Lars Jørgen Tvedt, Elisabeth Lien og Øyvind Eide
Universitetet i Oslo

Ordbokshotellet – varig lagring og formidling av norske ordsamlingar

Norske ordsamlingar er normalt prenta bøker med dialektord frå eit forholdsvis lite område. Forfattarane er lokale eldsjeler som ønskjer å dokumentera den lokale dialekten. Interessa for dialektar er stor i Noreg, og det vert årleg utgjeve fleire ordsamlingar.

Dei lokale ordsamlingane vert vanlegvis utgjevne i små opplag, og er ofte berre tilgjengelege i lokale utsal, eller gjennom historielag eller kulturforeiningar. Ordsamlingane er ein viktig dokumentasjon av den lokale språk- og kulturhistoria, men sidan det lokale også er viktig nasjonalt, bør dei gjerast lettare tilgjengelege for folk over heile landet.

Ordsamlingane er også ei viktig kjelde for det nasjonale ordboksverket Norsk Ordbok (Almenningen, 1994), ikkje minst fordi dei ofte kan dekkje hol i tidlegare innsamla materiale (Jenstad, 2005).

Eining for digital dokumentasjon (EDD) har i mange år arbeidd med samlingane og datasystema til Norsk Ordbok. Etter kvart som ein fekk meir røynsle med digitale arkiv og samanknyting av desse, vart det også aktuelt å digitalisere ordsamlingane. Målsettinga var å lagre desse i eit ope dataformat tilgjengeleg for både fagfolk og publikum.

Ingen ordsamlingar er like, men vi kunne ikkje bygge eitt system for kvar ordsamling. Vi måtte lage noko generelt, og etter mal av omgrepet "web-hotell", tok vi i bruk arbeidstittelen "ordbokshotell", som vi definerer slik:

> *"Eit ordbokshotell er ei teneste for elektronisk lagring, indeksering og publisering av innhaldet i ordsamlingar og ordbøker."*

I denne artikkelen går vi først gjennom bakgrunnen for ordbokshotellet, og så ser vi på dei tekniske sidene av prosjektet.

1. Digitalisering av kjeldematerialet til Norsk Ordbok

Norsk Ordbok er eit ordboksprosjekt med mål om å dokumentere det norske talemålet og det nynorske skriftmålet i eit 12-bandsverk.

Arbeidet med ordboka starta i 1930. Til grunn for det opphavlege arbeidet med ordboka ligg Grunnmanuskriptet, to setelarkiv med til saman 3,2 millionar belegg frå norsk talemål og nynorsk litteratur, og diverse andre ordbøker og ordsamlingar.

Første bandet kom i 1960, og band 3 kom i 1994. Utgjevingstakten var låg, og i eit samarbeid mellom Universitetet i Oslo og norske styresmakter satsa ein stort på ei revitalisering, som inneber både ein større redaktørstab, digitalisering av eksisterande kjeldemateriale, oppbygging av nytt kjeldemateriale, og utvikling av digitale reiskapar for redigering.

Digitalisering av setelarkiva starta på 1990-talet i regi av Dokumentasjonsprosjektet (noverande EDD), og i 2002 vart prosjektet Norsk Ordbok 2014 (NO 2014) oppretta. Målet var å fullføre 12-bandsverket innan år 2014. Året var valt fordi dette er 200-års-jubileet for den norske grunnlova.

Overgangen til ei databasebasert redigering har ført til at ordboka no står fram som ei elektronisk ordbok, der det er langt større rom for å dokumentere innhaldet i artiklane ved direkte kopling til det digitale kjeldematerialet, og NO 2014 ønskjer å dra nytte av dette ved å publisere ordboka og kjeldene på web (Bakken, 2006).

Som ei følgje av at ein no kunne kople dei digitale kjeldene direkte til definisjonsstrukturen, vart innsamling og organisering av digitale ordsamlingar straks meir aktuelt.

2. Innsamlingsarbeidet

Oppbygginga av ein database over lokale ordsamlingar starta med innsamling av manuskript frå forfattarane. Dei siste åra er det blitt samla inn over seksti elektroniske manuskript. Dette har skjedd ved direkte kontakt via telefon eller brev der forfattarane vert informert om verdien av ordsamlingane sett frå NO 2014 si side.

NO 2014 kan ikkje tilby økonomisk godtgjering for ordsamlingane, men kan tilby å gjere dei meir verdfulle ved å legge til ytterlegare opplysingar, som til dømes normerte former av oppslagsorda. I tillegg kan NO 2014 tilby sikker lagring i eit ope format. Dette vert nedfelt i skriftlege avtalar som regulerer korleis NO 2014 kan bruke materialet.

Dei aller fleste forfattarane NO 2014 har kontakta, har vore positive til å gje frå seg materialet sitt. Dei har alle lagt ned eit stort arbeid i å dokumentere daglegspråket og kulturarven på heimstaden sin, og dei fleste er glade for at deira arbeid vert brukt som kjeldemateriale for eit nasjonalt, vitskapleg ordboksverk. Mange ser det også som ei ære å verte sitert i Norsk Ordbok. Dei fleste ser også det positive i at arbeidet dei har lagt ned vert sikra gjennom trygg lagring i databasar.

3. Opphavsrett og publisering

NO 2014 kan ikkje utan vidare leggje ut lokale ordsamlingar på web på same viset som med anna kjeldemateriale. Ordsamlingane er verna av norsk *Lov om opphavsrett til åndsverk* (2005). Åndsverklova slår fast at skaparen av eit verk har einerett til å styre over eksemplarframstilling og tilgjengeleggjering av verket. I norsk lov oppstår denne retten automatisk når eit verk har såkalla verkshøgd. Verkshøgd får ein når det ligg eit sjølvstendig, kreativt arbeid bak eit verk. Eineretten til å styre over eit verk går ut 70 år etter at opphavspersonen er død.

I norsk rett er det grenser for denne eineretten. Ein har lov til å framstille eksemplar til privat bruk, undervisning og liknande, og det er lov å sitere frå eit verk.

NO 2014 har ei sjølvsagt plikt til å handtere ordsamlingane på ein slik måte at ein tek omsyn til opphavsretten i åndsverklova. Norsk Ordbok kan sitere ordsamlingane i artiklane sine, men målet om å publisere ordboka kopla til kjeldematerialet på web kan ikkje oppfyllast utan særskilt løyve frå kvar ordsamlingsforfattar.

4. Vitskaplege konsekvensar av å samle samlingar

Den vitskaplege nytten av å samle einskildbidrag til faget i større samlingar er opplagt. Det er heller ikkje noko nytt at NO 2014 samlar, indekserar og katalogiserar ordsamlingar. Tidlegare har ein nytta setelarkiv som indeksar for slike ordsamlingar. Ei digital samling av målføresamlingar må difor minst gje den same funksjonaliteten, helst litt meir.

Det er gjort eit stort arbeid med innhenting av ordsamlingar ved Norsk Ordbok. Vi har også laga rutinar og malar for å standardisere dokumentformatet for ordsamlingane. Dette er dei første, og kanskje viktigaste tenestene som er knytt til ordbokshotellet. Dette arbeidet sikrar lagring av ordsamlingane i eit ope format som gjer det mogleg å hente fram informasjonen også i framtida. Vi ønskjer likevel å gå lenger, og har difor utvikla eit digitalt system for felles indeksering og presentasjon av innhaldet i ordsamlingane.

Ordsamlingane er laga for å dokumentere særskilte målføre, og målgruppa er hovudsakleg brukarar av målføra eller personar som vil orientere seg om dei einskilde målføra. Ein konsekvens av at ordsamlingane dekkjer ulike målføre og har ulike målgrupper, er at dei varierer mykje både i omfang, innhald og presentasjon. Vi prøver å ta omsyn til dette når vi kodar samlingane, slik at vi tek vare på så mykje som mogleg av informasjonen i kvar samling. Men for å få nytte av dei einskilde samlingane, krev vi alltid at vi kan markere minst eitt oppslagsord i kvar artikkel.

Når vi no vel å samle ulike ordsamlingar i eitt system, riv vi artiklane laus frå den opphavlege samanhengen, og dei inngår i ei større eining. Dei framstår som sjølvstendige artiklar i ei ordsamling som etter kvart skal femne om heile det norske folkemålet. Med dette endrar vi både målsettinga med samlinga og målgruppa for henne. Frå å vere ordsamlingar som skal dokumentere einskilde målføre, vert det no meir aktuelt å leite på tvers av målføra for å finne likskapar og ulikskapar. Denne overgangen krev informasjon som normalt ikkje ligg i dei opphavelege samlingane. Dei viktigaste informasjonskategoriane vi treng er normaliserte oppslagsord, og normaliserte ordklassenemningar. Dette er tilleggsinnformasjon som må registrerast av fagleg kompetente folk.

Med denne informasjonen på plass har vi utvikla eit system der vi kan søkje på kryss og tvers av alle ordsamlingane etter målføreformer, normaliserte former og ordklasseopplysningar, i tillegg til informasjon om målføret artikkelen omtalar.

5. Populærvitskaplege konsekvensar

Vi håper med tida at det skal verte mogleg å publisere mesteparten av ordsamlingane fritt på web. I Noreg er det alt ei stor interesse for språk

og målføre, og det å gjere slike samlingar tilgjengelege for eit stort publikum vil sannsynlegvis auke denne interessa. Det er som nemnt tidlegare ein del juridiske problem knytte til fri elektronisk publisering, men vi håper å kunne legge til rette for at ei slik publisering kan gjerast slik at også interessa for å kjøpe dei trykte samlingane aukar. Ved å referere til utgjevar på web-sidene, og også informere om korleis ein kan få kjøpt papirutgåvene av ordsamlingane, håper vi å kunne publisere fleire samlingar på web.

6. TEI-koding, eit ope tekstformat

Når ei ordsamling kjem inn til oss, kan ho vere i mange ulike dataformat. Nokre er laga i moderne verktøy, tekstbehandlarar som Microsoft Word og satsprogram som FilemakerPro eller QuarkXPress, mens andre har lege i nokre år sidan dei var ferdige, og kjem i eldre format som MacWrite 3.

Eitt av måla våre er å ta vare på desse samlingane for framtida. Då er det ikkje nok å ha filer på ei kompaktplate eller eit platelager. Vi må òg kunne lese filene i framtida på ein rask og enkel måte.

Vi tek vare på filene i den forma dei kjem inn til oss, i alle slags ulike format. Men vi lagar òg filer som vi enkelt skal kunne lese i framtida. Slike filer må lagrast i eit format som må vere ope, slik at ein ikkje må lite på ein leverandør av programvare. Det må også vere eit format mange andre brukar, og det må vere rikt nok til å kunne lagre ordsamlingane på ein meiningsfull måte. Sist, men ikkje minst viktig, må det vere eit format som er følgt av god dokumentasjon.

Text Encoding Intitative (TEI) står bak eit arbeid som har vore i gang sidan slutten av 1980-talet. Dei har utvikla eit sett retningsliner for korleis ein skal kode dokument, særleg innan språk- og kulturfaga. Desse retningslinene er bygd på standarden XML. TEI har eit eige kapittel som syner korleis trykte ordlister og ordbøker kan kodast. Eit slikt koda dokument er bygt opp av to viktige delar. Den eine er den bibliografiske delen, som kallast “TEI Header”. Den inneheld bibliografiske opplysningar om den digitale teksten, kva for ei papirutgåve han er bygt på, ansvarlege personar, utgjevar, og anna.

Den andre delen av eit TEI-dokument er sjølve teksten i boka. Han er bygt opp slik at ein skal kunne sjå kva dei ulike delane er, ikkje berre korleis satsbiletet tok seg ut i utgåva. Eit døme vil gjere det

enklare å skjøne dette. Det er henta frå ei ordsamling frå Eidsvoll-området (Ljødal, 2002). Først artikkelen omlag slik han ser ut i den trykte boka, så TEI-versjonen av same artikkelen.

Apal **F** , appall *m* →	Apal, epletre. Gn: *apaldr.*

Døme frå ordsamling

```
<entry id="EidsvollLjoedal_orig34">
      <form type="simple">
            <orth n="0">Apal</orth>
            <usg n="0"><hi rend="bold">F</hi></usg>,
            <orth>appall</orth>
      </form>
      <gramGrp>
            <pos><hi rend="italic">m</hi></pos>
      </gramGrp>
      <def>Apal, epletre. Gn:
            <hi rend="italic">apaldr.</hi>
       </def>
```

TEI-koding av artikkelen

Her ser ein at heile artikkelen ligg i eit XML-element som vert kalla **entry**. Den har òg ein **id**-verdi som ein kan vise til frå andre stader. I elementet **form** finn ein opplysningar om oppslagsordet. Dei to oppslagsformene finn ein i to **orth**-element. Det første har ein **F** knytt til seg i artikkelen, som er ein opplysning om at forma vert nytta i Feiring kommune. Dette finn ein att i **usg**-elementet.

Etter dette kjem dei grammatiske opplysningane, som ligg i **gramGrp**. I dette dømet er det ordklassa **m**. Ho vert lagt i elementet **pos** (engelsk "part of speech"). Til slutt har vi ein definisjon som er plassert i **def**-elementet.

TEI-standarden har mykje som ikkje er vist i dette dømet, og ein kan òg legge til element eller endre oppskrifta på andre måtar om ein treng det for sine data. Det som er viktigast for oss er at ein på ein dokumentert måte kan vise kva dei ulike elementa i ei ordsamling er for noko. Når ei ordsamling er lagra slik, vil ein med TEI-dokumentet, databeskrivinga og dokumentasjonen lett kunne sjå kva slag informa-

sjon dei ulike datatypane inneheld. For oss er dette viktigare enn å kunne skape på nytt ordsamlinga slik ho såg ut i sin originale sats. Vi tek jo vare på dei trykte samlingane, og ved ei mogleg publisering kan ein skanne ordsamlingane og legge dei ved som faksimile, om ein ønskjer det. Samstundes er det slik at vi tek vare på opplysningar som kursiv og feit i **hi**-element. Dette er kodar som er knytte tett opp til oppsettet av kategoriane i ordsamlinga, og det gjer samlinga lettare å lese når ho kjem inn i applikasjonen.

Ei mogleg digital ivaretaking av utforminga må vi vurdere på nytt om og når hotellet tek inn ordsamlingar som ikkje er utgjevne på papir i det heile, som webdokument.

7. Normeringsarbeid

Vi har så langt sett på kva som er gjort av arbeid med ordsamlingane for å kunne lagre desse digitalt for framtida. Det vidare arbeidet med artiklane er hovudsakleg knytt til ønskje om å kunne indeksere og søkje på normalisert informasjon.

I prinsippet skil ikkje denne oppgåva seg nemneverdig frå det å lage eit papirbasert setelarkiv basert på ekserpering av ordsamlingane. Måten vi gjer sjølve normeringa på i ein elektronisk kvardag, må likevel gje oss ein del ekstragevinstar. Vi har forsøkt to ulike framgangsmåtar.

Dei første ordsamlingane vi klargjorde vart påførte ekstra XML-kodar for å registrere kva som var normerte former. Denne jobben vart gjort manuelt i tekstfilene, og utan anna støtte enn eventuelle ordbøker. Vi har så langt klargjort fire ordsamlingar på denne måten, og lagt dei inn i ordbokshotellet.

Erfaringa vår med dette er at denne metoden er tidkrevjande. Det er tungt å redigere XML-koda filer, og det viste seg etter kvart at dei som stod for normeringsjobben ikkje alltid hadde sams oppfatting av kva som var rett normering av eit ord.

Vi er difor i gang med å prøve ut ein annan reiskap som er utvikla for Norsk Ordbok 2014, nemleg Metaordboka (Ore, Tvedt, Bjørnstad, 2002). Metaordboka er eigentleg ei lemmaliste, der kvart lemma i teorien kan ha ulike normerte former knytte til seg, og kvart lemma kan ha peikarar til eit sett artiklar i ulike ordbøker, ordlister, eller andre objekt som skildrar ord, t.d. digitaliserte setlar frå eit setelarkiv.

Eit døme på ein "artikkel" i Metaordboka kan vere ordet "innanfor". I Metaordboka ligg det informasjon om at dette er ein preposisjon som er samansett av "innan" og "for". Artikkelen peikar så til 122 setlar i setelarkivet til Norsk Ordbok, til to artiklar i Nynorskordboka, til fem setlar i arkivet til Trønderordboka, til ein artikkel i Skards ordliste, og til ein artikkel i Grunnmanuskriptet. Metaordboka er klargjort for at ein kan legge inn alternative normeringar, t.d. kan ein normere ordet til bokmål. Med denne enkle tilleggsinformasjonen kan ein indeksere dei 130 belegga også på bokmål.

Til Metaordboka er det laga eit normeringsverktøy. Dette er eit enkelt system som gjer det lett å endre dei normerte formene av eit lemma, og det er enkelt å flytte koplingar mellom ulike artiklar i Metaordboka.

Vi har teke i bruk Metaordboka til å normalisere ordbokshotellet ved å utvide Metaordboka til å kunne peike på artiklar i ordbokshotellet. Ved hjelp av ei automatisk kopling/generering av artiklar i Metaordboka basert på informasjonen i ordbokshotellet, får vi eit første utgangspunkt for normalisering. Oppgåva til den som skal normalisere ordbokshotellet vert no å kontrollere dei automatiske koplingane som vert gjort, eventuelt å kople saman nygenererte artiklar med Metaordboksartiklar som alt er normaliserte. Fordelen med å gjere dette er at ein med enkle klipp-og-lim-funksjonar på datamaskinen kan foreta normeringa, og at den som skal gjere jobben har tilgang til 500.000 tidlegare normerte former. Sjansen for at ein får ulik normering av orda vert difor langt mindre.

8. Søking, presentasjon og datautveksling

Ordbokshotellet er implementert i ein relasjonsdatabase som vert drifta av EDD. Det er også kopla mot eit generelt søkje- og presentasjonssystem som er utvikla ved EDD. Dette systemet er utvikla slik at ein lett kan publisere innhald i databasar både i eit spesialutvikla Microsoft Windows-basert system, og på web. Systemet kan velje å passordbeskytte einskilde databasar, og førebels er ordbokshotellet ikkje tilgjengeleg for alle på grunn av opphavsrettsproblem.

Søkjesystemet er mogleg å endre ved enkle grep, men førebels er det laga eit system for ordbokshotellet som gjer det mogleg å søkje

etter normaliserte ordformer, målføreformer og målførenamn. Dette følgjer ein standard som brukarane av EDD sine databasar kjenner att.

Så langt har vi basert oss på å presentere artiklane i HTML, kodesettet for framvising av dokument på web. Dette gjer vi ved å transformere TEI-dokumenta ved hjelp av Extensible Stylesheet Language (XSL). Dette er eit språk for å transformere mellom ulike XML-dokument. Dei siste versjonane av HTML er eit delsett av XML, og ein transformasjon mellom TEI og HTML er difor ganske enkelt å definere. Brukaren av ordbokshotellet vil difor sjå ein normal web-vising av artiklane frå hotellet.

Det er mogleg å knytte særskilte XSL-dokument til kvar einskild ordsamling i hotellet, slik at ein kan spesialsy framvisinga for kvar samling slik at ho mest mogleg samsvarar med den originale presentasjonen.

9. Kopling mot Norsk Ordbok

Arbeidet med ordbokshotellet hadde sitt utgangspunkt i dei behova som Norsk Ordbok hadde for digitale kopiar av dette materialet, og sjølvsagt har ein skjegla til dette under utviklinga av ordbokshotellet.

Ved å kople einskildartiklane til Metaordboka vert innhaldet i ordbokshotellet ein del av datagrunnlaget for artiklar i Norsk Ordbok. Redigeringssystemet til Norsk Ordbok gjer det mogleg å referere direkte til grunnlagsmaterialet i dei artiklane redaktørane skriv, og redaktørane har verkty for å sortere og ordne materiale før artiklane vert utarbeidde. På dette viset vil alle artiklar frå ordsamlingane verta direkte tilgjengelege for redaktørane under arbeidet med Norsk Ordbok.

10. Annan bruk av hotellet

Teknologien som ligg bak ordbokshotellet kan med fordel utvidast til å handtere andre typar ordlister enn målføresamlingar. Døme på dette kan vere terminologisamlingar, fagbøker som t.d. floraer, eller andre typar bøker. Om dette skal inngå i det same hotellet, eller om ein bør opprette parallelle system avhengig av typen informasjon, må ein vurdere.

11. Konklusjonar

Vi har så langt gode erfaringar med eit system som gjer bruk av opne standardar og standard databaseteknologi. Når dette er teke i bruk innanfor avanserte løysingar for redigering av ordbøker, og er ein del av publiseringsløysingar for web, har leksikografane fått eit nytt hjelpemiddel som vil auke den faglege integriteten til ordboka.

I ordbokshotellet ligg det no 11 samlingar med til saman 23.000 artiklar. Det verkar som om løysinga vi har valt for å normalisere samlingane fungerer, og samlinga vil vakse jamt og trutt i åra som kjem.

Vi håper at vi seinare via web kan tilby samlingane for folk flest.

Litteratur:

Almenningen, Olaf (1994): Ordsamlinger på dialekt – problem og utfordringar. i: Anna Garde og Pia Jarvad (red.): *Nordiske Studier i Leksikografi II*. Rapport fra Konference om Leksikografi i Norden 11.-14. maj 1993. København. 9-17.

Bakken, Kristin (2006): The Dictionary and its Sources: The Ideal of Integration and the Example Norsk Ordbok. i: Elisa Corino, Carla Marello, Cristina Onesti (eds.): *Proceedings XII EURALEX International Congress*. Volume I. 117-122.

Dokumentasjonsprosjektet: http://www.dokpro.uio.no/

Grønvik, Oddrun og Lars Jørgen Tvedt (2006): Norsk Ordbok 2014 – presentasjon av eit komplekst leksikografisk verktøy. i: Henrik Lorentzen og Lars Trap-Jensen (red.): *Nordiske Studier i Leksikografi 8*. Rapport fra Konference om Leksikografi i Norden, Sønderborg 24.-28. maj 2005. København. 143-150.

Jenstad, Tor Erik (2005): Norsk Ordbok som dialektordbok. *Nordiske studiar i leksikografi 7*. Rapport frå Konferanse om leksikografi i Norden, Volda 20.-24. mai 2003. Oslo. 221-227.

Ljødal, Ola (2002): *Artukt me dei gamle talæmåta. Eidsvolldialekten – ord og uttrykk*. Moelv.

Norsk Ordbok 2014: http://no2014.uio.no/

Lov av 12. mai 1961 nr 2 om opphavsrett til åndsverk m.v. (Åndsverkloven): med endringer, sist ved lov av 17. juni 2005 nr 97 (i kraft 1. juli 2005), Oslo.

Ore, Christian-Emil og Lars Jørgen Tvedt (2006): Digital sats eller digital satsing? i: Henrik Lorentzen og Lars Trap-Jensen (red.): *Nordiske Studier i Leksikografi 8*. Rapport fra Konference om Leksikografi i Norden, Sønderborg 24.-28. maj 2005. København. 315-322.

Ore, Christian-Emil Smith, Lars Jørgen Tvedt og Tone Bjørnstad (2002): *The Meta Dictionary*. ALLC/ACH 2002; 24.07.2002 - 28.07.2002.
http://www.edd.uio.no/artikler/leksikografi/meta_dictionary.html

TEI, *Text Encoding Intitiative*: http://www.tei-c.org/

XML, *Extensible Markup Language*: http://www.w3.org/XML/