik Hovmark, 2009: Måske husker De noget alle andre har glemt. I: Gudiksen, Asgerd m.fl. (red.): Dialektforskning i 100 år. København. S. 13-64.

Hovmark, Henrik, 2004: Jagten på det autentiske citat. I: Danske Talesprog 5. S. 3-21.

Hovmark, Henrik, 2009: Hund og menneske imellem. I: Farø, Ken (red.): Sprogvidenskab i glimt. Odense. S. 397-402.

Køster, Finn, 2009: Fra fortid og nutid. I: Hovmark, Henrik m.fl. (red.): I mund og bog. København. S. 193-204.

Svensén, Bo, 1987: Handbok i lexikografi. Principer och metoder i ordboksarbetet. Stockholm.

Svensén, Bo, 2004: Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik. Stockholm.

Winchester, Simon, 2004: The Meaning of Everything. The Story of the Oxford English Dictionary. Oxford.

ØMO = Ømålsordbogen. En sproglig-saglig ordbog over dialekterne på Sjælland, Lolland-Falster, Fyn og omliggende øer, 1-, 1992 ff. København.

ØMO, Tillægsbind, 1992. Indledning, nøgler, forkortelseslister, kort. København.

Henrik Hovmark
Ømålsordbogen, Nordisk Forskningsinstitut, Afdeling for Dialektforskning, Københavns Universitet
hovmark@hum.ku.dk