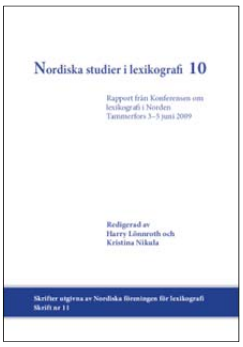
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Internettpublikasjonar som kjelde i dokumentasjonsordbøker - Status og bruk

Forfatter: Åse Wetås & Knut E. Karlsen

Kilde: Nordiska Studier i Lexikografi 10, 2010, s. 522-529
Rapport från Konferens om lexikografi i Norden, Tammerfors 3.-5. juni 2009

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.

Åse Wetås & Knut E. Karlsen

Internettpublikasjonar som kjelde i dokumentasjonsordbøker
Status og bruk

The purpose of this article is to discuss the use of Internet texts as a source in documentary lexicography. Texts published in print are stable and unchangeable in form, and they are mostly subject to long term storage in public national collections or archives. On the other hand electronic texts published on the Internet can be consecutively changed or even removed from one day to another. This is an obvious problem for the documentary lexicographer. To meet the demand for scientifically verifiable data, the lexicographer himself therefore needs to ensure that the text is properly stored for posterity. In *Norsk Ordbok* this is done by saving the web-page from which a specific word or collocation is harvested, as a PDF-file. The lexicographer should thoroughly consider the *type* of text he is referring to, not only the publishing format (printed vs. virtual). A lot of Internet texts have never been subject to quality controls such as proofreading. Texts from established publishing houses tend to be better than newspaper articles in terms of both text quality and stability of storage, while no such control is applied to personal web sites.

Nøkkelord: Internett-kjelde, dokumentasjonsleksikografi, langtidslagring, innhausting

Innleiing

Denne artikkelen tek opp Internett-kjelder til bruk i dokumentasjonsordbøker. Innleiingsvis vil vi sjå på tilhøvet mellom publisering av tekst på papir versus virtuell publisering på Internettet. Deretter vil vi peika på to problemfelt som må takast stilling til for leksikografen som bruker nett-tekst. Det første gjeld Internettet som lagringsmedium for tekst, og det andre gjeld spørsmål om sjangerdiversitet og kvalitetskontroll. Så skal vi sjå på om det finst nokon praksis for korleis leksikografen møter desse kjeldene, før vi til slutt rundar av med å visa kva *Norsk Ordbok 2014* har gjort for å møta dei problemstillingane vi reiser her. Artikkelen er meint å ha ei praksisnær tilnærming til ei relevant og samstundes relativt ny leksikografisk problemstilling.

Utgangspunktet for denne vesle studien er at vitskaplege dokumentasjonsordbøker som dekkjer skriftspråkleg tilfang, i stadig sterkare grad møter utfordringar med bruk av kjelder frå Internettet. Utviklinga har dei seinare åra gått i retning av at stadig fleire tekstar som tidlegare blei formidla med papir som

medium, no *både* kjem i ein nettversjon og som prenta tekst, eller dei blir rett og slett *utelukkande* publiserte på nettet. I tillegg er sjangervariasjonen stor i nettuniverset, og dette gjev ekstra utfordringar for leksikografen.

Når ein skal vurdera nettkjelder til bruk i dokumentasjonsleksikografien, er det to ulike problemstillingar som melder seg. Den eine av desse gjeld dokumentasjonen og lagringa av dei elektroniske tekstane for å møta vitskaplege krav til etterprøvbarheit. Den andre gjeld sjangervurderingar og utøving av kjeldekritikk. Vi vil i det vidare først drøfta problem knytt til lagring og etterprøvbarheit, før vi kjem nærare attende til det som gjeld sjanger, kvalitetskriterium og grunnleggjande kjeldekritikk.

Dokumentasjon, lagring og etterprøvbarheit

Tekst som er prenta og utgitt på papir, er stabil og uforanderleg i form, og i tillegg blir han (i det minste i prinsippet) langtidslagra gjennom innhaustingsordningar til etablerte institusjonar. I Noreg er det Nasjonalbiblioteket som står for denne langtidslagringa. I og med at teksten eksisterer i eit fysisk format, vil han vera tilgjengeleg over tid på ein måte som gjer at bruken av han i forskingssamanheng ikkje utfordrar vitskaplege grunnkrav til etterprøvbarheit. Det same er ikkje tilfellet med tekst som berre blir publisert i eit virtuelt rom som det Internettet utgjer. Tekst formidla med Internettet som medium kan stendig endrast. Desse endringane kan skje utan at url-lenkja til den aktuelle nettsida blir endra. I praksis betyr dette at ein nett-tekst som dokumenterer bruken av ei interessant ordform eller ein kollokasjon i går, ikkje nødvendigvis inneheld den gitte ordforma eller den aktuelle kollokasjonen i dag. Dette utgjer eit openbert problem for dokumentasjonsleksikografien. Om det ikkje er lagra nokon kopi av den virtuelle teksten på det tidspunktet leksikografen hausta inn orddokumentasjonen, er det altså ikkje mogleg å dokumentera og etterprøva informasjonen leksikografen har henta ut ved seinare høve.

Nyhendesaker i nettaviser er typiske døme på ein sjanger der innhaldet i tekstane kan endra seg raskt. Eitt av hovudfortrinna ved nettpublisering av nyhendesaker, sett frå nyhenderedaksjonane si side, er at sakene kan oppdaterast og utviklast heile tida, og etter kvart som hendingane skrid fram eller etter kvart som journalisten får nye saksopplysningar eller nye kjelder til rådvelde. Då blir teksten endra, og det "gamle" tekststykket forsvinn utan at det er blitt lagra for ettertida. Nyhenderedaksjonane har ikkje eigne system for lagring av alle tekstversjonar dei lagar under ei url-lenkje. Dette inneber at eldre versjonar av slike tekstar som regel er tapte.

Nett-tekst frå etablerte tidsskriftredaksjonar viser seg oftast å vera identiske med ein prenta versjon av same teksten. Med unntak av dei reine nyhendesakene, ser vi at mange mediehus driv samdrift mellom nettutgåve og prenta utgåve av same publikasjonen. Då er tekstane sikra lagring fordi den prenta utgåva blir teken vare på. Men i tillegg er det eit vell av andre tekstprodusentar og tekstprodukt på Internettet. Tekstproduksjonen er i stadig større grad brukargenerert, og store nettaktørar som til dømes Wikipedia, baserer seg på publikumsbidrag til artiklane sine. Også mange nettaviser inneheld brukargenerert stoff, og i tillegg omfattar nettet store mengder bloggar og samtalesider som er i stendig endring og som har mange tekstprodusentar bak seg.

Ei dokumentasjonsordbok som tek mål av seg til å vera vitskapleg, er avhengig av at påstandane som blir formulerte om eit ords tyding, bruk og utbreiing, i prinsippet er fullstendig etterprøvbare for ettertida. Når dokumentasjonsleksikografen skal bruka tekstar frå Internettet som kjelde, kan ein tenkja seg ulike tilnærmingsmåtar. Vi har nedanfor peikt på tre aktuelle tilnærmingar:

Konservativ tilnærming

Den mest konservative tilnærminga er heilt og fullt å avvisa bruk av nettkjelder til dokumentasjon på ordbruk, med utgangspunkt i at problema knytte til langtidslagring er så store. Denne tilnærminga vil for ei moderne dokumentasjonsordbok innebera at tilfanget av tilgjengelege kjelder krympar frå år til år, rett og slett fordi fleire og fleire tekstar etter kvart berre blir publiserte elektronisk. Frå ein lingvistisk og leksikografisk synsvinkel kan det då diskuterast om leksikografen verkeleg har tilgang til kjelder som gjev ein så mangefasettert dokumentasjon på faktisk ordbruk som han eller ho bør ha i dokumentasjonssamanheng.

Radikal tilnærming

Den mest radikale tilnærminga er å ikkje ta omsyn til at nett-tekstane er mindre stabile enn prenta tekst, og berre referera til url-lenkje i ordboka eller underlagsmaterialet som ligg til grunn for ordboksteksten. Då vil ein risikera at kjeldene i ein god del tilfelle aldri vil kunna rekonstruerast, og leksikografen har dermed formulert ein påstand om språkbruk som ikkje er etterprøvbar.

Ein farbar middelveg?

Ei fornuftig løysing er etter vårt syn å leggja til grunn at Internettet er eit flyktigare medium enn bøker og andre publikasjonar prenta på papir. For å garantera at både ordform og kontekst er mogleg å konservera, må leksikografen sikra seg at den aktuelle tekstsekvensen blir lagra på forsvarleg vis. I Noreg haustar Nasjonalbiblioteket regelmessig inn alle nettstader med eit domenenamn som endar på .no for langtidslagring, men sjølve innhaustinga skjer kvartalsvis. Det inneber at svært mykje nettpublisert tekst aldri blir langtidslagra, men forsvinn med sider som blir lagde ned, eller som blir endra ein eller fleire gonger mellom desse kvartalsvise innhaustingane. For at dokumentasjonsleksikografen skal vera sikker på at dei analysane han eller ho gjer av internettpublisert materiale skal vera fullstendig tilgjengelege for etterprøving, må leksikografen sjølv syta for at dokumentasjonen blir henta inn og forsvarleg lagra for ettertida.

Sjangervurdering og kjeldekritikk

I seg sjølv er Internettet berre ein publiseringskanal, og sjangermangfaldet er ikkje mindre enn det er i meir tradisjonelle kanalar for publisering av tekst. I praksis er det nok snarast slik at sjangermangfaldet på nettet er *vesentleg større* enn det er for prenta tekst, og nettet formidlar også det svært talemålsnære og spontane uttrykket som sjeldan blir sett på papiret og publisert. Det er dessutan ei vanleg oppfatning at tekstane på Internettet i mindre grad enn prenta tekst har gått gjennom kvalitetssikring i form av korrekturlesing og redaksjonell handsaming.

Nokre typar nett-tekst står likevel fram som intuitivt meir akseptable til bruk som ordbokskjelder enn andre. Dette gjeld særleg tekst som både er publisert på nettet og i prenta form. Slike tekstar er oftast tekstar frå ulike periodiske publikasjonar eller dei kan vera fagartiklar (t.d. elektroniske konferanserapportar og artiklar frå fagtidsskrift). Denne typen tekstar har eit statisk uttrykk på nettet, og dei blir samla inn og lagra både i fysisk og elektronisk form. Dermed har ein stetta kravet om langsiktig dokumentasjon.

Andre typar tekst publisert på Internettet vil vera meir ustabil over tid, fordi han i større grad baserer seg på dei moglegheitene nettet faktisk gjev til rask endring og omredigering. Dette gjeld mellom anna tekst som kan påverkast av brukarane, slik vi til dømes ser med leksikonartiklar i Wikipedia. Mange av desse tekstane er nok relativt stabile over tid, men dei er i prinsippet opne for endring når som helst. Som eit ytterpunkt i tekstmangfaldet står tekst som til

dømes jamleg oppdaterte nyhendesaker i aviser. Desse tekstane kan bli totalt omredigerte eller nyskrivne på kort tid og er såleis svært ustabile.

Ein god del nett-tekst har aldri vore gjennom korrekturlesing og språkvask, slik trykte bøker, artiklar og avistekstar oftast har. Dette pregar sjølvsagt tekstmaterialet. Den korte tidsavstanden mellom innskriving og publisering av tekst er ein del av eigenarten ved Internettet og eit stort pre for alle som vil formidla nyhende raskt. Den korte tidsavstanden fører samstundes til at tekstkvaliteten blir meir varierande. Vi vil hevda at leksikografen i møtet med mangfaldet av tekst som ligg tilgjengeleg på nettet, bør gjera ei grundig vurdering av *teksttype*, snarare enn berre å ta stilling til publiseringsformatet. Med ein tidsskriftartikkel frå ein etablert tidsskriftsredaksjon kan ein vera rimeleg trygg på kvaliteten og lagringsstabiliteten. Avisartiklar kan vera meir ustabile, medan det med tilfeldige, private heimesider ikkje er kontroll med langtidslagring av eit gitt tekstutrykk i det heile. Ein kan heller ikkje rekna med at denne typen nettstader har noka språkleg kvalitetssikring av tekstinnhaldet.

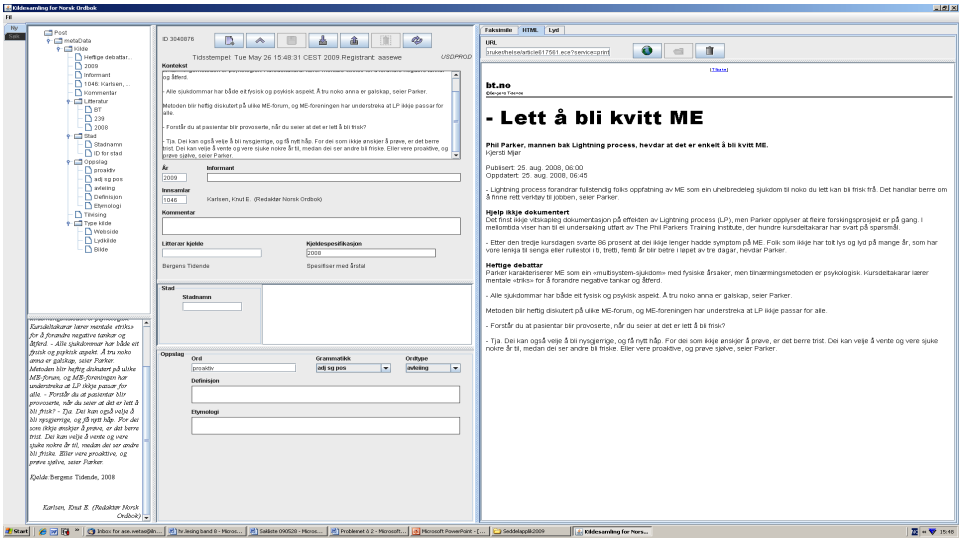
Praksis i Norsk Ordbok og i samanliknbare ordbøker

Inntil no har praksisen i *Norsk Ordbok* når det gjeld bruk av nettkjelder, vore å bruka desse kjeldene som supplement til dei andre materialkjeldene. Men strenge krav til etterprøvbarheit avgrensar bruken av nettkjelder. Nettpublikasjonar representerer sjølvsagt ei viktig kjelde til faktisk språkbruk som vi ønskjer å utnytta, men dokumentasjonskravet har – iallfall til no – gjort at vi oftast har nytta nettkjelder som òg har ei (parallell)utgåve i papirversjon. Dei redaksjonelle reglane og praksisen vår gjer at det sjølvsagt òg er høve til å bruka nettet for å finna bruksdøme som er så allment kjende at dei kan stå som redaksjonelle døme. Når ein bruker materiale frå Internettet, er det eit krav til redaktørane at dei skal oppretta ein elektronisk ordbokssetel som lagrast i ein database. "Setelen" skal innehalda så nøgne opplysningar om teksten som råd, det vil si fullstendig url-lenkje og publiseringsdato. Ordbelegget blir sett inn med ein kontekst med høve til å lagra inntil 4 000 teikn. Dersom leksikografen hentar ei opplysning eller eit sitat frå ei nettkjelde som krev sidetal (alle kjelder utanom aviser), må ein føra opplysningar om pagineringa i den prenta versjonen. Elles kan kjelda ikkje nyttast.

Ein tilsvarande praksis i møtet med nettkjeldene som den vi har skildra her, har òg *Den Danske Ordbog* (DDO). Til skilnad frå *Norsk Ordbok* (NO) er DDO ikkje noka dokumentasjonsordbok. Dette gjer kjelde- og lagringsproblema som er skildra ovanfor mindre akutte. På førespurnad svarer DDO-redaksjonen at

dei vurderer nettkjeldene som eit viktig supplement til kjeldematerialet dei elles rår over (inkludert *KorpusDK*), og dermed er innstillinga mykje den same som i NO.[1]

NO implementerte sommaren 2009 ein ny setelapplikasjon som ein del av det redaksjonelle redigeringsverktyet. Denne applikasjonen legg til rette for at redaktørar som hentar belegg frå nettkjelder, kan lagra ei PDF-fil av den aktuelle nettsida (jf. figur 1). Når denne nye setelapplikasjonen er ferdig utvikla, vil PDF-lagringa gjera det enklare å bruka nettkjelder. Samstundes vil dette tiltaket stetta krava til kjeldestabilitet slik at teksten blir tilgjengeleg for etterprøving. Det vil medføra at vi kan nytta Internettet som kjelde i større grad enn tidlegare. Ein slik måte å hausta inn kjeldedokumentasjonen på vil gjera nettet til meir enn eit supplement for leksikografen i det daglege arbeidet. Tiltaket vil truleg vera naudsynt i møte med ei utvikling der ein stadig større del av tekstproduksjonen er tilgjengeleg utelukkande i elektronisk versjon.



Figur 1. Den nye setelapplikasjonen til Norsk Ordbok 2014. Vindauget lengst til høgre viser korleis det er mogleg å lagra ei PDF-visning av den aktuelle nettsida informasjonen er henta frå.

1 Redaksjonen i DDO er merksame på problema som knyter seg til langtidslagring av nettkjeldene dei nyttar. I Danmark er det Det kongelige Bibliotek som har ansvaret for innhausting av nettekst, og praksisen deira svarer til den som Nasjonalbiblioteket i Noreg har.

Kjelder som dokumenterer skrift versus talemålskjelder

Som undertittelen "Ordbok over det norske folkemålet og det nynorske skriftspråket" viser, skal *Norsk Ordbok* også beskriva det norske folkemålet – altså dei norske dialektane. Vi har til no primært diskutert nettet som kjelde til dokumentasjon av skriftspråkbruk, men nettet kan også vera kjelde til talemålsopplysningar. Det finst store mengder virtuelt publiserte dialektordlister og lokale ordsamlingar som er tilgjengelege med få tastetrykk. Krava vi stiller til kvaliteten på kjelder til dokumentasjon av talemål, skal sjølvsagt vera like store som dei krava vi stiller til kjeldene som dokumenterer skriftspråkleg bruk av eit ord eller ein kollokasjon. Men korleis kan kjelder til talemålsopplysningar på nettet kvalitetssikrast? Kva krav kan og skal stillast til bruken av desse i ein dokumentasjonsleksikografisk samanheng?

Brukarar av *Norsk Ordbok* skal kunna sjå ordboksredaksjonen i korta og dermed kunna vurdera det totale materialet som ligg bak ordboksartiklane våre. Dette kravet er like sjølvsagt for talemålskjelder som for skriftspråklege kjelder. I motsetning til den skriftspråklege dokumentasjonen, som må ha ei viss kvantitativ tyngd (dvs. at fleire kjelder må visa bruken av eit gitt språkleg uttrykk), gjeld ikkje kvantitetskravet automatisk for talemålskjelder. I praksis kan eit dialektbelegg som berre har éi kjelde, danna grunnlag for ein ordboksartikkel dersom redaktøren etter ei fagleg totalvurdering kjem til at opplysningane er til å stola på og representerer ein etablert bruk.

Talemålskjeldene til *Norsk Ordbok* består i all hovudsak av setlar med innsamla målførebelegg. Desse setlane er kartotekkort med dialektopplysningar som kan heimfestast til eit geografisk område. Ei anna vesentleg kjelde er geografisk større eller mindre avgrensa lokale ordsamlingar. Alle setlane, og i tillegg ein del av desse ordsamlingane, er digitaliserte og gjort tilgjengelege. Andre ordsamlingar finst berre i papirversjon. I tillegg til desse to hovudkjeldene til opplysningar om talemålet, rår *Norsk Ordbok* også over ein talemålssynopsis (*Norsk dialektatlas*) som dokumenterer dialektal fonologi og morfologi for heile landet heilt ned på kommunenivå. Synopsisen og det innsamla arkivmaterialet er utarbeidd og/eller kvalitetssikra av skolerte målføregranskarar, og er derfor av god kvalitet. Trykte lokale målføresamlingar spring i mange tilfelle ut av eit større kollektivt arbeid utført av eit lokalt mållag, eit historielag eller språkinteresserte entusiastar med stor kunnskap om lokalmiljøet. I nokre tilfelle har innsamlarane formalisert språkvitskapleg skolering, i andre tilfelle ikkje.

Opplysningane i slike lokale målføresamlingar kan seiast å ha ei viss grunnleggjande kvalitetssikring ved at det ofte er fleire som har samla inn, vurdert og analysert materialet før samlinga er prenta. Når *Norsk Ordbok*-redaksjonen vurderer å bruka slike lokale målføresamlingar, blir det også gjort ein fagleg

kvalitetskontroll av innhaldet før samlinga eventuelt blir integrert i grunnlagsmaterialet.

På nettet finst det eit utal av private nettsider, bloggar og andre diskusjonsforum som inneheld informasjon om lokale målføre og dialektord. Korleis skal ein vurdera slike kjelder jamført med dei andre målførekjeldene *Norsk Ordbok* rår over? På same måte som med skriftmålskjeldene, kan det argumenterast for eit grunnleggjande skilje mellom dei kjeldene som også finst i ein prenta versjon, og som oftast har vore gjennom ein viss kvalitetskontroll, og meir tilfeldige kjelder på nettet som ein ikkje veit noko om kvaliteten på. Det er ofte uklårt kva slag kompetanse innsamlaren har, belegga er ikkje sanksjonerte i eit større brukarkollektiv, og dei er ofte ikkje korrekturlesne.

I slike tilfelle må ordboksredaktøren vera svært varsam, og til sjuande og sist vil målførekunnskapen til redaktøren vera avgjerande. Dersom opplysningane stadfester eller utvidar den geografiske dekninga til allereie kjende og kvalitetssikra data, eller dersom opplysningane kan stadfestast av kompetente fagfolk, treng ikkje publiseringsformatet i seg sjølv desimera den språklege opplysninga.

Konklusjon

Drøftinga vår munnar ut i ein konklusjon i tre punkt:

- For det første kan det ikkje understrekast sterkt nok at for ei vitskapleg dokumentasjonsordbok er kravet til etterprøvbarheit ufråvikeleg. Dette gjev spesielle utfordringar i møte med Internettet.
- For det andre er det eit faktum at dei institusjonaliserte innhaustingsordningane ikkje sikrar dokumentasjon av stendige endringar i nettpublisert tekst, og difor må lagringa skje lokalt.
- Sist, men ikkje minst, må det presiserast at kvaliteten på nettkjeldene heile tida må vurderast grundig. Nettet er ein publiseringskanal for tekst på line med andre kanalar, og nett-tekst bør ikkje av prinsipp haldast borte frå den leksikografiske granskinga.