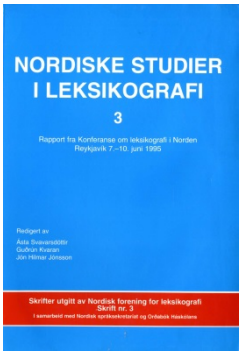
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Korpus og seddelarkiv, fredelig sameksistens mellom det beste og det gode?

Forfatter: Christian-Emil Ore

Kilde: Nordiske Studier i Leksikografi 3, 1995, s. 331-338
Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995

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.

Christian-Emil Ore

Korpus og seddelarkiv, fredelig sameksistens mellom det beste og det gode?

This paper focuses on how to save the information in a traditional citation archive for the computerized world. The sample archive is a citation archive at the Department of Lexicography at the University of Oslo, Norway. Originally the archive was to be keyed and given a SGML markup. Due to reasons discussed in the paper this approach was abandoned. The archive is now converted into a indexed facsimile data base which seems to be a better solution both from a scientific and from an economic point of view. The research, the program development and the actual conversion of the information in the archive forms a sub project of the Documentation project, which is a collaboration between the four Norwegian universities.

1 Innledning

Seddelarkivet er den tradisjonelle systematiske metoden for å samle belegg og opplysninger om ord til bruk i redigeringen av en ordbok. Korpuset og de tilhørende KWIC-konkordanser med ulike ordningskriterier er blitt muliggjort i stor stil ved hjelp av datamaskinene. Den nye teknikken har skapt et skille i bruk av materialet i ordboksredigeringen. For noen står seddelarkivet som en tilfeldig samling opplysninger eller deler av en konkordans, og arkivet er kun et eksempel på hva gårsdagens teknikk kunne produsere. For andre representerer seddelarkivet en skattekiste der hver seddel er valgt med omhu, mens det med et korpus kan være vanskelig å få med det spesielle og sjeldne språket. Etter forfatterens mening er korpusmetoden en åpenbart mer rasjonell måte å samle språklige data på.

I denne artikkelen vil det imidlertid bli fokusert på hvorledes en kan redde informasjonen i et tradisjonelt seddelarkiv over i den datamaskinelle verden. Det vil bli tatt utgangspunkt i Dokumentasjonsprosjektets arbeid med det 3,2 millioner store seddelarkivet til Norsk Ordbok ved Avdeling for leksikografi ved Universitet i Oslo. Jeg har tidligere gjort rede for prosjektet og for arbeidets gang i form av foredrag ved leksikografikonferansene i Oslo og København, Ore (1991) og Ore (1993).

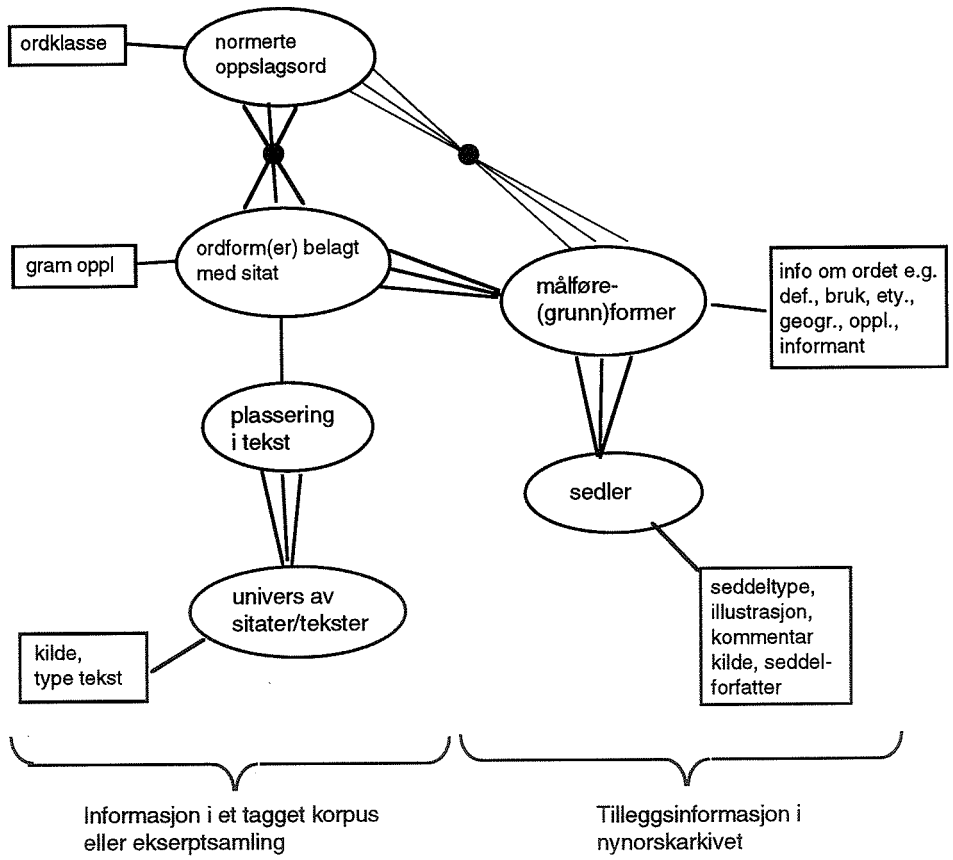
2 Det opprinnelige prosjektet

Informasjonen i et seddelarkiv kan på mange måter sammenlignes med den vi finner i et tagget tekstkorpus. I arkivet er det for et utvalg ord fra et utvalg tekstbrokker gitt opplysninger om bøyningsformen, ordets rot og annet. I et tagget tekstkorpus finner en gjerne de samme opplysningene knyttet til hvert enkelt av ordene i større tekstfragmenter (fra 25 sider løpende tekst og oppover). Denne observasjonen lå til grunn da vi planla konverteringen av nynorskarkivet. Vi ønsket å erstatte de rene ekserptsedlene med løpende elektronisk tekst :

De resterende sedlene skulle skrives inn og SGML-kodes (Sperberg-McQueen et al. 1994) slik at de kunne lastes inn i et databasesystem slik det er skissert i figur 1.

På bakgrunn av tidsstudier utført ved Nynorskavdelingen høsten 1989, ble avskrivingsarbeid og korrektur anslått til om lag 250 årsverk. Nynorskprosjektet ville dermed utgjøre en tredjedel av Dokumentasjonsprosjektet.

Konverteringsarbeidet begynte sommeren 1992 i Sigdal, Lier og Røyken, og i Mo i Rana i oktober samme år. Ved årsskiftet 1992/1993 var alt konverteringsarbeid flyttet til Mo i Rana. Konverteringsarbeidet ble gjort ved bruk av skannere og programmer for optisk lesning for de løpende tekstene. Sedlene ble skrevet av og kodet ved bruk av et vanlig tekstbehandlingsprogram.



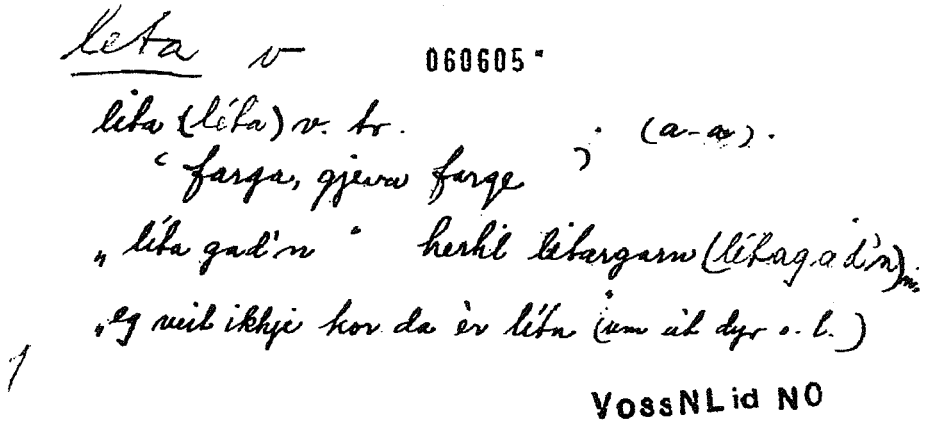
Figur 1: *Skisse av et kombinert seddelarkiv og tagget tekstkorpus*

2.1 Problemer forbundet med den opprinnelige konverteringsmodellen

Det vil alltid være en avveining mellom hvilket tempo en ønsker, og kompleksiteten i arbeidet som skal gjøres. I 1993 erfarte vi at arbeidet gjort i Mo i Rana hadde en akseptabel

kvalitet, men at det gikk for langsomt. Hastigheten var om lag halvparten av den forventede.

Nynorskavdelingen tillot ikke at vi sendte originalsedlene til Mo i Rana. Av den grunn måtte sedlene fotokopieres. Fra innskrivingshold ble det hevdet at fotokopiene var så dårlige at det sinket arbeidet kraftig. Fra andre hold ble det hevdet at kodeskjemaet var for komplekst.



Figur 2: *Faksimile av seddel*

Det er klart at det var en god del å hente ved enten å bruke originalsedlene eller bedre kopier. Men kopieringen og nummerering av sedlene er ganske kostbar, om lag 50 øre pr. seddel eller kr. 1 600 000 for hele arkivet. Det var derfor lite tiltalende å investere mer i en ren produksjon av papirkopier av sedlene som i neste omgang skulle kastes.

Kodingen av kortene var i løpet av 1993 blitt forenklet, og det er nok mulig at tempoet kunne ha økt ved at innskriverne kodet mindre mens de faglærte assistentene kodet mer under korrekturlesingen. På den annen side viste det seg at innskriverne gjorde relativt lite kodefeil. En ytterligere forenkling av kodesystemet ville etter min mening ha undergravd hensikten med å konvertere sedlene til søkbar elektronisk tekst.

Det er neppe riktig å si at dårlige kopier og komplisert tagging var hele årsaken til den lave produksjonen. Man må også huske på at oppstarten av registreringen av nynorskmaterialet falt sammen med oppstarten av en helt ny type sysselsettingstiltak. Det har altså vært satt i gang to helt nye prosjekter på en gang.

På bakgrunn av tidsstudier utført ved Nynorskavdelingen høsten 1989, ble avskrivingsarbeid og korrektur anslått til om lag 250 årsverk. Tatt i betraktning prosjektets størrelse var det umulig å skaffe ressurser til veie for å kunne fullføre det med halv hastighet. Det var derfor nødvendig å foreta en gjennomgang av prosjektet for å se hvordan det var mulig å få et akseptabelt resultat med de tilgjengelige ressurser.

3 En omlegging av registreringsarbeidet

3.1 Tekst og faksimile

Med elektronisk tekst menes det her den måten tekst lagres på i en datamaskin i for eksempel WordPerfect eller Microsoft Word. Hvert tegn i en tekst lagres som et tall på disken. Det er en kompakt og enkel lagringsmetode for vanlig tekst. Metoden gjør det mulig å søke etter bestemte tegnstrenger og å redigere teksten slik vi alle kjenner det fra vårt daglige arbeid med tekstbehandlere.

Hvis målet kun er å lese teksten, er det strengt tatt ikke nødvendig å lagre annet enn et bilde av teksten. Et bilde av en tekstside tar imidlertid mye mer lagringsplass enn en elektronisk tekstversjon, i beste fall om lag 100 ganger så stor plass. Men dersom selve innleggelsesprosessen for bildene er hurtig og rimeligere enn avskrift, kan kostnadene ved denne ekstra lagringsplassen lett forsvares.

Da Dokumentasjonsprosjektet ble startet, fantes det verken tilgjengelige høyhastighetsskannere eller tilstrekkelig rimelig lagringsplass. En inntasting av innholdet på sedlene var derfor det eneste mulige alternativet. I løpet av de seneste årene har imidlertid dette bildet endret seg totalt. Våren 1994 viste det seg at det var billigere å få skannet og lagret elektroniske faksimiler av alle sedlene enn å gjennomføre den nevnte utsorteringen og fotokopieringen av sedlene. Lagringsplass er fremdeles relativt kostbar, men vil antakelig innen prosjektets avslutning i 1997 være så rimelig at det er mulig å la alle de 3,2 millioner sedlene ligge „on line" samtidig.

<NSET NR=60605>
<OPPF GRM=v>leta</OPPF>
<ORDF>lita<ORDF>léta</ORDF>
<KOM>tr</KOM><BFORM>a - a
<DEF>farga, gjeva farga
<SIT T=USET>
<ORDFS GR=inf>líta</ORDFS>gad'n<KOM>heril litargarn (lítagad'n) n.
</KOM></SIT>
<SIT T=SETN>eg veit ikje kor da èr <ORDFS GR=prp>líta</ORDFS><FORK>um eit dyr o.l.</FORK></SIT></ORDF>
<KJEL>VossNLid NO

Figur 3: *Kodet versjon av seddelen vist i figur 2*

3.2 Tekstdatabase eller tekst/bildedatabase

Nynorsksedlene er laget av personer med og uten filologisk skolering over en 60 års periode og er svært forskjellige. Sedlenes eneste fellestrekk synes å være at de har et oppslag, en midtdel og en kilde. Det har vært hevdet at innskrivningen nettopp bør følge dette skjemaet. Det er lett å lære og enkelt å følge. Med dette skjemaet vil innholdet på sedlene bli tilgjengelig som søkbar elektronisk tekst. I tillegg vil oppslagsord, ordklasse og kilde være innganger til materialet.

Virksomheten i Nynorskprosjektet de to siste årene har imidlertid gitt en rekke erfaringer om problemer i registreringsarbeidet og om forhold i sedlenes utforming. På denne bakgrunn kan man spørre seg om det er verdt de ekstra kostnadene det er forbundet med å gjøre teksten på sedlene elektronisk tilgjengelig som elektronisk tekst i forhold til de lavere kostnadene ved

å lage en faksimiledatabase med oppslagsord, ordklasse og kilde som søkenøkler. Under er det listet opp seks sentrale punkter som taler mot en ren tekstbase med eller uten omfattende tagging:

1. Taggeskjemaet dekker innholdet på sedlene, men mange av sedlene er så lite systematisk satt opp, at en tagging krever en total innholdsfortolkning av hva som står på seddelen.

2. Det finnes flere lydskriftssystemer og mange ikke konsistente forsøk på slike, på sedlene.

3. En rekke av sedlene er ført med vanskelig leselig hånd. En avskrift blir dermed å likne med manuskripttranskripsjon.

4. Mange av sedlene i arkivet er det man kan kalle „huskelapp"-sedler. De inneholder kun informasjon om at det eksisterer et ord slik og sånn og hvor det er funnet brukt.

5. Om lag 25% av sedlene er basert på ekserpter fra ordlister, ordsamlinger og fra tett strekede bøker. I tillegg til det normerte oppslagsordet inneholder sedlene sjelden annen informasjon enn sitatet.

6. En rekke av sedlene har illustrasjoner.

Punktene 1–3 dreier seg om den formaltekniske kvaliteten på sedlene. Ved å registrere sedlene ved ord, ordklasse, ordform og kilde er man sikret at det alltid er mulig å hente frem originalsedlene og lese teksten slik den står. For „huskelapp"-sedlene vil man få registrert nettopp den informasjon de er ment å gi, nemlig koblingen mellom ord og kilde.

Det er liten hensikt i å lagre lydskrift som elektronisk tekst dersom det ikke finnes en konsistent definisjon av hvilke lyder tegnene angir. Det er vanskelig å formulere fornuftige søk i noe man ikke vet hva er. Det synes derfor som en bedre løsning å kunne hente bilder av de sedlene som inneholder lydskrift, eventuelt merke sedlene i basen slik at lydskriftsedler senere kan tas ut og behandles separat. På tilsvarende måte er det mulig å hente ut faksimiler av sedler etter en eller flere av inngangene oppslagsord, ordklasse, til dels også heimfesting, skriftlig kilde, tidspunkt for belegget og oppskriver (NO-medarbeider). Faksimilene av disse sedlene brukes så som punchegrunnlag for en utvidet registrering av seddelopplysningene der dette er ønskelig.

I de senere år har datateknikken muliggjort en mye mer effektiv oppbygging av den informasjonen som et seddelarkiv representerer. Elektronisk lesning av tekst (OCR), konkordansprogrammer og hjelpemidler for (halv-)automatisk markering av grammatisk informasjon på ord i løpende tekst kan nevnes. Det er derfor grunn til å vurdere om en avskrift av sitatene på sedlene nevnt under punkt 5, i noe fall ville vært formålstjenlig.

3.3 Registreringsarbeidet i dag

Beslutningen om å gå over til en faksimiledatabase ble fattet sommeren 1994. I løpet av det siste året er samtlige 3,2 millioner sedler sendt til Kodak Norge A/S for innskanning og mikrofilming.

I dag foregår registreringen av Nynorskarkivet ved hjelp av et spesiallaget program. Innskriverne får vist et seddelbilde på skjermen og skriver så inn oppslagsordet, ordklassen og kilden slik de klarer å tyde den. Deretter søker de i en kildeliste etter den „offisielle" betegnelsen på kilden og kobler denne mot seddelbildet. Resultatet er at hver seddel får knyttet til seg et oppslagsord, en ordklasse og en standardisert kildereferanse og også en såkalt „NO-medarbeider" (seddelskriver for Norsk Ordbok) der dette er oppgitt.

I forbindelse med fotokopieringen skulle disse rene ekserptsedlene uten tilleggsinformasjon og sedler basert på eldre ordbøker og ordlister bli frasortert. Underveis ble det klart at en slik utvalgsmetode har praktiske ulemper idet utvelgelsen av sedler tar lang tid. Å treffe riktige valg er vanskelig uten at arbeidet utføres av personer med inngående kunnskap om hele arkivet. En slik kunnskap er neppe å oppdrive for et arkiv bygd opp over 60 år av hundrevis av personer. Av den årsaken går nå alle sedlene gjennom innskrivningsprosessen. Det ville rett og slett koste for mye å foreta en sortering.

4 Konklusjon

4.1 Hva mister vi?

Faksimilemetoden gjør det umulig å søke i teksten på seddelen, da særlig etter ekstra informasjon om uttale, bruk, definisjon, bøyning og etter spesielle konstruksjoner og former i sitatene.

Det er intet å gjøre med sitatene. De kan kun nås gjennom kilder og oppslagsord. Men her kan man merke seg at mange av de litterære beleggene i arkivet er valgt ut etter mekaniske metoder. Enkelte leksikografer har valgt kun en forekomst fra hver bok og da gjerne den første regnet fra side 1. Beleggene er således ikke nødvendigvis de mest interessante. Videre har man ønsket å finne de ord som ikke forekommer ofte. Det kan her nevnes at under bokstaven A er det 25 sedler fra Arne Garborgs roman Bondestudentar. Om lag halvparten av disse er eksempler på ord som forekommer bare én gang i hele romanen. Det er videre i overkant av 120 lemma som begynner på A i denne romanen. Slik kan en muligens argumentere for at arkivet ikke kan betraktes som en sitatsamling som gir et representativt bilde av de omgivelser et ord blir brukt i.

Hva angår de andre kategoriene, viser det seg at en stor del av de sedlene som inneholder slik informasjon, kan gjenfinnes ved hjelp av kilde- og medarbeider-opplysningene. En grov måte å finne disse „interessante" sedlene på, er ved å søke etter sedler uten litterær kilde og som ikke kommer fra visse deler av Norsk målførearkiv. Denne hypotesen har vi verifisert ved å bruke de 120 000 sedlene som ble skrevet av og fullkodet som en fasit. Denne fasiten viser oss også at stedfesting av ordopplysninger på sedlene i over 80% av tilfellene er lik seddelskriverens hjemtrakter. Det er klart at slike grove mål ikke er fullgodt med en total gjennomgang av sedlene.

4.2 Hva vi har

De erfaringene som er høstet de siste årene, peker entydig i retning av at det er riktig å nøye seg med å lage en indeksert faksimilebase av sedlene supplert med en fullstendig innskriving av sedlene til noen få utvalgte medarbeidere. Rett nok finnes det ikke søkemuligheter for

bildene. Men selv bare den enkle registreringen av oppslagsord, ordklasse og kilde for hver seddel gir mange nye innganger til materialet. Arkivet kan dessuten brukes uansett om man sitter i Volda, Tynset, Oslo, Kirkenes eller Minnesota. Det kan om ønskelig også publiseres som 75 CD-plater.

Dette bør etter forfatterens mening representere den endelige avslutningen av arbeidet med å vedlikeholde seddelarkivet slik vi kjenner det i dag. I fremtiden burde man konsentrere seg om å bygge opp leksikalske databaser basert på elektroniske tekster og om mulig en systematisk landsomfattende undersøkelse av talemålet. De frivillige medarbeidernes virksomhet som nær sagt trøffelsvin kan man vel aldri klare seg uten. Men jeg synes det er overordentlig viktig at en slik virksomhet foregår etter mest mulig standardiserte metoder som kan gjøre det mulig å etterprøve resultatene.

Litteratur

Ore, C.-E. 1991: Dokumentasjonsprosjektet ved Det historisk-filosofiske fakultet, Universitet i Oslo. I: R. V. Fjeld (udg.): *Nordiske studier i leksikografi.* Rapport fra konferanse om leksikografi i Norden, 28.–31. mai 1991. Oslo: NFL, 403–408.

Ore, C.-E. 1993: Blant fire millioner sedler, En situasjonsrapport fra Dokumentasjonsprosjektet. I: A. Garde/P. Jarvad (udg.): *Nordiske studier i leksikografi.* Rapport fra Konference om Leksikografi i Norden, 11.–14. maj 1993. København: NFL, 243–247.

Sperberg-McQueen C. M./L. Bernard (eds.) 1994: *Guidelines for the Encoding and Interchange of Machine-Readable Texts (TEI P3).* Chicago/Oxford.