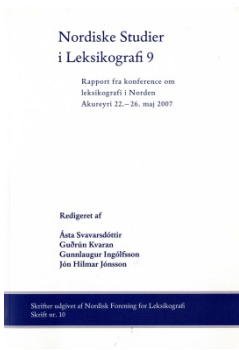


NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Materialsortering på digital plattform - Eit steg mot høgare dataintegritet i den vitenskaplege leksikografien	
Forfatter:	Kristin Bakken og Oddrun Grønvik	
Kilde:	Nordiska Studier i Lexikografi 9, 2008, s. 31-42 Rapport fra Konference om leksikografi i Norden, Akureyri 22.-26. maj 2007	
URL:	http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive	

© Nordisk forening for leksikografi

Betingelser for brug af denne artikel

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

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

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.

Materialsortering på digital plattform

Eit steg mot høgare dataintegritet i den vitskaplege leksikografien

For Norsk Ordbok 2014 linguistic documentation comes in the forms of scanned paper slips, electronic slips and corpus quotations, organised through the index system called the Meta Dictionary. Each entry in Norsk Ordbok is directly linked to the corresponding materials in the Meta Dictionary. However, with up to several thousand instances to sort into a complex entry structure, editors needed a sorting tool for classifying instances and storing the sorting itself. This would both improve editorial stringency, allow for variation in approach method and save time. This article, entitled *Classifying instances of linguistic documentation on a digital platform – one step further towards data integrity within academic lexicography*, describes the tool developed for Norsk Ordbok to meet this need, and sums up user experiences through the first two years as positive but still with space for improvement.

1. Innleiing

I 2002 blei redaksjonen ved Norsk Ordbok omorganisert. Omlegginga innebar nyordning både når det galdt administrasjon og struktur, finansiering, faglege løysingar og ikkje minst digitalisering og utvikling av nye verktøy. Éi viktig erfaring vi har gjort desse fem første åra, er at dei faglege omleggingane og alle tiltaka som har med digitalisering å gjere heng uløyseleg saman. Vi vil her sjå nærare på det verktøyet som vi i redaksjonen av Norsk Ordbok kallar *Sorteraren*. Dette er kanskje det beste dømet på korleis utviklinga av tekniske løysingar hos oss både heng tett saman med faglege prioriteringar og får faglege konsekvensar.

Eit krav redaksjonen hadde på seg i 2002 var å dra nytte av at kortarkivet vårt alt var digitalisert. Under eit slikt krav låg sjølvstøtt primært ønsket om effektivisering. Dei første åra vart det då utvikla ein redigeringsapplikasjon som hadde direkte kontakt med det digitaliserte arkivet via den digitale indeksen for språksamlingane, kalla Metaordboka (for nærare omtale sjå Ore 2000). Ordutval og dimensjonering blir no gjord delvis automatisk med utgangspunkt i det digitaliserte arkivet, og underlagsmaterialet er tilgjengeleg på skjerm. I teorien treng altså ikkje lenger redaktørane forhalde seg til skuffar med arkivmateriale. Men eitt problem stod att: korleis sortere digitale arkivkort? Mange av redaktørane

såg seg framleis nøydde til å sortere underlagsmaterialet på skrivebordet i små bunkar med arkivkort. Dette problemet blei akutt for den delen av redaksjonen som arbeider ved universitetet i Trondheim, for det hadde ikkje tilgjenge til det fysiske arkivet, og var såleis nøydde til å bruke dei digitaliserte arkivkorta, som dei berre fekk sjå eitt for eitt på skjermen.

Mot denne bakgrunnen gjekk vi i dialog med utviklarane ved Eining for digital dokumentasjon. Vi ville ha ein digital sorterar som løysing på eit praktisk problem. Sorteraren skulle vere eit tenleg verktøy i materialanalysen, og helst skulle ein spare tid i denne arbeidskrevjande fasen av det redaksjonelle arbeidet. Vi ønskte også å hente kvalitative gevinstar, dvs. å sikre at kvar og ein gjorde så godt empirisk grunnarbeid som mogleg. Metodisk såg vi her ei moglegheit for at Norsk Ordbok i større grad kunne bli empirisk etterprøvbar, med andre ord at dataintegriteten kunne bli høgare med ein digital sorterar på plass. Til sist var det svært viktig for oss at verktøyet vi etablerte ikkje innebar ei einsretting av arbeidsformene. I forkant hadde vi nemleg kartlagt korleis redaktørane faktisk arbeider med materialet. Vi hadde sett at dei arbeider forbausande ulikt. Og her vil vi understreke at dette dreier seg om heilt grunnleggjande tilnæringsmåtar som heng saman med personlegdomen, og at det er umogleg å seie at den eine tilnæringsmåten er "betre" enn den andre. Dette kjem vi tilbake til, i denne omgangen held det med å understreke at vi ønskte å ta vare på denne metodiske diversiteten. Vi ville at det nye sorterarverktøyet skulle bli eit fleksibelt verktøy som kunne tilpassast ulike, individuelle arbeidsformer. Det måtte for all del ikkje bli ei metodisk tvangstrøye. Dette var då ønsket vårt: Vi ville ha eit fleksibelt verktøy som opnar for metodisk diversitet, som sparar tid, hevar kvaliteten og aukar dataintegriteten.

2. Løysing: Eit funksjonsrikt system

2.1 Plassering i NO-systemet

IKT-løysinga for Norsk Ordbok 2014 er eit system av databasar og IT-verkty til ulike føremål. Norsk Ordbok blir redigert i ein applikasjon som matar data rett inn i ein database. Det er lagringsstaden for ordboksartikkelen, som skal vere resultatet av ei mogleg, men ikkje ei obligatorisk sortering. Tilfanget kjem frå Metaordboka og blir hekta rett på artikkelhovudet, og på denne måten er det direkte samband mellom Metaordbok og ordboksdatabase.

Sorteraren er eit tilbod, eit slags siderom på dette sambandet, der ein kan ordne og annotere tilfanget. Sorteraren ligg derfor utanfor redigeringsapplikasjonen og ordboksdatabase. Sorteraren er tilgjengeleg berre frå applikasjonen, via den einskilde artikkelen.

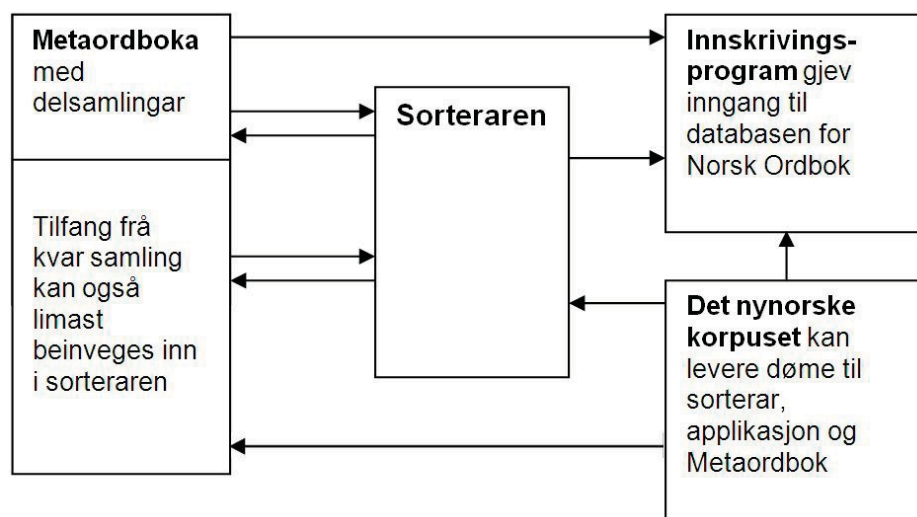
2.2 Utforming

Først eit ord om ein tanke vi forkasta. Det hadde vore mogleg å lage eit simultert "skrivebord" på skjermen, der ein brukte musa til å dra belegg over til rett "bunke". Dette ville vere visuelt attkjenneleg. Men setelbunkane har aldri vore noka ideell løysing. Dei har den veikskapen at ein berre ser øvste setel, og er avhengig i postit-lappar på bunkane for å minnst sorteringskriterium. Det var også eit poeng å unngå musbruk og halde seg mest mogleg til tastaturet.

Vi valde i staden ei løysing som tek vare på den mentale sida ved sorteringsprosessen, nemleg sjølve kategoriseringa. Poenget er då å kunne (1) merkje eller annotere kvar setel, (2) sjølv velje sorteringskriterium, (3) kunne merkje opp setlane etter så mange kriterium som ein vil, (4) kunne sortere framlengs og baklengs etter eitt eller fleire kriterium samstundes, (5) lagre sorteringa som grunnlag for framtidig arbeid.

Sorteraren ser ut som eit rekneark, der venstre kolonne listar opp alle belegg linje for linje. Kvar belegg frå Metaordboka er representert med eit symbol. Når ein klikkar på symbolet kjem setelen opp til venstre på skjermen. Ein kan leggje til kolonnar (for sorteringskriterium) og gje dei tittel etter eige skjøn i høve til kvart ord.

Ein stor føremon ved denne arbeidsmåten er at når tilfanget er ferdig annotert og sortert ser ein straks kva ein har i kvar "bunke", og ein ser heilskapen i sitt eige sorteringssystem. Sorteraren gjer dermed tilfanget langt meir oversiktleg enn ein nokon gong før har hatt det.



Figur 1. NO 2014: Kopling av tilfang til og frå sorteraren. Pilene syner kva veg tilfang kan koplast.

3. Resultat: erfaringar og evalueringar

Intensjonen ein har med ein applikasjon blir ikkje alltid realisert av brukarane slik ein hadde tenkt. Etter knappe to år med sorteraren på plass var det då svært interessant for oss å kartleggje bruken av han. 25 redaktørar i to ulike byar har fått eit tilbod om å bruke sorteraren. Dei har ikkje fått noko *påbod* om det, for i utgangspunktet skulle sorteraren dekkje eit uttalt behov hos nokre, men det var jo slett ikkje alle som hadde etterlyst ein digital sorterar. Etter knappe to år er det akkumulert 1446 artiklar som har sorterarar knytte til seg. Det er opplagt at redaktørane i ulik grad og på ulik måte har tatt dette verktøyet i bruk.

Artiklar med sorterar	færre enn 5	10–60	60–100	fleire enn 100
redaktørar	4	13	4	4

Figur 3. Redaktørane ordna i grupper etter kor mange artiklar dei har oppretta ein sorterar for.

Ved å gå databasen etter i saumane kan ein finne ut ganske mykje om bruken av sorteraren. I tillegg har vi intervjuet nokre av redaktørane i kvar gruppe for å få svar på desse spørsmåla:

- Når opprettar ein redaktør ein sorterar?
- Korleis blir sorteraren nytta?
- Sparar ein tid?
- Er kvaliteten på arbeidet blitt høgare?
- Er dataintegriteten blitt betre?

Når ein analyserer svara på desse spørsmåla, ser ein om intensjonane bak utviklinga av sorteraren er blitt realiserte.

3.1 Når opprettar ein redaktør ein sorterar?

Som venta er det det slik at ein sorterar blir oppretta når ein redaktør skal redigere eit ”stort” ord, dvs. eit ord med stort materiale bakom seg i arkivet. Dette hadde vi rekna med, for ein når jo ei smertegrense for kor mykje informasjon ein kan greie å ha oversyn over i hovudet. På eit eller anna tidspunkt må ein lage seg ”hugselappar”. Men det vi ikkje hadde rekna med, er kor store forskjellar det er på kva redaktørane reknar som ”store” ord, med andre ord kor mykje materiale den einskilde tykkjer det er greitt å halde oversyn over utan hjelpemiddel. Nokre definerte eit stort ord som eit ord med meir enn 20 belegg attom seg i arkivet, medan andre definerte eit stort ord som eit ord med fleire enn 250 belegg attom seg.

Ein tilleggsgrunn for fleire var forventninga om kompleks tydingsstruktur. Om ein raskt ser at ordet kjem til å bli redigert med mange tydingsnummer fordi tydingsstrukturen er kompleks, opprettar redaktøren ein sorterar sjølv om ordet ikkje har svært mange belegg attom seg i arkivet. Det same gjeld for ord som må

homografseparerast, dvs. at belegga under eitt oppslag i arkivet faktisk skal fordelast på to forskjellige homografar og dermed to forskjellige oppslagsord.

Til sist nemnde fleire at typiske dialektord, dvs. ord med stor målførevariasjon, ofte får knytt ein sorterar til seg. Norsk Ordbok skal jo omfatte både den normerte skriftlege nynorsken og dei munnlege dialektane. Det gjer ein ved å ordne og føre opp dialektvariasjon under kvart oppslag. I mange tilfelle er formvariasjonen svært stor, og i slike tilfelle blir sorteraren nytta for å kartleggje denne typen av variasjonsmønster.

Ikkje uventa ser det altså ut som om ein sorterar blir oppretta for å behalde oversyn over komplekst materiale. Fleire seier eksplisitt at dei alltid vurderer tidsbruken ved å opprette ein sorterar. Der dei trur dei greier å halde oversynet utan sorterar, lèt dei vere å opprette ein, for dei trur altså at dei sparar tid på denne måten. Det må også seiast at dei som relativt rutinemessig opprettar ein sorterar for alle ord som har meir enn 20-30 belegg attom seg, har integrert sorterarbruken i den vanlege arbeidsrutinen sin som redaktør. Vi kjem tilbake til erfaringane desse redaktørane har gjort seg.

3.2 Korleis blir sorteraren nytta?

Svaret på dette spørsmålet var det ikkje så lett å lese ut av databasen, men etter å ha intervjuet ulike redaktørar, teiknar det seg eit bilete av to ganske forskjellige måtar å bruke sorteraren på. Interessant nok er det ikkje eintydig om desse to ulike måtane fordeler seg på anten storbrukarane eller dei som brukar dette verktøyet minst.

Vi har valt å karakterisere dei to bruksmåtene for ei induktiv tilnærming og ei deduktiv tilnærming. Den reinaste induktive metoden blir av ein av brukarane også kalla "notatblokkmetoden". Redaktøren går her til materialet utan førehandsoppfatningar om struktur, og opprettar ein sorterar "nedanfrå" så å seie. Det vil blant anna seie at kategoriane som vert oppretta får ganske innhaldstunge overskrifter, og at samstundes med at materialet blir gjennomgått, blir kategoriseringssystemet raffinert og justert ved at kategoriar blir delte eller at nye kjem til. Redaktøren har ikkje primært fokus på eit ryddig sluttprodukt, og det blir typisk nok ikkje eit ein-til-ein-forhold mellom artikkelstrukturen og strukturen i sorteraren. Om redaktøren skulle ønskje å kople relevant grunnlagsmateriale til kvart tydingsnummer i artikkeltreet, krev det ein eigen ryddeprosess til slutt. Dei redaktørane som arbeider slik, viser til fordelane ved å gå til materialet utan førehandsoppfatningar, for slike oppfatningar kan kanskje vere galne og stå i vegen for den analysen som høver beste med det materialet faktisk viser. Dei meiner nok også at metoden yter materialet størst rettferd.

Den reinaste deduktive metoden går ut på at ein i forkant har laga seg eit grovt kart over forventa relevant semantisk og dialektbasert variasjon, og at ein opprettar ein sorterar med kategoriar som speglar desse forventa variablane. Slike hypotesar om struktur blir etablerte på grunnlag av grunnmanuskriptet til Norsk

Ordbok (ein kompilasjon av alle dei eksisterande ordbøkene over nynorsk pr. 1940), og på bakgrunn av andre ordboksverk, grammatikkar og dialektoversyn.

Med denne arbeidsforma blir sorteraren nytta til å teste og justere hypotesar som er formulerte på førehand, og ein relaterer sin eigen materialgjennomgang meir eksplisitt til tidlegare arbeid. Metoden gjer det lettare å etablere eit ryddig sluttprodukt, og den gjev typisk nok eit større samsvar mellom sorterarstruktur og ordboksstruktur. I neste omgang blir det lettare å kople materialet i sorteraren til nodane i artikkelstrukturen.

Vi vil likevel understreke at i praksis er det sjølvsagt oftast slik at metodane i nokon mon glid over i kvarandre. Dei som primært arbeider deduktivt får plutseleg auge på eit særtrekk i materialet som gjer at dei må justere den opphavlege strukturen. Typisk nok er det kanskje dei faste ordsambanda, dvs. dei konvensjonaliserte kollokasjonane som slik "materialiserer seg" undervegs. På den andre sida er ikkje dei som arbeider mest induktivt heller utan føresetnader, og det er sjølvsagt slik at dei konsulterer tidlegare arbeid før dei set i gang. Dei lèt likevel tidlegare arbeid i mindre grad styre eigen analyse.

Det som er heilt klart er at intensjonane våre med å lage eit fleksibelt verktøy som opnar for metodisk mangfald, er oppfylde. Reint konkret er det slik at nokre nummerer tydingane i sorteraren med éin gong, t.d. "1", "2", "2a", "2b" osv. Andre gjev tydingane titlar i form av meir innhaldstunge stikkord, t.d. "lokativ", "temporal", "durativ" osv. Nokre plukkar berre ut målføresetlane og markerer dei t.d. med "Målf.", medan andre annoterer dei ulike dialektformene for seg. Somme gonger blir tekstdøme merkte med å føre opp kollokat, og såleis blir det lagt til rette for ei strengt kontaktbasert sortering. Nokre tilpassar sorteraren slik at dei kan føre inn alle heimfestingane, og alle belegg frå eldre kjelder, medan andre igjen markerer alle idiom. Nokre hakar av dei belegga dei tykkjer er gode kandidatar for å bli tatt med i den endelege ordartikkelen som sitat. Nokre brukar mange kolonner eller fleire sorterarar for å halde orden på alle informasjonstypene, medan andre heller lagar ulike innhaldstunge annoteringar innanfor same kolonne i sorteraren. Felles for alle systema er prinsippet om at all konsekvent annotering gjer det mogleg å søke fram dei aktuelle delmengdene av materialet etterpå.

Dette inneber ei raffinering av tilfanget som også aukar eigenverdien av ord-samlingane for seinare forskning. Om eit stort underlagstilfang for eit einskildord er merkt opp i sorteraren med kollokat, til dømes, blir det enkelt å plukke ut og bruke dei setlane til andre føremål seinare.

Den store fleksibiliteten systemet opnar for, er slik vi ser det ein stor føremon. Det gjer det mogleg for den einskilde redaktøren å vidareføre si eiga arbeidsform. På den andre sida kan det tenkjast at dette opne systemet kan framstå som krevjande for dei av redaktørane som ikkje, eller i svært liten grad, har tatt sorteraren i bruk.

3.3 *Sparar ein tid?*

Det er viktig å nemne at fire av redaktørane så å seie ikkje har tatt sorteraren i bruk. Dette er fire av dei eldste redaktørane (over 60 år), og sams for dei er at dei viser til at dei for lenge sidan har etablert gode rutinar som fungerer for dei. Dei har halde fram med å sortere papirsetlar på skrivebordet, og meiner det ville føre til unødig tidsbruk å forsere den tekniske terskelen som sorteraren ber med seg. Dei forventar altså ikkje at bruk av sorteraren skal gje nokon tidsgevinst. I samband med denne undersøkinga har vi sett at vi bør bruke litt meir tid på å gje innføring i bruk av sorteraren for dei som argumenterer med tidkrevjande teknisk terskel.

Bortsett frå desse fire redaktørane meiner alle dei andre at sorteraren gjev tidsgevinstar. Som alt nemnt gjeld dette for mange med visse atterhald. Sorteraren gjev tidsgevinstar for visse (store) ord, men ikkje generelt, og redaktørane gjer slike vurderingar av moglege tidsgevinstar frå ord til ord.

Tre tidssparingspunkt blir nemnde. For det første ligg det potensiell tidssparing i det å ikkje miste oversynet over ei stor materialmengd. Utan merking er faren der alltid for at ein må gjere operasjonar fleire gonger fordi ein ikkje finn ting att, eller fordi ein gløynde noko i førre runden. Sorteraren hjelper slik til med at ein raskt kan orientere seg i materialet og utan omvegar finne fram til relevante døme.

For det andre er det svært viktig for mange at sorteraren gjer det mogleg å annotere og kategorisere etter fleire kriterium samstundes. Når ein sorterer arkivbelegg i haugar på eit skrivebord, blir det eit problem om det finst fleire måtar å kategorisere materialet på som ikkje er synkroniserte. Typisk nok er dette tilfellet med dialektbasert, kjeldebaserst eller heimfestingsbasert kategorisering i møte med semantiske kategoriar. Papirløysinga er å sortere i fleire omgangar. Ulempen med denne metoden er at sorteringane blir borte etter kvart. Ein får ikkje lagra og slik rekapitulert sorteringane om ein har bruk for det. Med den digitale sorteraren kan ein sortere etter like mange kriterium som ein i utgangspunktet har brukt i annoteringa. Og ein kan samsortere etter fleire kriterium, t.d. etter heimfestingsstad og dialektform, eller etter dialektform og tydingsnummer. Ved dei store orda gjev dette vesentlege tidsgevinstar.

Det faktumet at desse sorteringane blir elektronisk lagra saman med artikkelen dei utgjer det empiriske grunnlaget for, blir no utnytta i opprettings- og kontrollfasen. Dette er det tredje punktet der mange redaktørar viser til tidssparing. Når ein kontrollør skal vurdere materialanalysen, eller ein i opprettingsfasen ikkje lenger hugsar kva grunnlaget var for vala ein har gjort, er det ein føremon at materialgjennomgangen faktisk er lagra. Både i funksjon av å vere redaktør som skal rette opp eigne manuskript eller i funksjonen som gjennomlesar som skal kvalitetssikre manusa til andre, peikar redaktørane på tidsgevinstane ein sorterar fører med seg.

Vi meiner då at det er riktig å konkludere med at bruk av sorteraren gjer at ein sparar tid. Men det føreset altså at ein brukar sorteraren slik som no, dvs. etter ei ganske gjerrig vurdering av tidsbruken når ein vel å opprette ein sorterar i utgangspunktet. Eit par av dei som i størst grad har integrert bruk av sorteraren i arbeidsrutinen sin, ser tidsgevinstane i opprettingsfasen som så viktige at dei også vel å opprette sorterarar til etter måten små ord.

3.4 Er kvaliteten på arbeidet blitt høgare?

Alle dei som brukar sorteraren meiner kvaliteten på arbeidet er blitt høgare. Dei viser til litt ulike element i kvalitetsomgrepet. Éin viktig grunn for mange er at dei kjenner at det empiriske grunnlaget for den analysen ordboksartikkelen er eit uttrykk for, er meir solid. Dei er tryggare på at dei har tatt omsyn til alle nyansane i materialet, og sikrare på at dei ikkje har gått glipp av vesentlege døme eller kollokasjonar. Éin redaktør viste til at sorteraren hjelper til med å etablere ei naturleg rekkjefølgd på tydingsnumra. Ein får lettare oversyn over kor prominente dei ulike tydingane er i materialet, og reint visuelt får ein indikasjonar på kva for tydingar som er primære og sekundære.

Dei som arbeider mest induktivt, held det fram som ein kvalitativ føremon at ein kan møte materialet utan hypotesar og slik yte det betre rettferd. Det kan kanskje lesast ut som at den induktive arbeidsforma gjer redaktørarbeidet til eit meir sjølvstendig vitskapleg forskararbeid, med resultat som er etablerte med uavhengig evidens i forhold til tidlegare ordboksarbeid. Vi vil understreke at det same gjeld for ei moderat deduktiv tilnærming. Om ein er open for å justere og endre dei hypotesane om struktur ein hadde då ein byrja å sortere materialet, vil ein ved hjelp av sorteraren gjere empirisk grunnlagsarbeid uansett.

Eit svært viktig punkt i høve til dataintegritet er det som har med kvalitetskontroll å gjere. Det er heilt opplagt at kontroll og oppretting av eigne manuskript er blitt betre fundert. Ein slepp å byrje på nytt med materialet når ein kvalitetskontrollerer manuskripta sine. I staden kan ein lett ta ein titt i sorteraren om det er noko ein lurar på. Det same gjeld for dei som kvalitetsikrar andre sine manuskript. Spesielt viktig er kanskje dette for redaktørar som er i ein opplæringsfase. Ein svært viktig del av opplæringa er jo å lære seg å analysere materiale rett. No kan dei som har ansvar for manuskripta til nybyrjarane, på ein lettvint måte også kvalitetssikre materialanalysen deira. Dette er blitt eit viktig punkt i redaksjonen vår, for så langt har vi i desse fem åra hatt nye redaktørar under opplæring kontinuerleg. Sjølv om vi no nærast har full bemanning, ser vi at det i ein stor redaksjon vil vere utskifting jamleg, og då gjeld det å gjere opplæringa av nye redaktørar mest mogleg effektiv.

Vi ser slik at det som i utgangspunktet var tenkt som ei løysing på eit praktisk problem, faktisk også har gitt kvalitetsgevinstar vi ikkje hadde sett så klart for oss då vi byrja med arbeidet. Vi *hadde* sett for oss at det ville bli lettare å skaffe seg oversyn og verre å gå glipp av relevant materiale, men vi hadde ikkje så klart sett

kva for gevinstar dei lagra sorteringane ville ha i kontroll- og kvalitetssikringsarbeidet.

3.5 Er dataintegriteten blitt betre?

Norsk Ordbok er ei vitenskapleg ordbok. Grunnen til det er at ein tek sikte på å kartleggje og dokumentere og analysere ord som anten ikkje er kartlagde før i det heile, eller som ikkje er kartlagde på denne måten før. Dokumentasjon er her eit nøkkelord. I ei dokumentasjonsordbok skal ein ved hjelp av materiale eller kjelder underbygge dei påstandane om ordtyding, utbreiing og formell variasjonsbreidd ein kjem med i ordartiklane. Eit krav til god vitenskap er at den er empirisk etterprøvbar. Det vil seie at dei påstandane eller konklusjonane ein legg fram, skal kunne prøvast mot grunnlagstilfanget, og avkreftast om tilfanget er brukt urett. Det skal vere mogleg å finne og leggje fram motprov mot den konklusjonen eller den påstanden ein kjem med.

I eit vitenskapleg dokumentasjonsordbok er påstandane eller konklusjonane som ein legg fram om språk, implisitt formulert. I den ferdige ordartikkelen vil tydingshierarkiet som ein har ordna tydingane til ordet i typisk nok vere ein slik konklusjon. Ein annan konklusjon vil vere det oversynet over distribusjonen av dialektvariantar som blir presentert. Ein tredje vil vere det sambandet som er etablert mellom oppslagsform, språkhistorisk opphavsform og dialektvariantar i artikkelen, det vil seie sjølv avgrensinga av lemmaet. I mange tilfelle vil dette sambandet bli etablert for første gang i ei ordbok av vår type som handterer både målføre, normert tale og skrift.

Spørsmålet blir då om sorteraren har gjort Norsk Ordbok meir *vitenskapleg* enn ho var før. Ein annan måte å stille dette spørsmålet på er å spørje seg om dataintegriteten er blitt betre. Med dataintegritet meiner vi her at det er eit godt samsvar mellom konklusjonane om materialet i ordboka og materialet sjølv. Og vi vil leggje til – dette samsvaret skal også kunne etterprøvast.

Det viktigaste einskildet elementet i den nye Norsk Ordbok-redaksjonen som betrar dataintegriteten, er faktisk ikkje sorteraren, men Metaordboka. At det digitaliserte arkivet gjennom Metaordboka er direkte kopla til den elektroniske ordboksdatabasen, gjer at ein kan gå fram og tilbake mellom ordartikkel og arkiv, og slik *etterprøve* ordartiklane. Dette er ein revolusjonerande ny måte å tenkje dokumentasjonsordbok på, og det er ikkje før Norsk Ordbok er tilgjengeleg for alle som elektronisk ordbok, at dette potensialet blir fullt ut realisert. For oss i redaksjonen er det likevel allereie ein realitet, vi har ei direkte kopling mellom empiri og ordartikkel, og vi utnyttar dette til etterprøving i det daglege kvalitetssikringsarbeidet.

Sorteraren representerer i denne samanhengen ei raffinering av denne koplinga mellom materiale og ordbok. Koplinga mellom materialet i Metaordboka og artiklane i ordboksdatabasen er jo ikkje nærare spesifisert og fortolka. Den som skal gå analysen etter i saumane må i og for seg gjere heile det leksikografiske

grunnlagsarbeidet om igjen. Sorteraren bøter på dette, for i sorterarane er det jo nettopp spesifiserte og fortolka analysar av tilfanget som blir lagra saman med artikkelen. Ved å kople sorterarane til ordboka, ja gjerne til dei enskildte nodane i artikkelstrukturen, er det etablert ei moglegheit for etterprøving på meir detaljert nivå. I dei tilfella der ein sorterar er kopla direkte mot kvar node i artikkelstrukturen, er kvar node empirisk fortolka mot eit større materiale enn det illustrerande dømet som før fekk ivareta denne funksjonen.

Før vi konkluderer på dette punktet, må vi understreke at undersøkinga vår har vist at denne eksplisitte oppkoplinga mellom nodane i ordartiklane og sorteraren til vanleg ikkje blir gjort. Med eitt unnatak er det ingen redaktørar som tek den ekstrajobben det er å kople sorteraren opp mot ordartikkelen, node for node. Hovudgrunnen har med tidsbruk å gjere. Slik oppkopling tek tid som ikkje vil gje gevinstar for produksjonen av papirutgåva. Omsynet til ei framtidig elektronisk utgåve er ein luksus redaktørane ikkje unner seg med dei krava til framdrift som dei har å forhalde seg til. Den andre grunnen er at dei som i størst grad brukar sorteraren som ei nedanfrå-og-opp notatblokk, ikkje har etablert ein så tydeleg struktur i sorteraren at ein utan vidare kan kople den opp mot ordboksdatabase. Her må det verkeleg eit reelt tidkrevjande ryddearbeid til før slik kopling kan gjerast.

Idealet om auka dataintegritet er først og fremst realisert på den digitale plattformen i og med Metaordboka. Opprettinga av sorteringar i samband med artiklane i databasen inneber at etterprøvarheita er realisert på eit meir spesifisert nivå. I og med oppkoplingsmoglegheita som ligg i systemet, kunne ein tenkje seg ei endå tettare og meir fortolka kopling mellom materiale og ordartikkel, men av grunnar som har med tidsbruk å gjere, er det berre éin redaktør som gjer slik oppkopling. Det er viktig å understreke at dette idealet om dataintegritet berre lèt seg realisere på den digitale ordboksplattformen.

4. Konklusjon

For oss er det opplagt at ei digital kopling mellom materiale og ordbokslemma via Metaordboka gjev auka dataintegritet. Det er blitt mogleg å etterprøve det redaksjonelle arbeidet ved å gå fram og tilbake mellom ordboksdatabase og det digitalisert arkivet. I tillegg meiner vi at dei lagra sorterarane som no i mange tilfelle er hekta på ordboksartikkelen i databasen, gjer dei konklusjonane som ordboksartikkelen er eit uttrykk for, etterprøvbare på eit endå meir spesifisert nivå. I og med sorteringane er materialet fortolka, og desse fortolkingane er lagra. Det høgaste nivået av integrering mellom materiale og ordbok får ein ved å kople sorteringane til artikkeltreet node for node. Dette er ei moglegheit som redaksjonen så langt i liten grad nyttar seg av.

Det er slik eit stort arbeid som ligg framfor oss om den moglegheita vi no har fått til å integrere kjelder og ordboksdatabase skulle realiserast gjennomgåande.

I ein framtidig digitalisert versjon av ordboka kunne vi eksplisitt kople sorteringsresultatet til nodane i artikkeltreet slik at materialanalysen også blei tilgjengeleg for brukarane. Men dette set store krav til sorteringane. Dei må då vere ryddige og ferdiganalyserte med ein struktur som er i samsvar med den ferdige artikkelen. Dette inneber heilt klart eit stort meiarbeid. Om ein verkeleg skulle ha som ambisjon at brukarane av ordboka ved alle ord skulle få tilgjenge til oppkoplale materiale, ligg det dessutan ei kjempestor utfordring på metodesida. Etter vår meining måtte ein då uniformere sorterarbruken i større grad enn no, og dette ville gripe radikalt inn i arbeidsformene til den einskilde. Vi ser altså at dei ideelle krava til dataintegritet er i konflikt for det første med ønsket om effektiv tidsbruk og for det andre med ønsket om individuelt tilpassa verktøy og metodisk mangfald.

Vi vil likevel konkludere med at dei intensjonane vi hadde når vi sette i gang arbeidet med å utvikle ein sorterar, langt på veg er realiserte. Vi har fått eit fleksibelt metodisk hjelpemiddel, verktøyet har gitt tidsgevinstar, kvaliteten på ordboksproduktet er blitt høgare og til sist, dataintegriteten er blitt betre. Vi vil halde fram med å oppmuntre redaksjonen til å realisere det potensialet for betre metode som ligg i sorteraren, men vi vil av omsyn til metodisk fleksibilitet og effektiv tidsbruk førebels ikkje fullt ut realisere den optimale etterprøvbarheita for eksterne brukarar.

Litteratur

Ore, Christian-Emil Smith 2000: Metaordboka. I: *Ord om Ord* 6, 30–32.

Kristin Bakken
prosjektdirektør, f. 1959
kristin.bakken@iln.uio.no

Oddrun Grønvik
hovudredaktør, f. 1947
oddrun.gronvik@iln.uio.no

Norsk Ordbok 2014
Universitetet i Oslo
postboks 1021 Blindern
NO-0315 Oslo