Seksjon for nordisk
språk og litteratur

NOR-*skrift*

Arbeidsskrift for nordisk
språk og litteratur

Nr. 97/1998

NOR*skrift*. **Arbeidsskrift for nordisk språk og litteratur** blir utgitt av Seksjon for nordisk språk og litteratur, Institutt for nordistikk og litteraturvitskap, Universitetet i Oslo. Spørsmål om abonnement kan rettast til Mette Rønning, telefon 22 85 62 27.

Adressa til redaksjonen er Seksjon for nordisk språk og litteratur
Institutt for nordistikk og litteraturvitskap
Boks 1013 Blindern
0315 Oslo

ISSN 0800.7764

FORORD

Ruth Vatvedt Fjeld og Boye Wangensteen ved Seksjon for leksikografi og målføregransking ved INL er gjesteredaktører for dette heftet av NORskrift. Artiklene utgjør i hovedsak en rapport fra et dagsseminar ved seksjonen 15.1.1998.

Redaksjonen
v/Arne Torp

INNHOLD

Leksikografiens rolle i det moderne kommunikasjonssamfunnet

RUTH VATVEDT FJELD & BOYE WANGENSTEEN

Dette nummeret av NOR-Skrift er et spesialnummer som inneholder fem av foredagene fra det seminaret som ble arrangert ved Seksjon for leksikografi og målføregransking 15.1.98 i forbindelse med professor Dag Gundersens 70-årsdag.

Dag Gundersen har mer enn noen i Norge i etterkrigstiden vært med på å utforme faget leksikografi. På seminaret var det meningen å gi seg tid til å se både bakover og framover og å drøfte hvordan faget best kan utvikles videre. Målet var å få ideer til hvordan det materialet som fins på seksjonen, best kan utnyttes for å dekke de krav som settes til faget i framtida, og for å møte de behovene for leksikografisk kunnskap som et moderne kommunikasjonsamfunn krever.

Leksikografisk arbeid har like lange tradisjoner som språkvitenskapelig arbeid ellers, men er av mange sett på som ikke-vitenskapelig filologisk arbeid. Faget har enten nærmest vært betraktet som et slags håndverk som hvem som helst kunne lære seg, eller som en kunstart man måtte ha spesielle nådegaver for å kunne utøve.

Som vitenskapelig fagdisiplin er leksikografien forholdsvis ung. Det har de siste 20-30 årene skjedd en intensiv teoriutvikling i faget, ikke minst på grunn av teknologiske nyvinninger som gjør selve datapresentasjonen mindre krevende, mens datainnsamling og bearbeiding og systematisering av dataene blir de viktigste oppgavene.

Mange har ment at det å lage en teoretisk beskrivelse av den leksikalske komponenten i et språks grammatikk og det å lage en beskrivelse av et språks leksikalske inventar, er to helt forskjellige og uavhengige språkvitenskapelige oppgaver. Dette synet har preget både den teoretiske beskrivelsen av leksikon og den praktiske leksikografien. Gramatikerne har ofte laget eksempler som skal illustrere og understøtte de reglene de har formulert, ut fra egen kompetanse og intuisjon (og har konstruert mye rart språk med det), ofte ut fra et vel

sterkt ønske om at den grammatiske beskrivelsen skulle gå opp. På den andre siden har leksikografene for en stor del arbeidet i et faglig isolat der et lite om enn systematisk innsamlet materiale har fått styre beskrivelsen av hele språkets ordforråd. Begge deler er uheldig. Systematisk innsamlet leksikalsk materiale kan gi gramatikerne bedre grunnlag for å finne fram til språkets regler. Gramatikerne på sin side har et oversyn over språkets helhet som bør gi leksikografene hjelp til å lage mer konsistente beskrivelser ut fra de innsamlede belegg og språklige data de har tilgjengelig.

Den moderne fagutviklingen foregår ikke i et vakum, men er avhengig av og videreutvikler de generelle teoriene innen annen språkvitenskapelig beskrivelse. Mye av de mest spennende nyvinninger i lingvistikk skjer i dag i miljøer med tilknytning til leksikografi. Å holde seg på høyden i moderne leksikografi forutsetter dermed innsikt i moderne lingvistisk teori og metode. For å komme så langt det er mulig i beskrivelsen av språkenes leksikalske enheter, er det derfor svært nyttig å samarbeide med andre språkvitenskapelige og informasjonsteknologiske fagmiljøer. Seminaret la opp til en markering av viktige forbindelseslinjer mellom deldisiplinene.

Leksikografenes nitide tekstgransking og løpende vurdering av de leksikalske enheter som til enhver tid opptrer i et språk, samt systematisk dokumentasjon av denne, kan likevel verken erstattes av automatiske analyser eller av generell regelbeskrivelse. Samarbeid over faggrenser og skjeling til andre disipliner betyr ikke at leksikografi ikke er en genuin og selvstendig fagdisiplin som trenger å utvikle sine egne teorier og metoder. Kjente forskere som arbeider med temaer som er leksikografisk relevante og utfordrende, ble derfor invitert til å holde foredrag. I det moderne vitenskapssamfunnet er det umulig å være seg selv nok, gode prosjekter forutsetter gode samarbeidspartnere. Seminaret viste både hvilke samarbeidspartnere vi allerede har, og det var også uttrykk for et "frieri" til institusjoner som vi ønsker å utvikle et nærmere samarbeid med.

Det er viktig for utviklingen og videreføringen av leksikografi som vitenskapsfag at leksikografene ikke bare er "passive" dataleverandører for andres prosjekter, men selv makter å utarbeide faglige prosjekter som kan drive faget videre.

Seksjon for leksikografi og målføregransking har som en av sine hovedoppgaver å dokumentere utvikling og status for norsk språk. Det skjer i stor grad ved hjelp av innsamling og systematisering av språklige data. Professor Dag Gundersen har fra 1962 ledet Det norske litterære ordboksverk, som i 1972 ble en egen avdeling ved Leksikografisk institutt ved Universitetet i Oslo, seinere en avdeling under Institutt for Nordistikk og litteraturvitenskap. Avdelingen er med den nye administrative strukturen ved fakultetet blitt borte, men kalles uformelt for "Subseksjon for bokmål". Faglig sett utgjør den en selvstendig enhet som har ansvaret for å bygge opp leksikalske arkiver over det norske bokmålet eller bokspråket fra 1500 til i dag. Det er samlet og arkivert mellom 3,5 og 4 millioner ekserpter, en samling som har og vil ha stor nasjonal og vitenskapelig verdi. Dag Gundersen gjorde i sitt innledningsforedrag rede for bakgrunnen for innsamlingen av nyordsmaterialet ved seksjonen, og for hvordan tilgangen til dataene er i dag. Ved hjelp av Dokumentasjonsprosjektet er hele basen nå blitt gjort elektronisk tilgjengelig.

Elektroniske hjelpemidler har lenge vært brukt i det leksikografiske arbeidet, men langt på vei er de elektroniske redskapene selvutviklet eller de er laget ved hjelp av eksterne konsulenter. Når konsulentene forsvant, levde hjelpemidlene sitt eget liv videre, noe som kunne føre til en viss utvikling, men også lett kunne gi stagnasjon. Leksikografer trenger en stadig oppdatering på metoder og modeller innenfor feltet automatisk tekstbehandling. Dokumentasjonsprosjektet har vært til uvurderlig hjelp for seksjonen på dette området. Christian Emil Ore gjør i sin artikkel rede for hvordan dette arbeidet har foregått, og presenterer tanker om hvordan basene kan utnyttes og videreutvikles i framtida.

Det leksikografiske arbeidet skjer i stadig større grad ved hjelp av it-teknologi, og vi har lenge følt savnet av bedre kompetanse i elektronisk databehandling og automatisert språkanalyse. Et samarbeid med Tekstlaboratoriet er derfor viktig og nødvendig.

Tekstlaboratoriet holder til i lokaler ved siden av Seksjon for leksikografi, og vi anser ikke den plasseringen som tilfeldig. Det er allerede tatt initiativ til et samarbeid, og ønsket er at samarbeidet skal

fortsette. Janne Bondi Johannessen gir i sitt bidrag flere gode eksempler på hvilken nytte vi kan ha av de metodene og modellene som er utviklet der, men også at avhengigheten går den andre veien, slik det skal være i et reelt samarbeid.

Seksjon for leksikografi er deltaker i NorKompLex-prosjektet, som ble satt i gang i 1995. Først og fremst har seksjonens bidrag vært å levere det leksikalske utgangspunktet gjennom basene til Bokmålsordboka og Nynorskordboka, men de produktene som prosjektet utarbeider, skal også være tilgjengelige for seksjonens videre arbeid. Torbjørn Nordgaards foredrag viser hvordan man i en leksikalsk base systematisk representerer verbenes argumentstruktur. Denne strukturmerkingen er bare noen av de mulighetene som prosjektet byr på, og bør bli styrende for noe av vårt fremtidige arbeid med det norske leksikon.

Seksjonen har lenge vært opptatt av å utvikle et tilbud i tospråksleksikografi, situasjonen på fakultetet er i dag slik at det lages tospråklige ordbøker ved de enkelte språkinstituttene på grunnlag av god kompetanse i to forskjellige språk, men ofte uten den nødvendige innsikt i leksikografi og metode. Arbeidet med tospråklige ordbøker foregår ofte som fritidsaktivitet for språkforskere i fremmedspråkene. Et nærmere og bedre faglig fundert samarbeid her ville være til fordel for alle parter. Videre arbeides det med relaterte problemer ved oversettingsenheten ved Seksjon for anvendt lingvistikk, og ved fakultetsprosjektet om parallellkorpus. Vi er derfor spesielt glad for at Stig Johansson kunne referere fra dette arbeidet og komme med noen ideer om hvordan parallellkorpus kan benyttes i utarbeidelsen av tospråklige ordbøker. Hans fremtidsvisjon av en kobling av ordbok, grammatikk og korpus er i høy grad sammenfallende med drømmer som ofte drømmes ved seksjonen og ved instituttet for øvrig (jf bl a Vannebos artikkel i festskriftet *Normer og regler)*.

Leksikografiens endelige mål er fortsatt å lage ordbøker og ordboksliknende produkter. Som et grunnleggende arbeid før igangsetting av leksikografiske prosjekter må det alltid gjøres grundige analyser av behovene for de produktene som prosjektene skal føre fram til. Seksjon for leksikografi har særlig siden den første konferansen i nordisk leksikografi ble arrangert i Oslo i 1991, hatt et utstrakt

samarbeid med de øvrige leksikografiske miljøene i Norden. Med tanke på igangsetting av nye tospråklige ordboksprosjekt mellom to eller flere nordiske språk, er Ilse Cantells foredrag et eksempel på hvordan slike forundersøkelser kan gjøres.

*

Dette nummeret av NOR-Skrift er bare delvis en rapport fra Dagsseminaret, da noen av foredragsholderne enten ikke hadde anledning til å omarbeide sine foredrag til artikler, eller hadde avtale om å trykke dem annetsteds. Det er likevel å håpe at de foreliggende artiklene gir et inntrykk av noen av de mange muligheter og oppgaver som ligger i fagfeltet, til inspirasjon både for leksikografer og andre som er interessert i kunnskap om språks leksikon.

Seksjon for leksikografi og målføregransking, Dokumentasjonsprosjektet og elektronisk leksikografi

CHRISTIAN-EMIL ORE, DOKUMENTASJONSPROSJEKTET

Seksjon for leksikografi og målføregransking (SLM) er den eneste leksikografiske avdelingen ved noe universitet i Norden. Alle de fire avdelingene som nå utgjør SLM, har i årenes løp bygd opp store samlinger med opplysninger om norske ord. Seksjonen har den desidert største samlingen av opplysninger om norsk språk.

Når man leser forordet til de ulike heftene av Norsk Ordbok som er kommet til nå, slår det en at oppbygningen av seddelsamlingen har vært svært sentral hele tiden helt til i dag. Seddelsamlingen er det tradisjonelle systematiske verktøyet for å samle belegg og opplysninger om ord til bruk i redigeringen av en ordbok. Et typisk kort i en slik samling inneholder et ord i grunnform, en liten tekstbit som gir et eksempel på en bruk av ordet, hvor ordet er brukt, opplysninger om grammatikk, uttale og eventuelle andre forhold rundt dette eksempelet. Samlinget er sortert alfabetisk etter ordenes grunnform. I de fleste ordboksredaksjoner har det vært et ønske om å få seddelsamlingen størst mulig for nettopp å kunne ha mange eksempler for hvert eneste ord, også de mer sjeldne. Seddelsamlingen til Norsk Ordbok rommer nå om lag 3,2 millioner sedler. Til sammen har SLM i underkant av 8 millioner sedler. Dette er ikke spesielt mye. Tilsvarende nasjonale prosjekter i Sverige og Danmark har mer en 10 millioner hver, mens arkivet til "Oxford English Dictionary" rommer mer enn 30 millioner sedler.

I de siste 40-50 årene har imidlertid datateknologien muliggjort alternative metoder for å fange inn og lagre tilsvarende informasjon. Den nye teknikken har skapt et skille i synet på bruk av materialet i ordboksredigeringen. For noen står seddelsamlingen som en tilfeldig samling opplysninger og er kun et eksempel på hva gårsdagens teknikk

kunne produsere. For andre representerer seddelsamlingen en skattkiste der hver seddel er valgt med omhu.

Ved SLM ble datateknologien introdusert på slutten av 1970 tallet gjennom arbeidet med Nynorskordboka og Bokmålsordboka, to håndordbøker med henholdsvis om lag 90000 og 65000 oppslagsord. Det ble utarbeidet innskrivnings- og kodeformater for ordboksmanuskriptene og for det såkalte Avis- og tidskriftarkivet over nyord i bokmål (Nyordsarkivet). De elektroniske ordboksmanuskriptene og det elektroniske Nyordsarkivet har siden blitt vedlikehold og utvidet. Men det store potensialet den elektroniske lagringsformen ga, ble i svært liten grad utnyttet. Bruken av datamaskiner ved SLM utviklet seg i løpet av 1980-årene likevel ikke særlig lenger enn til en erstatning for skrivemaskinene i ordboksredaksjonene.

DOKUMENTASJONSPROSJEKTET

De første planene for Dokumentasjonsprosjektet ble lansert i 1989, og tanken var å gjøre et krafttak for å gi alle samlingsavdelingene ved HF-fakultetet ved Universitetet i Oslo anledning til å ta i bruk datastøttede metoder. De store papirarkivene hadde på sett og vis blitt sin egen fiende idet enhver form for systematisk reorganisering av materialet var blitt så kostbart at det i seg selv krevde ekstra bevilgninger. Dette gjør de manuelle arkivene svært lite anvendelige til forskning. Se Ore 1991 for en nærmere beskrivelse av samlingsavdelingene ved Universitetet i Oslo

Hensikten med Dokumentasjonsprosjektet var å legge et moderne datateknisk grunnlag for samlingsavdelingenes behandling av informasjon og effektivisering av det interne samlingsarbeidet, og å bedre eksternt samarbeid og utveksling av informasjon samt innhenting av ny informasjon. På denne måten ønsket man både å utløse et forskningspotensiale, men også å avdekke svakheter i rutiner og systemer og å forbedre disse. Det var også et viktig mål å tilgjengeliggjøre informasjonen for andre forskere, for studenter, for undervisning, for offentlig forvaltning og for allmennheten så langt det er forsvarlig ut fra personvern, sikkerhet, opphavsrettigheter og eventuelle kommersielle hensyn, se Ore 1994 og Ore 1998 for en

nærmere beskrivelse av metoder og løsninger i Dokumentasjonsprosjektet.

Dokumentasjonsprosjektet ble i 1991-1992 utvidet til å omfatte alle de fire universitetene i Norge. Prosjektet kan grovt deles inn i en museumsdel og en språklig orientert del. I den museumsrettede delen av prosjektet arbeider vi med å bygge opp forskningsdatabaser for de arkeologiske museene i Bergen, Oslo og Tromsø. I Tromsø arbeider vi også med nyere kulturhistorisk materiale. Den språklige delen består i hovedsak av tilrettelegging av bakgrunnsmateriale for ordboksavdelingene (bokmål, gammelnorsk og nynorsk) ved SLM i Oslo og for Trønderordboka ved Norges teknisk-naturvitenskapelige universitet (NTNU) og for navnegranskingsmiljøene i Oslo og Tromsø.

Om lag halvparten av ressursene i Dokumentasjonsprosjektet har blitt brukt på å lage en elektronisk erstatning for seddelarkivene ved SLM. Det var tre delprosjekter innen leksikografi, ett for bokmål, ett for gammelnorsk og ett for nynorsk. De seks årene delprosjektene varte, har gitt oss mye erfaring: De teknologiske forutsetningene i verden rundt oss har også utviklet seg voldsomt i disse årene. Det som kunne virke utopisk i 1990, er i dag hverdagslige realiteter. Vi er alle knyttet til Internett, der vi kan hente informasjon om de fleste emner. Det foregår en stadig økende elektronisk publisering av ordbøker og tekstarkiv både via nettet og via CD-ROM.

DOKUMENTASJONSPROSJEKTETS SPRÅKDATABASER

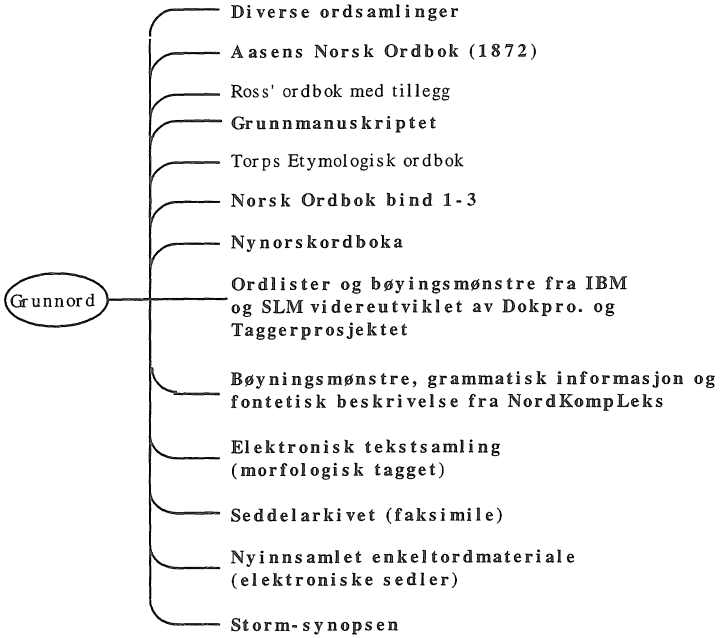
Det originale målet for Dokumentasjonsprosjektets arbeid ved SLM var "å gjøre seddelsamlingene tilgjengelig på elektronisk form". Dette noe upresise målet ble etter en del tids arbeid delt opp i en rekke delmål. Det viste seg raskt at det ikke var hensiktsmessig å skrive av hele seddelmassen og lagre alt sammen i en database.

Om lag halvparten av alle sedlene er enten skrevet av og lagt inn i tekstlige databaser eller konvertert til faksimiledatabaser (se senere i artikkelen). Dette kan vi kalle leksikografske grunnlagsdatabaser.

En stor del av ordbokssedlene som er basert på litterære kilder, er det ikke gjort noe med. I stedet er det laget elektroniske tekstsamlinger. En lang rekke eldre ordbøker og ordboksmanuskript er konvertert til

elektronisk form. Dette danner en tekstsamling på om lag 100000 tekstsider, hvorav det meste er eldre litteratur. Fra et leksikografisk synspunkt må vel dette sise å være uberabeidede rådata.

Kjernen i språkdatabasene er de leksikalske databasene for bokmål og nynorsk. Med utgangspunkt i håndordbøkene Bokmålsordboka og Nynorskordboka og materiale fra IBM er det laget to moderne leksikalske databaser. Basene vil senere bli utvidet med materiale fra NorKompleks-prosjektet ved NTNU. De leksikalske databasene er laget for å kunne brukes i ulike språkteknologiske formål (se senere i artikkelen).



Figur 1: *En tematisk skisse av nynorskdatabasen. De deler som er digitalisert eller planlagt digitalisert gjennom Dokumentasjons-prosjektet eller andre, er satt i halvfet.*

DATABASER OVER ORDBØKER

For bokmål finnes bare Bokmålordboka tilgjenglig i elektronisk form. Det har vært gjort visse forsøk på å få inkludere Norsk Riksmålsordbok, men dette har strandet av forskjellige grunner. For nynorsk finnes det derimot et stort utvalg av elektroniske ordbøker og ordlister. Vi kan blant annet nevne Aasens Norsk Ordbog (Aasen 1872), Aasens Norsk Målbunad, Nynorskordboka (Hovdenak et al. 1994) og Grunnmanuskriptet (Skard et al. 1932), som er et første utkast til Norsk Ordbok fra 1930-tallet. I tillegg finner vi 20-30 ordsamlinger fra 1600-, 1700- og 1800-tallet. Det er også planer om å gjøre Ross' ordbok med tillegg tilgjenglig.

For å få full nytte av den nynorske ordbokssamlingen må det foretas en felles normering av oppslagsord for de ulike ordsmalingene og ordbøkene. Dette krever norsk-filologisk og helst leksikografisk ekspertise. Sammenkoblingsarbeidet krever altså at alle oppslagsordene i de ulike samlingene og verkene også får påført en variant som følger den moderne rettskrivningen eller eventuelt 1938-rettskrivningen. Fagkonsulentene i Dokumentasjonsprosjektet har begynt på dette standardiseringsarbeidet både når det gjelder ordsamlingene og nynordskarkivet.

I databasen over nynorskarkivet beholder vi det originale oppslagsordet. I tillegg har vi introduserer et nytt felt som inneholder den normaliserte varianten. På denne måten vil det alltid være mulig å finne sedler i databasen ved å søke etter det oppslagsordet som faktisk står på den originale papirseddelen. I de gamle ordlistene legger vi de nye opplysningene inn som SGML-koder i den elektroniske fullteksten, se figur 2. Dette svarer i og for seg til den tradisjonelle ekserperingen, men har den fordel at den orignale teksten ikke blir "klippet opp" i fragmenter. Men dersom man ønsker tradisjonelle sedler enten i elektronisk form eller på papir, kan disse produseres maskinelt på grunnlag av den kodede fullteksten.

<ORDF NORM="kljåstein" KL="m"
GRM="pl(dsk)"><HFEIT>Kliaasteene</HFEIT></ORDF> kaldis de Steene / som
er fest paa
Grunden / af en Nod / at det skal snart siuncke for Fisken /
ellers brugis samme Steene til <ORDF NORM="oppstodevev"
KL="m"></ORDF>opstaa Væff / neden ved Gulffuit
paa Varpen / <KURS>sive</KURS> <ORDTILV>Renne Garnit</ORDTILV> /
at det er stifft oc stragt /
at igiennem drage i Sletten.
</ARTIKKEL>

Figur 2: *Et utsnitt av den elektroniske versjonen av Christen Jensøns 'Den Norske Dictionarium' fra 1646 med to ekserperinger i den egentlige artikkelen for 'Kliaasteene'.*

Med dagens utgivelsesstrategi for Norsk Ordbok vil nynorskdatabasen de facto være den eneste samordnede presentasjon av grunnlagsmaterialet fra A til Å i de neste 50 årene. Det er derfor et spørsmål om ikke fagleksikografene burde delta sterkere i dette arbeidet. Så lenge denne normaliseringen ikke er foretatt, må brukeren bruke sin egen kreative språksans for å formulere søk som får tak i oppnåelig informasjon.

LEKSIKALSKE DATABASER

Med en leksikalske database menes en database som inneholder informasjon om ords oppbygning og bøyning (morfologi), grammatisk funksjon, mening, relativ frekvens og så videre. En leksikalsk database er altså datateknikkens svar på en ordbok. Den skiller seg fra en tradisjonell ordbok også ved at det ikke er meningen at all informasjonen skal leses av mennesker. I mange tilfeller kan leksikalske databaser inneholde informasjon som er kodet til bruk i ulike språkteknologiske verktøy, så som morfologiske analysatorer, syntaksanalysatorer og oversettelsesstøttesystemer. Men en leksikalsk database vil typisk kunne inneholde teksten fra en eller flere

tradisjonelle ordbøker som hjelp og supplement til menneskene som bruker basen.

Ordbøker og ordlister er skrevet for å brukes av mennesker og er ofteganske inkonsistente i måten opplysningene er oppført på. Dette skyldes dels at ordboksforfatterne ikke har hatt mulighet eller tid til å kontrollere manuskriptet samlet, dels at mangelen på konsistens ikke spiller noen særlig rolle for den jevne bruker. I produksjonen av trykte ordbøker er hensynet til plass viktig. Dette medfører at ekstra informasjon om for eksempel bøyning og grammatisk funksjon ofte er helt utelatt eller bare ført opp der det i prinsippet ikke er mulig å finne informasjonen på annen måte. Bøyningsinformasjon er for eksempel utelatt ved de aller fleste sammensetninger i både i Bokmålsordboka og i Nynorskordboka. Dette har lite å si i vanlig bruk, men gjorde arbeidet med å lage de leksikalske databaser noe plundrete.

En leksikalsk database trenger ikke inneholde flere opplysninger enn en vanlig trykt ordbok, men et vesentlig poeng ved å lage leksikalske databaser er å lage grunnlagsmateriale som også kan brukes i datamaskinelle redskaper for språkanalyse og i andre språkteknologiske anvendelser som stavekontroll og talegeneratorer. Språkteknologiske systemer kan deles i normative og deskriptive/analyserende. De deskriptive/analyserende verktøyene krever en størst mulig dekningsgrad. Ordbaser som skal brukes i slike verktøy, bør derfor helst inneholde flest mulig ord og bøyningsformer, også slike som ikke er i henhold til gjeldende rettskrivning, men som er i hyppig bruk. Ordbaser for normative verktøy som stavekontroller må på sin side være begrenset til ord og bøyningsformer som er i henhold til gjeldende rettskrivning. For begge typer anvendelser gjelder det at ordbasene må inneholde fullstendige og konsistente opplysninger for hvert enkelt oppslagsord.

grunnord	artikkelnr	ordbok	ordbokskode	ibm-kode	bøyningskode
...	...	...	...	...	...
lage	39397	b	V01,V03	ibm-kode	001,010,011,030
...	...	...	...	...	...

001	010	011	030	opplysninger brukt i taggerprosjekt	linjenr
e	e	e	e	verb inf <trans1>	01
er	er	er	er	verb pres <trans1>	02
es	es	es	es	verb inf pres pass <trans1>	03
a	et	et	de	verb pret <trans1>	04
a	et	et	d	verb perf-part <trans1>	05
a	et	et	d	adj <perf-part> nøyt ub ent <trans1>	06
a	et	et	d	adj <perf-part> mask fem ub ent <trans1>	07
a	ete	ede	de	adj <perf-part> mask fem nøyt be ent <trans1>	08
a	ete	ede	de	adj <perf-part> ub be fl <trans1>	09
end	ende	ende	ende	adj <pres-part> mask fem nøyt ub be ent fl	10
e				<trans1>	
-	-	-	-	verb imp <trans1>	11

Figur 3: *Et forenklet utsnitt av bokmålsdatabasen med bøyningsinformasjon, øverst grunnordtabellen, nederst noen verbparadigmer.*

De leksikalske databasene som er laget i forbindelse med Dokumentasjonsprosjektet, er til nytte for både normative og deskriptivebeskrivelser. Kjernen i de leksikalske databasene for bokmål og nynorsk er laget på grunnlag av materiale fra IBM (se f.eks. Engh 1991), Bokmålsordboka og Nynorskordboka, samt materiale laget i prosjektet NorKompLeks ved Lingvistisk institutt, NTNU (se Nordgaard 1995). Fra IBM har vi overtatt rettigheter til fritt å bruke grunnordlister og sett med bøyningsmønstre bokmål og nynorsk. For hvert ord i grunnordlistene er det markert et nummer som angir det tilhørende bøyningsmønstret.

I forbindelse med Taggerprosjektet (Johannessen 1995) har Dokumentasjonsprosjekt og Tekstlaboratoriet tatt utgangspunkt i IBM-materialet, utvidet grunnordslistene med oppslagsordene i Bokmålsordboka og Nynorskordboka samt foretatt en ombygging av systemet med bøyningsmønstre slik at hvert mønster ikke har alternativer, og slik at sideformer er tatt med. Hvert ord er koblet til den eventuelt tilhørende artikkelen i Bokmålsordboka eller Nynorskordboka. Ved hjelp av bøyningsmønstrene er det mulig å generere alle mulig bøyningsformer av et ord. Denne genereringsprosessen kan selvfølgelig snus, slik at det ut fra en bøyd ordform er mulig å finne hvilket eller hvilke grunnord ordet kan ha. Denne teknikken brukes allerede i den såkalte multitaggeren, som inngår som en del av taggerprosjektet. Den kan fritt prøves av interesserte via Dokumentasjonsprosjektets hjemmesider. Taggerprosjektet ved Tekstlaboratoriet arbeider med teknikker for å gjøre analyseresultatet entydig, for eksempel å kunne avgjøre om 'baker' kommer av verbet 'å bake', substantivet 'en baker' eller er flertall av legemsdelen 'en bak'. En analysator som klarer å skille mellom slike former, vil være svært nyttig i arbeidet med å analysere løpende tekster med henblikk på å finne interessante forekomster av ord og fraser. De mange rettskrivningsendringene i norsk gjør det imidlertid komplisert å få en slik tagger til å virke tilfredstillende på eldre sakprosa og skjønnlitteratur generelt. Et annet problem er blanding av dialekter og standardspråk. Nynorske tekster inneholder mange dialektvendinger, men også svært mange ord og uttrykk som helst klassifiseres som bokmål. Jeg prøvde i 1994 å "vaske" romanen "Hildegunn" av Haldis Moren Vesaas med en tidligere utgave av multitaggeren for nynorsk. I romanen viste det seg å være stor mengde bokmålsord som taggeren ikke ville vedkjenne seg.

I NordKompLeks-prosjektet har det også blitt laget bøyningsmønstre som dekker oppslagsordene i Bokmålsordboka og Nynorskordboka. I tillegg har hvert ord fått uttaleinformasjon og grammatisk informasjon. Dette vil også bli lagt inn i databasene.

Databasene er allerede i bruk i Taggerprosjektet og vil senere i år bli tatt i bruk av Norsk språkråd som en ordregistrant for norsk. Da vil databasene bli utvidet med felter som indikerer om ordet og en eventuell form av det er i henhold til gjeldende rettskrivning. Siden databasene

allerede bygger på ordbøkene og på IBM materialet (ajourført med læreboknormalen frem til om lag 1987), vil ikke dette være uoverkommelig. Man vil dermed oppnå å ha en ajourført oversikt over den nye rettskrivningen som kommer i år 2000. I første omgang vil en slik database lette arbeidet til Språkrådet og til SML i forbindelse med ugivelsen av nye utgaver av Bokmålsordboka og Nynorskordboka. De ajourførte databasene vil også være nyttige for andre ordboksutgivere, og ikke minst for produsenter av stavekontrollssytemer.

DOKUMENTASJONSPROSJEKTET OG DE LEKSIKOGRAFISKE SEDDELSAMLINGENE

Det er til sammen 8 - 9 millioner sedler i samlingene ved SML. En typisk ordseddel inneholder et ord i grunnform, en liten tekstbit som gir et eksempel på en bruk av ordet, og opplysninger om grammatikk, uttale og eventuelle andre forhold rundt dette eksempelet. I de siste 130 årene har ordsedler vært den tradisjonelle systematiske metoden for å samle belegg og opplysninger om bruk av ord for redigeringen av ordbøker. Den opprinnelige målsetningen for delprosjektene ved SLM var som nevnt tidligere, "å gjøre seddelsamlingene tilgjengelig på elektronisk form".

I de siste tiårene har imidlertid datateknikken muliggjort en mye mer effektiv oppbygning av den informasjonen som en seddelsamling representerer. Optisk lesning av tekst (eng. Optical Character Recognition, OCR) og datastøttede hjelpemidler for (halv-)automatisk markering av grammatisk informasjon til ord i løpende tekst brukes nå til å bygge opp elektroniske tekstsamlinger. Konkordansprogrammer og ulike søke- og analyseprogrammer for tekst brukes til å hente ut informasjon fra disse tekstsamlingene. På denne bakgrunn var det derfor riktig å vurdere om det var hensiktsmessig å skrive av alle sedlene og legge dem inn i en database. Etter en mer inngående analyse av seddelmaterialet ble det besluttet bare å konvertere nyordssamlingen til Bokmålsavdelingen, seddelsamlingen til Gammelnorskavdelingen og hele nynorsksamlingen (Seddelsamlingen til Norsk Ordbok), i alt 4 millioner sedler. De resterende seddelsamlingene skulle erstattes av elektroniske tekster.

I den leksikografiske seddelsamlingen er det for et utvalg ord fra et utvalg tekstbrokker (1-20 linjer) gitt opplysninger om bøyningsformen, rotform og annet. Et tagget tekstkorpus er derimot en samling større tekstfragmenter (fra 25 sider løpende tekst og oppover) der hver ordform i tekst(fragment)ene har fått markert grunnord, ordklasse og aktuell bøyningsform. Et seddelarkiv kan delvis sammenlignes med et tagget tekstkorpus. I et seddelarkiv kan man imidlertid bare søke etter enkeltord med eventuell morfologisk informasjon, mens man i et tagget tekstkorpus kan søke etter fraser og flere ord med tilhørende morfologisk informasjon.

Gammelnorskavdelingens seddelsamling er basert på hele diplomtekster og sagaer. Det er stort sett laget en seddel for hvert eneste ord i de tekstene som ligger til grunn. Seddelsamlingene viste seg å være så systematisk bygd opp at vi kunne bruke dem til å gjenoppbygge de originale sagaene som elektroniske tekster der hver ordform har fått markert grunnord, ordklasse og aktuell bøyningsform. I figur 4 er det øverst vist en typisk seddel. Nedenfor står det samme tekstfragmentet, men der er informasjonen fra sedlene ført inn i teksten. Den originale teksten kan gjenfinnes ved å lese ordene som står rett foran '</F>' til høyre på linjene. Til venstre står ordene i normalisert grunnform. Den taggede eller kodede teksten svarer til hva den morfologiske taggeren som lages i Taggerprosjektet, vil kunne levere for en moderne norsk tekst.

N 30 konungr m
konungenum bds

(1) Erkibyskupenum flytianda sem fyrr sagðez þætta mal við konungenn. (2) en konungenum væitanda. færr thomas ærkidiakn til konungsens ok (3) gerez hans kanceler.

Thom. 7

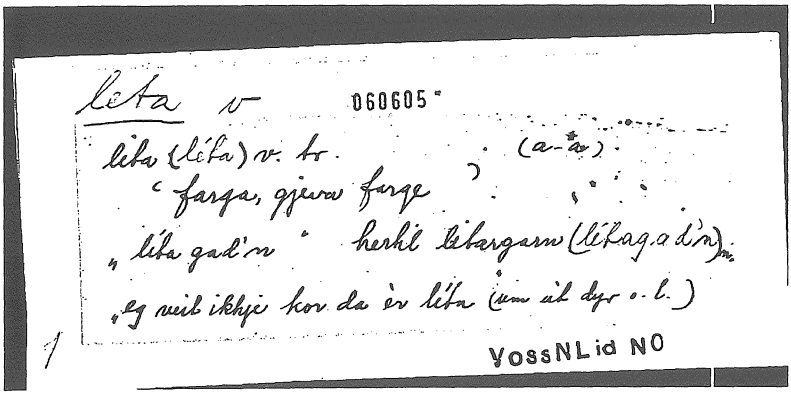
(1) <F O="erkibiskup" K="m" G="bds">Erki
<F O="biskup" K="m" G="bds">byskupenum</F></F>
<F O="flytia" K="v" G="ppt dsn">flytianda</F>
<F O="sem" K="konj" G="med i">sem</F>
<F O="fyrr" K="adv" G="komp">fyrr</F>
<F O="segia" K="v" G="3sii refl">sagðez</F>
<F O="siá" K="pron" G="asn">þætta</F>
<F O="mál" K="n" G="as">mal</F>
<F O="viðr" K="prep" G="med a">við</F>
<F O="konungr" K="m" G="bas">konungenn</F>. (2)
<F O="en" K="konj" G="">en</F>
<F O="konungr" K="m" G="bds">konungenum</F>
<F O="veita" K="v" G="ppt dsn">væitanda</F>.
<F O="fara" K="v" G="3spi">færr</F>
<F O="Thomas" K="m pr" G="ns">thomas</F>
<F O="erkidiákn" K="m" G="ns">ærki
<F O="diákn" K="m" G="ns">diakn</F></F>
<F O="til" K="prep" G="med g">til</F>
<F O="konungr" K="m" G="bgs">konungsens</F>
<F O="ok" K="konj" G="">ok</F> (3)
<F O="gera" K="v" G="3spi refl">gerez</F>
<F O="hann" K="pron" G="gs">hans</F> kanceler.

Figur 4: *Øverst original seddel, nederst den samme teksten etter at informasjonen for alle sedlene i seddelarkivet er skrevet inn.*

I Bokmålsavdelingen er den alt overveiende delen av ordsedlene laget på grunnlag av hele forfatterskap (Wergeland, Bjørnson, m.m.). På disse sedlene står det en liten tekstbit med det interessante ordet understreket, grunnformen av ordet, ordklassen og opplysning om hvor tekstbiten er hentet fra. Ser vi bort fra ordklassen og grunnformen, er dette akkurat den informasjonen vi kan få ut av en såkalt KWIC-konkordans (Key Word In Context). Slike konkordanser kan lages maskinelt ut fra en elektronisk tekst. En elektronisk tekst åpner også for mange interessante anvendelser. Vi valgte derfor å bygge opp en elektronisk tekstsamling som en erstatning for 90-95 % av sedlene i bokmålssamlingen (i alt 3-4 millioner). Tekstene er blitt kodet i overensstemmelse med anbefalingene til det SGML-baserte "Text Encoding Initiative" (Goldfarb 1991, Sperberg-McQueen and Burnard 1994), se siste avsnitt i artikkelen. Den eneste seddelsamlingen som er skrevet inn, er det såkalte nyordsmaterialet (200 000 sedler) for perioden før 1976. Dette materialet supplerer det allerede eksisterende elektroniske nyordsmaterialet for perioden etter 1976 (om lag 300 000 "sedler").

Seddelsamlingen til Norsk Ordbok skiller seg ut fra de to andre ved at det er bygd opp over en lang periode (om lag 60 år) av mange hundre frivillige uten spesiell leksikografisk utdannelse. Dette har resultert i en heterogen samling som består både av enkle sedler som for bokmål, og av sedler med mye ekstra informasjon om blant annet bruk og uttale. Den opprinnelige planen for konverteringen var at de enkle sedlene skulle frasorteres og erstattes av elektroniske tekster. Resten av sedlene skulle skrives inn og SGML-kodes slik det er vist i figur 3. Etter at om lag 10 % av sedlene var behandlet, viste det seg at det i praksis var vanskelig å sortere samlingen på en effektiv måte. Det til dels uryddige oppsettet av informasjonen på sedlene samt mye håndskrift gjorde at konverteringsarbeidet gikk for langsomt. Det ble også klart at dersom man skulle kunne stole på den elektroniske versjonen, måtte en faksimile være tilgjengelig i databasen. Vi valgte derfor å gå bort fra sortering og avskrift av sedlene. I stedet har vi laget en faksimiledatabase over samtlige 3 millioner nynorsksedler. Denne samlingen av faksimiler har oppslagsord, ordklasse og uttømmende

kildeopplysninger som søkenøkler. Man mister på denne måten mulighetene til å søke i den løpende teksten på sedlene, men har fremdeles muligheten til å finne sedler etter grunnord, ordklasse, sted i landet, kildetype og hvem som har skrevet seddelen. Se Ore 1996 for en inngående diskusjon av fordeler og ulemper ved denne løsningen.

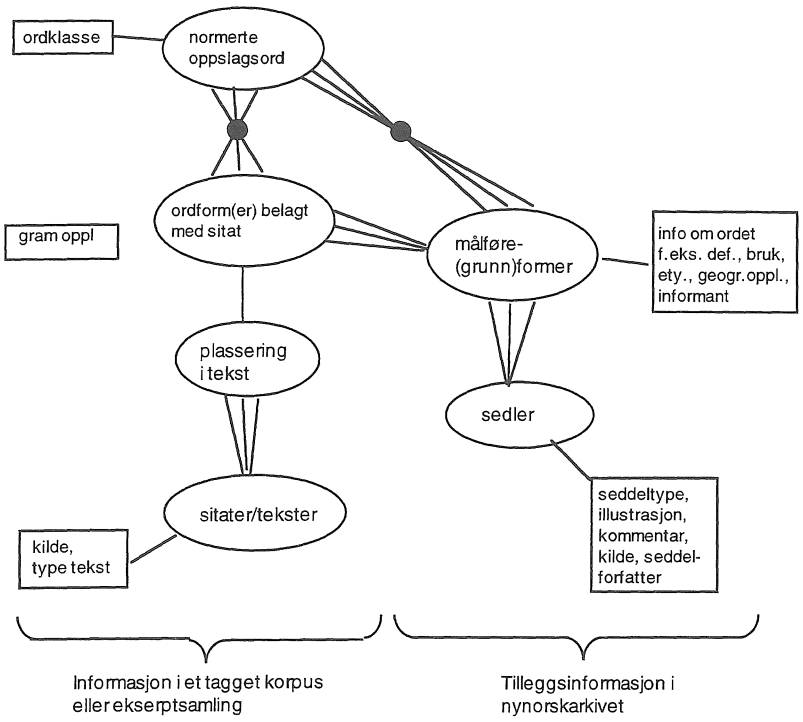


Figur 5: *Faksimile av nynorskseddel.*

```
<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 6: *Kodet versjon av seddelen vist i figur 5.*

Denne korte gjennomgangen av arbeidet med de ulike seddelsamlingene viser mangfoldet i seddelmassen. Da vi startet arbeidet, hadde vi nok en noe idealisert forestilling om "den leksikografiske seddel". Gjennom arbeidet med samlingene har vi fått en svært nøktern holdning til innholdet i seddelarkivene. Dokumentasjonsprosjektets medarbeidere har også opparbeidet en usedvanlig kompetanse på koding og lagring av leksikografisk grunnmateriale. De store seddelarkivene hører på mange måter fortiden til. Men informasjonen i dem vil være viktig for fremtidens leksikografer. Men det vil også i fremtiden være ønskelig å motta informasjon fra informanter som snubler over interessante ord og vendinger. Det vil også være ønskelig å drive selektiv ekserpering fra ulike tekster. Leksikografene selv vil derfor fortsatt ha behov for å lage egne "sedler" og å legge inn informasjon som kommer på tradisjonelle papirsedler. I den ferdige databasen vil det derfor være mulig å legge inn denne typen materiale. Dette vil kunne gjøres enten ved at det skrives SGML-kodede sedler eller "inline"-markering som vist i henholdsvis figur 5 og 4, ved bruk av et spesiallaget ekserperingsverktøy, slik det ble gjort i gammelnorskprosjektet, eller mer systematiske masselemmatiseringverktøy, slik som det er blitt gjort i Canterbury Tales-prosjektet (Robinson 1996).



Figur 7: *Skisse av et kombinert seddelarkiv og tagget tekstkorpus.*

Selve seddelbasen følger det noe skjematiske oppsettet i figur 7. Den venstre delen i figuren beskriver en enkel seddelbase, for eksempel gammelnorskdatabasen eller de litterære sedlene i bokmålsavdelingen. Den høyre delen beskriver hva som kommer i tillegg i sedlene fra nynorskarkivet. Her vil det kunne være en eller flere målføreformer koblet til oppslagsordet som i seddelen i figur 5 og 6. I tillegg vil en kunne ha informasjon om bruk og informant. I nedre høyre hjørne finner vi seddelfaksimilene sammen med eventuell annen tilleggsinformasjon.

TEKSTSAMLINGENE

Det ble allerede under arbeidet med forprosjektet til Dokumentasjonsprosjekt klart at det ikke var særlig mye å vinne ved å legge sedler basert på litterære tekster inn i databasene. De inneholder ikke informasjon ut over det man relativt enkelt kan hente ut av elektronisk tekst ved hjelp datamaskinelle metoder. De litterære sedlene i nynorsksamlingen har likevel blitt lagt inn i databasen siden det var for tidkrevende å sortere dem ut, og fordi antallet forskjellige kilder var så enormt (over 6000). Vi valgte likevel å lese optisk (OCR) ca 10000 sider eldre nynorsklitteratur.

Bokmålsedlene er stort sett basert på hele forfatterskap. Det var derfor mulig å lese optisk grunnlagstekstene. Opphavsrettigheter har imidlertid vridd utvalget i retning av eldre litteratur. Alt i alt består tekstsamlingen for bokmål av 60000 sider tekst hentet fra 1500-tallet til første halvpart av dette århundret. Tekstene er nøye SGML-kodet med hensyn på struktur og genre. De egner seg også godt for datamaskinelt støttede litterære studier og som basis i kulturhistoriske publikasjoner.

Gammelnorsksedlene er alle basert på diplomer eller litterære tekster. Men her gjorde en rasjonell innskrivning at vi fikk rekonstruert de originale tekstene. Som et supplement til seddelarkivet konverterte prosjektet det trykte Diplomatarium Norvegicum samt nyere diplomavskrifter gjort i perioden fra 60-tallet og oppover til 80-tallet. Disse tekstene er kanskje vel så interessante for historikere som for språkforskere.

Dokumentasjonsprosjektet har også konvertert en stor mengde andre tekster til elektronisk form. Dette omfatter arkeologiske rapporter, folkeviser (3400 oppskrifter) og en stor mengde selvbiografier eller såkalte minneoppgaver (64000 sider) skrevet av ulike personer på 1960-1990-tallet. Det er å anta at særlig minneoppgaven vil kunne være interesante som språklig grunnmateriale siden disse er skrevet av "alminnelige mennesker" og ikke har gått gjennom noen form for språkvask.

VIDERE ARBEID VED SEKSJON FOR LEKSIKOGRAFI OG MÅLFØRE

Dokumentasjonsprosjektet er nå avsluttet. De store samlingene ved Seksjon for leksikografi og målføre er dataførte. Det er fremdeles et stort behov for dataføring av Målførearkivets materiale. Men det er neppe sannsynlig at det i de nærmeste årene vil komme ressurser av den størrelsen vi hadde i Dokumentasjonsprosjektet. Det er nå viktig å forsette utviklingen av gode verktøy for å utnyttede og vedlikeholde de elektroniske datasamlingene. Det er også viktig å utarbeide effektive og databaserte rutiner ved SLM for nyinnsamling av data slik at det ikke skal bli behov for noe nytt Dokumentasjonsprosjekt.

Deler av Dokumentasjonsprosjektets stab er videreført i en liten etterorganisajon. Seksjonen selv har ansatt en egen samlingsansvarlig. Etterorganisasjonen har som delmål å videreutvikle databasene og metoder for SLM. Øverst på oppgavelisten står ordboksredigeringsverktøy, vedlikeholdsgrensesnitt mot de leksikalske databasene og opplegg og rutiner og metoder for nyordsinnsamling og annen tilvekst. Oppgavene er mange og interessante, men jeg tror det vil kreves mye arbeid med fornyelse og tilpassing for at SLM også i årene som kommer, skal kunne hevde seg som et leksikografisk tyngdepunkt.

LITTERATUR

Atkins, S. *The Hector Project*, i "Proceedings of Complex '92", Budapest 1992

Engh, J. *Lekikografi i IBM Norge.* I: Fjeld, R.V. (red): "Nordiske studier i leksikografi". Oslo 1992

Goldfarb, C. *The SGML Handbook*, Oxford University press 1991

Johannessen J.B. *Morfologiske taggere for norsk*, Prosjektbeskrivelse Oslo 1995

Nordgaard T. *NordKompLeks – et norsk komputasjonelt leksikon.* Prosjektbeskrivelse, Trondheim 1995.

Ore, C.-E. *Dokumentasjonsprosjektet ved Det historisk-filosofiske fakultet, Universitet i Oslo*, i Fjeld, R.V. (red): "Nordiske studier i leksikografi" Oslo 1992

Ore, C.-E. *Making an information system for the Humanities*, Computers and the Humanities, Journal of the ACH, 1994

Ore, C.-E. *Korpus og seddelarkiv, fredelig sameksistens mellom det beste og det gode?*I: Svavarsdottir, Kvaran, Jónsson (red): "Nordiske studier i leksikografi 3", Reykjavik 1996

Ore, C.-E. *Hvordan lage databaser for språk og kulturfag,* i Aukrust A. og Hodne B. (red.) *Fra skuff til skjerm,* Universitetsforlaget, Oslo 1998

Robinson P. (ed) *The Wife of Bath's Prologue* Cambridge University Press, Cambridge 1996

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

"Elektroniske hjelpemidler - leksikografisk fornying"

JANNE BONDI JOHANNESSEN, TEKSTLABORATORIET

1. INNLEDNING[1]

I dette foredraget vil jeg ta for meg to sider ved leksikografien. Den ene er den leksikografiske arbeidsprosessen - særlig arbeidet med å lage definisjoner til ordbøkene. Den andre er nyordsinnsamling. Jeg tror at begge deler vil kunne fornyes ved hjelp av elektroniske hjelpemidler, og gjerne ved nærmere samarbeid med Tekstlaboratoriet, som jo er nærmeste nabo til Seksjon for leksikografi og målføre. La meg straks legge til at jeg vet at elektroniske hjelpemidler ikke er ukjente for leksikografene. I årsskriftet fra seksjonen var det i hvert fall to leksikografer som presenterte arbeid som bygde på elektroniske hjelpemidler. Den ene var Anne Engø, som hadde en artikkel om karakteriserende personnavn i nyordsmaterialet, og den andre var Laurits Killingbergtrø, som skrev om bokstavfrekvens og bokstavkombinasjoner i tekster, ordbøker og kryssord. Jeg skal her konsentrere meg om hvordan man kan bruke elektroniske hjelpemidler til mer tradisjonelt leksikografisk arbeid.

2. DEN LEKSIKOGRAFISKE ARBEIDSPROSESSEN

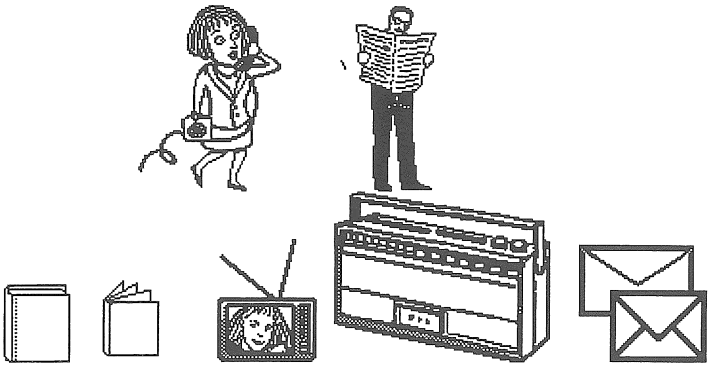
En viktig del av det leksikografiske arbeidet er å skrive definisjoner. Hva betyr egentlig et ord som *lun*, og hva er forskjellen på *tygge* og *knaske*? Når leksikografen spør seg om dette, konsulterer han eller hun gjerne seg selv, sine kolleger eller noen kyndige informanter i felten. Han eller hun har gjerne noen tekstutdrag også, som er funnet ved nitid lesing av aviser og tidsskrifter. De tradisjonelle leksikografiske kildene er illustrert i (1).

---

[1] Jeg vil gjerne takke Kristin Hagen og Anders Nøklestad for nyttige kommentrarer til en versjon av dette foredraget.

(1)

Tradisjonelle leksikografiske kilder



En slik arbeidsprosess har selvsagt lange tradisjoner, men det er ikke selvsagt at det er den eneste eller beste måten å arbeide på. Det synes i hvert fall helt klart at prosessen kan forbedres ved å bruke søking i tekstkorpora.

2.1 Søking i tekstkorpora

La oss begynne med å se på noen definisjoner jeg har funnet i ordbøkene. *Bokmålsordboka* (1994) definerer adjektivet *lun* slik:

(2)

Artikkelen i Bokmålsordboka om ordet 'lun'

lun a1 (sm o s *//logn*) **1** skjermet mot vær og vind, i ly *en l- krok / det l-ere her* / adv: *stedet ligger l-t til* **2** passelig varm, mild *l-e vinder* **3** stillferdig, sindig, behersket; hyggelig *et l-t smil / l- humor* / adv: *snakke*

Hva betyr *lun*? Er lun det samme som *lunken*?

(3) Mulig bruk av ordet *lun*?

a. ? det lune badevannet til babyen

b. ? den lune kokeplata på komfyren
c. ? barnet var lunt - termometeret viste 37°C

Her tror jeg leksikografene kunne hatt god nytte av å bruke et tekstkorpus med mange forekomster - altså et elektronisk korpus. Jeg vil gjerne understreke at det er viktig med store tekstmengder, slik at man virkelig har en sjanse til å se et ords bruk i mange sammenhenger, og slik at det kanskje peker seg ut visse hovedtendenser i bruken. Vi skal se på noen eksempler fra et aviskorpus (Dagbladet, Vårt Land og Arbeiderbladet) som riktignok ikke er så stort (ca 800 000 ord), men størrelsen er valgt fordi den er brukervennlig for dagens formål.

(4) Noen eksempler fra et aviskorpus (Dagbladet, Vårt Land og Arbeiderbladet 1981, fra HD, Bergen)

lun
(13254) og kvinne, og den er holdt | i en **lun**, personlig tone som |
(22585) Finn Hasselknippe | satser på **lun** hjemmehygge og | diskret
(22627) preget av en | hjemlig atmosfære og **lun** hygge. | Samtidig som

lune
(11721) ved siden av | olje, skal sørge for **lune** | stuer og varmt vann i
(6088) tid da familien | samlet seg i den **lune** sirkelen | rundt
(14382) høy, du er ikke pen, men har det **lune** glimtet | i øyet.
(16279) oppbrettede | frakkekraver, **lune** ovnskroker | og ramlende

lunt
(6373) jeg var i det militære og | lå i et **lunt** telt med parafin-ovn, |
(9954) | statsminister Helmut Schmidt | som **lunt** smålpratende titter | opp
(12446) til sengs, kommer | Schluter på **lunt**, dansk vis først | i

Det synes opplagt at alle eksemplene inkluderer både hyggedimensjonen, varmedimensjonen og kanskje t.o.m. skjermingsdimensjonen fra definisjonene i (2). Dette må være grunnen

til at ikke badevannet, kokeplata og barnet kan være lune. Definisjonen i ordboka ville antagelig blitt mer presis om leksikografen hadde brukt et tekstkorpus. Den ville kanskje ikke gitt tre adskilte betydninger av ordet, men heller én som inkluderte alle aspekter.

En sentral oppgave for en ordbok må være å gi svar på hvor grensen går mellom betydningen til de forskjellige leksemer i språket. En slik grensoppgang vil jo ikke minst gjøre definisjonene mer presise også. La oss f.eks. spørre hva som er forskjellen på ordene *bred* og *vid*? Det er ikke lett å finne svaret i ordbøkene.[2]

(5) Presise distinksjoner?

bred/brei:

BM: - med utstrekning til siden, i bredden, t forskj fra smal el. lang
- vidstrakt, som omfatter mye

NN: - som har ei viss vidd ut til sidene målt på tvert av lengde- el. høgderetninga
- som er vid ut til sidene i høve til høgd el. lengd
- som tøyer seg til alle kantar; vid, romsleg

NO: - som har en viss utstrekning mellom sidene
- ikke smal, vid

NIO: - med (stor) utstrekning til alle sider, vidløftig, som omfatter meget,
- med utstrekning til siden

vid:

BM: - med stor bredde el. utstrekning
- omfangsrik

NN: - med stor lengd og breidd
- omfangsrik; omfattande; vag

NO: - bred
- rommelig

NIO: - med stor bredde, utstrekning

[2] BM=Bokmålsordboka (1994), NN=Nynorskordboka (1994), NO=Norsk ordbok (1996), NIO=Norsk illustrert ordbok (1993)

- om det som dekker mange (enkelt-)tilfeller, spenner over et stort område
- SYN. bred, folderik, lang, løs(tsittende), omfangsrik, rom, sekkeaktig, sid, stor, utstrakt

Det kan se ut som om det er minimale forskjeller mellom de to ordene. De brukes jo også til å definere hverandre. Men kan dette være riktig? Er de virkelig synonyme? Er de følgende eksempler mulige?

(6) Mulig bruk av ordene *bred* og *vid*

a. ? en vid elv
b. ? bredåpne øyne
c. ? på bredt gap
d. ? en bred kjole
e. ? et vidt slips
f. ? å være vid over skuldrene

De tentative eksemplene i (6) virker ikke naturlige, og vi venter ikke å finne dem i noen norsk tekst. Da burde definisjonene i ordboka kanskje gitt et hint om forskjellen? Hva er sentrale og hva er mer perifere egenskaper? Å sitte på kontoret og lene seg tilbake i kontorstolen mens man grunner på forskjellen, er nok en lite effektiv fremgangsmåte. Igjen vil et elektronisk tekstkorpus hjelpe en langt på vei:

(7) Romantekster, bokmål (konkordanser hentet fra HD, Bergen):

Vid

EB1	fra I dypet passerte inn mellem dem i **vid** avstand, og efterpå I var
TS1	hår og skjegg vokse og skaffet seg en **vid** hatt. Utpå I vinteren
TS1	gjorde han en aldeles unødvendig I **vid** bue. I I En aften i
TS1	snart står de på en I avsats som er så **vid** og dyp at en hel familie

TS1 er," sier | ravnen høyt og gjør en **vid** bue over skogen for å få |

TS1 kappelignende nattskjorte som lå i en | **vid** pose om skulderen og fikk

AJ1 Og katedralens porter står alltid på **vid** vegg! | Hva! Hahahaha!

AJ1 Hahahaha! Hørte du det, Bobo! Alltid på **vid** vegg! | Rulle litt med

GB1 som fremdeles lå og sutret med **vid** | åpne øyne. Han var det

JHJ1 porten. | Vedkommende var iført en **vid** frakk eller kappe - i

JHJ1 unna. Mannen | er relativt høy, har **vid** frakk med oppslått krave

AB1 120 meter | opp som et landemerke i **vid** omkrets. Andre figur er

JBJ1 nede." | Hun nikket, og Cessnaen tok en **vid** sving nordover. | Thomsen

JBJ1 og betraktet flyet der det steg i en | **vid** sving mot nord. Det kom

KB1 Vinteren gjør byen trang men sko- | gen **vid** . Tankene blir kjøligere.

KB1 men løper bort | til døren, åpner den på **vid** gap, slynger armene rundt

KB1 om morgenen, ligger han på ryggen, med **vid** åpne | øyne, og stirrer

KB1 Han vrenger øynene og åpner munnen på **vid** | gap mens han setter seg

Bred

RA1 ha | det. | Henry dukket inn gjennem en **bred** port. En | varm og beisk

RA1 seige massen. | Trappa endte foran en **bred** dør. | Bak den

HSD1 den vei som andre optråkket god og **bred** . | - Han talte de

HSD1 | nedover høire kinn og vasket en **bred** , buktet vei | nedover i

NG1 kunde vel | være tredve år, **bred** , tettbygget, huden

NG1 det blonde | håret, den andre lignet en **bred** , sterk | bondejente med

EB1 og mørk, | eller hvor den var grund og **bred** med krusninger, hvor |

SC1 maurvei i skogen. Men | demningen var så **bred** at alle fikk plass.

SC1 hans mening var å få meg til å se den så **bred** | og dyp at jeg ville

TS1 å | grave grøft, 50 alen lang, 2 alen **bred** og 3 alen dyp, | per dag.

TS1 En dyp, skinnende kveld kommer de til en **bred** elv. | De stanser i

IH1 ble | fargene bedre og. Munnen var **bred** og kraftig, leppene |

GJ1 rundt dem | som om havet bare var en **bred** elv, og de kom så fort |

GJ1 Så var hun forbi ham. Gaten | var svært **bred** og en trikk kom og

AJ1 | Øynene hans satt dypt under en **bred** , tungsindig panne. | De

BV1 se ut som Johan Ole. For vennen var | **bred** over skuldrene, hadde

BV1 av øyene. | Lenger sør var der en **bred** renne så kystruten kunne

BV1 da han så det. Vest | for nuten skar en **bred** kløft seg ned mot

BV1 | mannfolkarbeid. Og jeg vil bli like **bred** over skuldrene | som du

BV1 | En nokså alminnelig gate. Ganske **bred** og fargerik. | Myldrende

BV1 ser og hører omkring meg. Den er ganske **bred** , men | virker likevel

BV1 hun og pekte på et hode som sto på en | **bred** hylle over sengen. Han

JEB1 | Jan reiste sig og kom frem, spenstig, **bred** og velkledd. | Han hadde

JEB1 stod en mann med ryggen til henne, **bred** | og skrå i skuldrene,
JEB1 hun tenker på en mann til, ung, mørk og **bred** , med | sammenvokste,
JEB1 hurtig, med lange | skritt. Han er **bred** og tung, men gangen er
GB1 | spørsmål det på forhånd hadde vært **bred** tverrpolitisk | enighet
GB1 halvannen meter | lang og ti centimeter **bred** i eregert tilstand. | Når
GB1 gir. . . ," sa Britobert | Ødeskjær. | - **bred** oppslutning, fortsatte
GB1 rødfarge | markerte omrisset med en **bred** kant, og innenfor: den
GB1 Det var klart at det måtte dannes | en **bred** front med et solid
KAL1 seg bakover. En underleppe som var **bred** , og en som var | smal. Et
KAL1 var alt halv ett. | | Bredo er stor. **bred** . Gutta kalte ham
AB1 at han ikke | virker farlig allikevel. **bred** og tett fyller han sin
AB1 først slippe løs et viktig spørsmål til **bred** , demokratisk | diskusjon
AB1 på | den måten." | - Det ble altså en **bred** , demokratisk behandling
JBJ1 | en tretti meter lang og fire meter **bred** sprekk. Trykket er | i
BK1 steder og røper en | dialekt som er like **bred** som Missisippi er lang. |
BK1 tekster: | "FUCK NIXON!" | skrevet med **bred** , rød tusj - | "The system
BK1 prest i 1961. | Veien er rett, **bred** og kjedelig. Store
KB1 å le, | for han arbeider selv med en **bred** tverrpolitisk aksjon |

Etter å ha studert eksemplene, tør jeg si at en sentral egenskap ved adjektivet *vid* må være: Så stor åpning at den ikke kan bli større (vid gap, vid åpne øyne), for stor (om klær). En viktig egenskap ved ordet

bred må være at det brukes om noe som også er langt (bred vei, bred elv, bred gate, bredt slips).

La oss også se på et par verbdefinisjoner:

(8) Presise distinksjoner?

tygge:
BM: - bevege kjevene fra hverandre og sammen flere ganger;
knuse med tennene, presse sammen med gjentatte tyggebevegelser
NN: - knuse, male, elte (mat) med tennene
NO: - bruke tennene på
NIO: - bearbeide, knuse med tennene

knaske:
BM: - tygge hørlig
NN: - tyggje høyrleg
NO: - tygge, spise knask (=godterier)
NIO: - tygge (hørlig)

Betyr disse definisjonene at hørlighet (eller godterier) er eneste kriterium for at man knasker? Hva med det følgende:

(9) Akseptable eksempler på knasking?

a. ? knaske tyggegummi
b. ? knaske bananer
c. ? knaske fikener
d. ? knaske seigmenn

Må ikke det man knasker være sprøtt eller knekkelig, altså ikke bare gi en hørbar lyd, men også en karakteristisk, knusende lyd? Hvis vi igjen går til tekstkorpuset, finner vi i hvert fall noen forekomster som kanskje underbygger mistanken.

(10)
knaske:
a. Bokmålsromaner
BK1 ville fortelle ham noe. | Karl begynte å knaske på eplet igjen. Da den knasket (3)
AJ1 Og jeg så på brystene | hennes og knasket på en nepe og rapte og
AJ1 i røysen og slept ned | til hytten) og knasket mahamzagryn. Den
KB1 suget ned i lungene, et eple som blir | knasket, og allikevel er det

b. Bokmålsaviser:
(14278) | gulrot eller ei skive kålrabi | å knaske på mens de venter |
950622, 18 på Øvregaten før krigen. Vi ungene knasket på frøene, og så ble det

Min mening er ikke å kritisere definisjonene slik de fremkommer i ordbøkene. Siden jeg aldri selv har arbeidet med å finne fram til brukbare definisjoner, har jeg ingen muligheter til å engang forestille meg hvor vanskelig dette arbeidet er. Men mitt mål er å vise at det faktisk går an å gjøre arbeidet enklere ved å bruke elektroniske hjelpemidler. Jeg kan tenke meg to måter som kan øke presisjonen i definisjonene.

Den ene måten å bruke elektroniske hjelpemidler på i leksikografiarbeidet, er altså å benytte et elektronisk tekstkorpus for å lage definisjoner. Det har vi allerede sett eksempler på.

2.2 Aktiv, elektronisk bruk og organisering av ordbøkene

En annen måte er å organisere den ordboka man lager, på en måte som gjør det mulig å foreta søking ikke bare etter oppslagsordet, men også etter ord i andre utvalgte felt av oppslaget, f.eks. definisjonsfeltet, samt ha tilgang til andre ordbøker elektronisk. På den måten kunne man søke etter alle oppslag som har tygging, tenner, tunge, knasking o.a. i enten oppslagsordet eller definisjonen. Slik får man forhåpentligvis samlet ord som brukes nærmest synonymt, og man kan lettere undersøke både hva som forener og hva som skiller de forskjellige ordene.

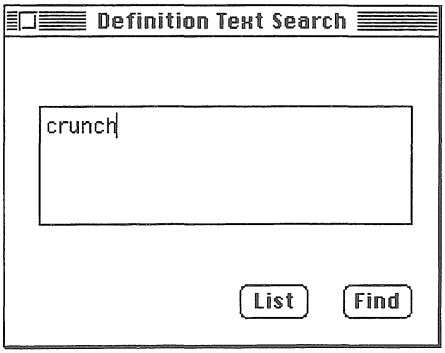
La meg eksemplifisere med den elektroniske utgaven av Oxford English Dictionary:

(11) Oppslagsordet crunch (=knaske) i Oxford English Dictionary CD-ROM

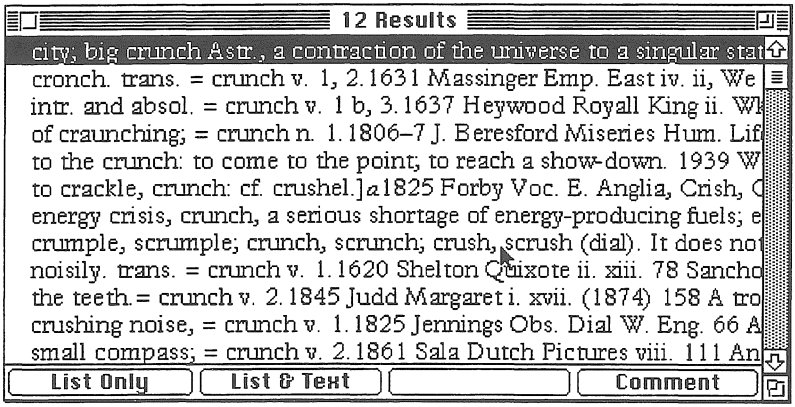
crunch (krʌnʃ), *v.*
[A recent variation of *cranch*, craunch, perhaps intended to express a more subdued and less obtrusive sound, perh. influenced by association with *crush*, *munch*.]
1. *trans.* **To crush with the teeth (a thing somewhat firm and brittle); to chew or bite with a crushing noise.**
1814 *Suppl. Grose's Provinc. Gloss.*, *Crunch*, *Cronch*, and *Cranch*, to crush an apple, etc. in the mouth. *North.*
1832 W. Irving *Alhambra* II. 201 'While I was quietly crunching my crust.'
1859 Kingsley *Misc.* (1860) I. 202 A herd of swine crunching acorns.
b. *intr. or absol.*
1816 Byron *Siege Cor.* xvi, Their white tusks crunch'd o'er the whiter skull.
1856 Kane *Arct. Expl.* II. x. 101 Our appetites were good; and..we crunched away right merrily.
2. *trans.* **To crush or grind under foot, wheels, etc., with the accompanying noise.**
1849 C. Brontë *Shirley* ii. 24 A sound of heavy wheels crunching a stony road.
1873 *Spectator* 23 Aug. 1069/1 You crunch little heaps of salt at every step.
b. *intr. or absol.*
c. *intr. for refl.*
1801 Southey *Thalaba* viii. xxii, No sound but the wild, wild wind, And the snow crunching under his feet!
1880 *Blackw. Mag.* Apr. 452 The animal's hoofs crunch on the stones and gravel

Etter å ha sett på selve oppslagsordet, kan man lete etter andre ord som måtte ha dette verbet i definisjonsfeltet:

(12) Finn ordet *crunch* brukt i definisjonsfeltene for andre oppslagsord



(13) Oppslagsord som har *crunch* i definisjonsfeltet



Herfra kan man klikke seg inn på oppslagsordet som tilhører definisjonen, og finne om det er en sammenheng, eller om det er homonymt, eller hva.

(14) Et oppslagsord som har *crunch* i definisjonsfeltet

scranch *v.*

2. = crunch *v.* **2.**

1845 Judd *Margaret* i. xvii. (1874) 158 A troop of boys and girls..were coming up the hill, goring and scranching the crust [of the snow] with their iron corks.

1853 G. J. Cayley *Las Alforjas* I. 261 [It] broke, being scranched in my pocket, when I fell off pony-back.

Hence

'scranching *vbl. n. and ppl. a.*

1846 W. Sandys [Jan Treenoodle] *Spec. Cornish Dial.* 38 (E.D.D.) Apples ripe for scranching.

1854 A. E. Baker *Northampt. Gloss.* s.v. *Scraunch*, A bow drawn in, an awkward, unskilful manner across a violin makes a scraunching noise.

Slik behøver ikke leksikografen å stole bare på seg selv og sine nærmeste kolleger ved definisjonsformuleringer.

3. NYORDSREGISTRERING

Et annet prosjekt som jeg vet leksikografene har, er å registrere nye ord (og nye bruksmåter av gamle ord) i språket. I dag sitter det folk rundt omkring som har gratis abonnement på et tidsskrift eller en avis, og sender inn funn de mener er nye, til leksikografene i Oslo. Jeg er blitt fortalt at det kan være svært uforutsigelig hva som kommer inn, både

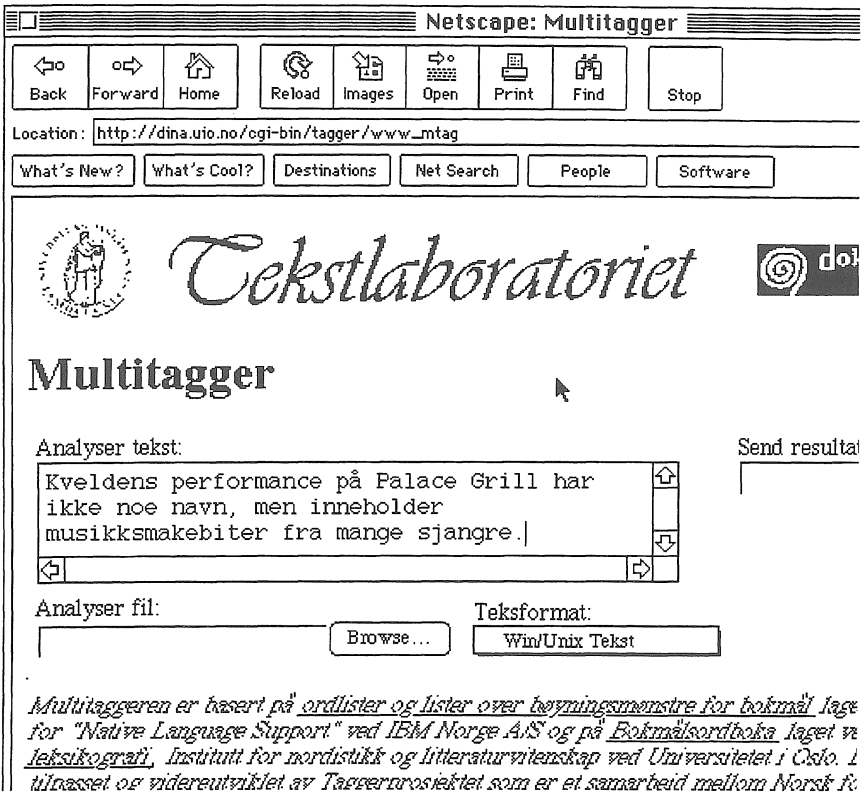
mht kvantitet og hyppighet - og kanskje også kvalitet. (Selv om det ikke er noen grunn til å tvile på at innsamlerne hver for seg er dyktige folk.)

En nyordsregistrering kan gjøres svært mye mer effektiv ved å bruke elektroniske metoder. La oss anta at vi fikk abonnere på elektroniske utgaver av avisene og tidsskriftene. Da ville leksikografene ha tilgang til alt materialet sentralt. Et forholdsvis enkelt datamaskinprogram ville raskt kunne sjekke for hvert ord i en gitt tekst om det var nytt eller ikke - altså om det sto i ordbøkene eller ikke.

For at en maskin skal kunne klare det, er det nødvendig med ordlister som inkluderer alle ordbøkenes oppslagsord, samt deres bøyde former - altså fullformsordslister. Slike fullformsordlister med nødvendig programvare er utarbeidet av Tekstlaboratoriet (spesielt innenfor Taggerprosjektet) og Dokumentasjonsprosjektet i samarbeid. Ordbøkene de bygger på, er først og fremst ordbøkene fra Seksjon for leksikografi og målføre, Bokmålsordboka og Nynorskordboka, men vi har også inkludert IBMs ordlister.

Nedenfor kan man se en måte fullformsordlistene kan brukes på: Man skriver inn en liten tekst, og får den analysert vha fullformsordlistene. Programmet kalles en multitagger, fordi det gir alle lesninger av de kjente ordene. (Alle interesserte kan prøve det selv på denne internett-adressen:
http://dina.uio.no/cgi-bin/tagger/www_mtag.)

(15) Et program for full tagging av tekst:



(16) Full analyse av multitaggeren

"<*kveldens>"
    "kveld" subst mask appell ent be gen
"<performance>"
    "performance" ukjent ord
"<på>"
    "på" prep
"<*palace>"
    "*palace" subst prop
"<*grill>"
    "*grill" subst prop

"<har>"
    "ha" verb pres <trans6> <auxp>
"<ikke>"
    "ikke" adv
"<noe>"
    "noe" adj pos mask fem nøyt be ent
    "noe" adj pos mask fem ub ent
    "noe" adj pos nøyt ub ent
    "noe" adj pos ub be fl
    "noen" det kvant nøyt ent
"<navn>"
    "navn" subst nøyt appell ent ub
    "navn" subst nøyt appell fl ub
    "navne" verb imp <trans1>
"<$,>"
    "$," CLB <KOMMA>
    "$," <KOMMA>
"<men>"
    "men" CLB konj
    "men" konj
    "men" subst nøyt appell ent ub
    "men" subst nøyt appell fl ub
    "mene" verb imp <trans1> <trans2>
"<inneholder>"
    "inneholde" verb pres <trans5>
"<musikksmakebiter>"
    "musikksmakebit" subst mask appell fl ub samset
"<fra>"
    "fra" prep
"<mange>"
    "mange" adj pos ub be fl
"<sjangre>"
    "sjangre" verb inf <refl4> <trans1> samset
    "sjangre" verb imp <refl4> <trans1> samset(NB! Analysert som sammensetning - en måte å re sengen på?)

"<$.>"
"$." CLB <PUNKT>

(Angående *sjangre*: De tillatte formene i ub fl er: *sjangere, [sjangerer], sjangrer*)

Det programmet som er vist her, har som formål å gi alle mulige analyser av hvert eneste ord. Til en nyordsregistrering ønsker man jo strengt tatt bare nye ord. Men vi kan se at den informasjonen et nyordsregistreringsprogram ønsker, finnes her - blant all den andre informasjonen. Ord programmet overhodet ikke klarer å analysere, som *performance,* får merkelappen (taggen) "ukjent ord". Ord som programmet klarer å analysere som sammensetning, gis taggen "samset", samt den grammatiske informasjonen til sammensetningens siste ledd, som i *musikksmakebiter*. Dessuten analyserer programmet hva det tror er egennavn, og gir dem også en tagg. Leksikografen vil altså allerede med den programvaren vi har idag, kunne finne fram til nye ord - altså slike som ikke finnes i ordbøkene. Om han eller hun ønsker det, kan man sile bort f.eks. egennavn allerede før registreringen. Det kan også skilles mellom sammensetninger og ord som ikke kan analyseres som sammensetning.

Jeg vil gjerne demonstrere med et større eksempel hvordan nyordsinnsamlingen kan foregå. Jeg tok en tekstsamling som besto av ca 120 000 ord, fra noen nyere aviser og romaner, og fikk laget lister over akkurat de tre kategoriene vi har snakket om. Her er resultatet:

(17) Ukjente ord, 37 av ca. 120 000. Et utvalg presenteres her:
abroad
af
bosniadina
breath
filius
fresh
h
happy

healthy
i'en
kke
me
mikta
n't
night
oh
paa
peace
perfect
poisonous
relle
sanctus
says
study
the
valium
with
yes

(18) Egennavn: Tilsammen 3985 funn av til sammen ca 120 000 ord.
Her presenteres et utvalg:

Aftenposten
"Aftenposten" subst mask prop
Tyrkia
"Tyrkia" subst prop
Penguin
"Penguin" subst prop
USA
"USA" fork subst prop
Boutros
"Boutros" subst prop
"Boutro" subst prop gen
Boutros-Ghali

"Boutros-Ghali" subst prop
Bill
"Bill" subst prop
Clinton
"Clinton" subst prop
FN
"FN" fork subst prop
Clinton
"Clinton" subst prop
FNs
"FNs" subst prop
"FN" subst prop gen
New
"New" subst prop
York
"York" subst prop
Sovjetblokken
"Sovjetblokken" subst prop

(19) Sammensetninger: 1357 nye sammensetninger av ca. 120 000 ord
Et utvalg, alle grammatiske analyser er inkludert:

Kongsberg-gruppen
"Kongsberg-gruppe" subst mask appell ent be samset
forsvarsmateriell
"forsvarsmateriell" adj pos mask fem ub ent samset
"forsvarsmateriell" subst nøyt appell ent ub samset
"forsvarsmateriell" subst nøyt appell fl ub samset
fem-seks
"fem-seks" det kvant fl samset
norsk-utviklet
"norsk-utvikle" verb pret <trans1> samset
"norsk-utvikle" verb perf-part <trans1> samset
"norsk-utvikle" adj <perf-part> nøyt ub ent <trans1> samset
"norsk-utvikle" adj <perf-part> mask fem ub ent <trans1> samset

invasjonsforsvar
"invasjonsforsvare" verb imp <trans1> <refl4> samset
"invasjonsforsvar" subst nøyt appell ent ub samset
"invasjonsforsvar" subst nøyt appell fl ub samset
industriallianser
"industriallianse" subst mask appell fl ub samset
våpenprosjekter
"våpenprosjektere" verb imp <trans1> samset
"våpenprosjekt" subst nøyt appell fl ub samset
leseangst
"leseangst" subst mask appell ent ub samset
ex.phil.-krampe
"ex.phil.-krampe" subst mask appell ent ub samset
universitetetsvankerne
"universitetetsvanker" subst mask appell fl be samset
idrettslokaler
"idrettslokal" subst mask appell fl ub samset
"idrettslokale" subst nøyt appell fl ub samset
aerobicstime
"aerobicstime" verb inf <trans1> samset
"aerobicstime" subst mask appell ent ub samset
gåturene
"gåtur" subst mask appell fl be samset
"gåture" subst mask appell fl be samset
enhåndsarmhevninger
"enhåndsarmhevning" subst mask appell fl ub samset
legathåndboken
"legathåndbok" subst mask appell ent be samset
tuberkuloserammede
"tuberkuloseramme" adj <perf-part> mask fem nøyt be ent <intrans2> <trans1> samset
"tuberkuloseramme" adj <perf-part> mask fem nøyt be ent <part1/inn> samset
"tuberkuloseramme" adj <perf-part> mask fem nøyt be ent <part1/ned> samset

"tuberkuloseramme" adj <perf-part> ub be fl <intrans2> <trans1> samset

"tuberkuloseramme" adj <perf-part> ub be fl <part1/inn> samset

"tuberkuloseramme" adj <perf-part> ub be fl <part1/ned> samset

jentedoen

"jentedo" subst mask appell ent be samset

studentkafeene

"studentkafe" subst mask appell fl be samset

"studentkafé" subst mask appell fl be samset

skjermregistrering

"skjermregistreringe" verb imp <trans1> <part1/inn> samset

"skjermregistreringe" verb imp <trans11/med> samset

"skjermregistrering" adj pos mask fem ub ent samset

"skjermregistrering" subst mask appell ent ub samset

registreringsautomatene

"registreringsautomat" subst mask appell fl be samset

kvinneforaktere

"kvinneforakter" subst mask appell fl ub samset

konjakkstinkende

"konjakkstinke" adj <pres-part> mask fem nøyt ub be ent fl <intrans2> samset

Hjortefaret

"Hjortefar" subst nøyt appell ent be samset

gudstjenestefellesskap

"gudstjenestefellesskap" subst nøyt appell ent ub samset

"gudstjenestefellesskap" subst nøyt appell fl ub samset

guddommeliggjøre

"guddommeliggjøre" verb inf <trans1> samset

kjerne-nåden

"kjerne-nåde" subst mask appell ent be samset

rødvinsfesten

"rødvinsfest" subst mask appell ent be samset

Sammensetningene er verd en liten ekstradiskusjon: For noen formål er det nok interessant å oppdage også nye sammensetninger. For andre

formål er det mindre interessant. Sammensetning i norsk har jo ikke bare den funksjonen å sette ord på nye begreper - det er også en måte å uttrykke noe som man mer naturlig uttrykker syntaktisk i andre språk. Altså:

(20) Sammensetningenes funksjoner

a. *Sette navn på et bestemt begrep:*

fange	-> innsatt
student	-> høyskolekandidat
splicer	-> filmskjøteapparat
cache (memory)	-> hurtigminne (Gundersen 1996)
designe	-> formgi (Gundersen 1996)

b. *Forenkle et komplekst syntaktisk uttrykk:*

ta podhia tis kareklas (gresk) ≈ stolbena
glavna saobracajnica (bosnisk) ≈ motorvei
innstramningene på HF
-> HF-innstramningene
seminaret om leksikografiens rolle
-> leksikografiseminaret
grensene for kjerneområdene
-> kjerneområdegrensene
anslag over en bestand
-> bestandanslag
bedrifter som videreforedler noe
-> videreforedlingsbedrifter
oppdretterne i EU-landene
-> EU-oppdretterne

Siden sammensetninger på denne måten har to roller å spille i det norske språket (i motsetning til mange andre språk - selv til døves tegnspråk), er det ikke overraskende at det virkelig er mange nyord av denne typen. Det er selvsagt en interessant og viktig oppgave for leksikografene å avgjøre hva slags og hvilke sammensetninger som skal registreres. Antagelig er det bare ord fra den første gruppen som er

interessante - altså der hvor sammensetningene navngir et bestemt begrep. Altså: Ord som *snøbrett, kikkehullskirurgi, skjermregistrering* og *jentedo* er antagelig interessante, mens *konjakkstinkende* og *gudstjenestefellesskap* kanskje ikke er så interessante i første omgang. Antagelig er leksikografenes interesse for et ord knyttet til dets frekvens. Her kan jo datamaskinen som kjent virkelig være til nytte. Kvantitative beregninger er nettopp maskinens forse. Man kan legge inn alle sammensetninger i en database og kjøre frekvenstellinger i den av og til. Hvis en bestemt sammensetning kommer over en viss frekvens, kan man betrakte den som et nytt ord i språket.

Det er interessant å se hva slags ord som er tatt med i ordboka *Nyord i norsk*. Målsetningen for boka er ganske klar:

(21) Fra innledningen i *Nyord i norsk*:

> Nyord kan såleis vere nye anten av språklege grunnar eller av saklege, dvs. at det er noko nytt i verda utafor språket. I boka her er det teke med ord av begge typar. Ordet radar er med fordi det er noko nytt både teknisk og språkleg. Ordet trafikkdøden er samansett av to velkjende gamle ord og dekkjer ikkje noko nytt, folk har blitt drepne i trafikken lenge før 1945. Men sjølve samansetjinga er ny og er teken med som eit døme på bruken av -død(en) som nytt produktivt etterledd. Ordet fireukersferie er språkleg sett ei ordinær nylaging bygd opp av gamle element. Men ordet viser til ei viktig sosial reform i norsk politikk og er derfor med. (*Nyord i norsk*, s. 15)

Fra denne innledningen kan vi trekke ut noen kriterier som er brukt:

(22) Kriteriene til *Nyord i norsk*:

1. Ordet er språklig helt nytt (ikke en sammensetning)
2. Ordet er en sammensetning bygd opp av gamle elementer, men ett av elementene er nytt som produktivt etterledd

3. Ordet er en sammensetning bygd opp av gamle elementer, men det betegner et nytt og viktig begrep

Hvis man skiller ut sammensetningene i et stort tekstkorpus, får man et viktig perspektiv på ordlaging. Sammensetning er en så produktiv prosess at kriterium nr 2 blir vanskelig å forstå. De aller fleste substantiver kan inngå i sammensetninger, og siden substantivene er den desidert største klassen av ord, vil det ofte være slik at et ord brukes som etterledd for første gang. Andre ordklasser er sjeldnere brukt som etterledd i sammensetninger, så her kunne man kanskje tenke seg en viss registrering av førstegangstilfeller.

Kriterium 1 er selvsagt ikke vanskelig å følge. Men hva med kriterium 3? Hvor kjent skal et begrep være før det oppfattes som mer enn en produktiv sammensetning? En måte å avgrense kriterium 3 på, er å innføre et tilleggskriterium: at summen av delene ikke automatisk gir forståelse av hva begrepet er.

(23) Kriterium 3 med et tilleggskriterium A: Helheten er ikke lik summen av delene:

> hanskerom, hagesenter, motorjournalist, midtstopper, maskeringsteip, takhøyde, venstrehåndsarbeid, øremerke, valgskred, tekstforfatter, supperåd, (*Nyord i norsk* 1982)

Uten tilleggskriteriet er det tydelig at kriterium 3 har vært vanskelig å følge. Det ser vi i de følgende eksemplene. Er de virkelig nye og viktige begreper?

(24) Kriterium 3: Nye og viktige begreper?

> mammutbakke (stor hoppbakke), høstfag (fag som tas om høsten), fødeaktiv, hoftebukse, kontinentalsokkelavtale, motekonsulent, løpetrene, reklameskatt, turistdiaré, stjerneskrue, sjømotell, råsykler, rullestolbruker, (*Nyord i norsk* 1982)

Det finnes selvfølgelig mange grunner til ikke å ha et tilleggskriterium som det som er foreslått. En viktig grunn er ikke minst den at vi går

glipp av mange sammensetninger som betegner tidstypiske begreper. En nyordbok som er gitt ut over et bestemt tidsrom, er jo et fabelaktig hjelpemiddel for å finne ut hva slags samfunn som utviklet seg i en viss periode. De følgende ordene er f.eks. enten noe utdaterte, eller så vanlige i dag at det er utrolig at de en gang var nye:

(25) Nyord fra en viss periode gir innsikt i

a. viktige begreper fra perioden, som nå er avleggse:
diggbar, boligvrak, nålefilt, kursplan, helsetruse, forbrukerombudsmann, filkjøring, femdagersuke, venstreintellektuell, skoleveik, reformgymnas

b. fremkomsten av begreper som er viktige også i dag:
popartist, pocketbok, karrierekvinne, kvartsur, kurssenter, identitetskrise, generasjonsbolig, sklisikker, skolebuss, samboer

Nyordene kan altså gi oss viktig innsikt i samfunnet før og nå. Om vi ønsker alle nye ord fra en gitt periode, kunne vi istedenfor tilleggskriterium A velge et tilleggskriterium B:

(26) Kriterium 3 med et tilleggskriterium B:
Ordet må ha forekommet i mer enn én kilde i et tidsrom som strekker seg over minst en måned.

For å kunne bruke tilleggskriterium B er det nødvendig å ha tilgang til mye materiale. Det må være elektronisk søkbart, og man må kunne sette opp flere søkekriterier, som kilde, tidsperiode osv. Igjen er det nødvendig med et elektronisk tekstkorpus for å kunne foreta nøyaktig de valgene og få nøyaktig de resultatene man er ute etter.

4. KONKLUSJON

Jeg har diskutert to sider ved leksikografiens arbeide, slik jeg kjenner det utenfra, i forhold til elektroniske hjelpemidler. Jeg har vist at arbeidet med definisjonsskriving kan forbedres vha to typer hjelpemidler: på den

ene siden tekstkorpora av en viss størrelse, på den andre en organisering av ordbøkene på en måte som gjør dem søkbare i definisjonsfeltene og som gjør at man lett kan hoppe fra ett sted til et annet i ordboka ved å klikke på enkeltord.

I tillegg til definisjonsskrivingen har jeg foreslått at nyordsregistreringen kan gjøres ved hjelp av datamaskinprogrammer og store og stadig nye tekstmengder som er elektronisk tilgjengelige. Slik kan man få ut absolutt alle forekomster som ikke finnes i ordbøkene fra før.

5. REFERANSER

Engø, A. 1997. Karakteriserende personnavn i nyordmaterialet. *I Ord om ord 3* Årsskrift for leksikografi, s. 80-82. Seksjon for leksikografi og målføregransking, Oslo

Gundersen, D., E. Simensen, L.S. Vikør, B. Wangensteen og G. Harildstad (red.) 1997. *Ord om ord 3.* Årsskrift for leksikografi. Seksjon for leksikografi og målføregransking. Institutt for nordistikk og litteraturvitenskap. Universitetet i Oslo.

Gundersen, D. 1996. Om engelsk i norsk. *Språknytt 2,* Oslo.

Guttu, T. (red.) 1993. *Norsk illustrert ordbok. Moderat bokmål og riksmål.* Kunnskapsforlaget, Oslo.

Hovdenak, M., L. Killingbergtrø, A. Lauvhjell, S. Nordlie, M. Rommetveit og D. Worren (red.). 1994. *Nynorskordboka. Definisjons- og rettskrivingsordbok.* Det Norske Samlaget, Oslo.

Killingbergtrø, L. 1997. Bokstavfrekvens og bokstavkombinasjonar. I *Ord om ord 3,* Årsskrift for leksikografi, s. 22-37. Seksjon for leksikografi og målføregransking, Oslo

Landfald, Aa. og K.M. Paulssen. (red.) 1996. *Norsk ordbok. Bokmål.* J.W. Cappelens Forlag, Oslo.

Landrø, M.I. og B.Wangensteen. (red.) 1994. *Bokmålsordboka. Definisjons- og rettskrivingsordbok.* Universitetsforlaget, Oslo.

Leira, V. (red.) 1982. *Nyord i norsk. 1945-1975.* Norsk Språkråd. Universitetsforlaget, Bergen.

Nasjonal leksikografisk database
Status og potensial[1]

TORBJØRN NORDGÅRD
LINGVISTISK INSTITUTT
NTNU

INNLEDNING

Sammenlignet med andre land har Norge en svakt utbygd portefølje av ressurser som elektroniske ordlister, korpus og programmer som foretar automatisk ordmerking ("taggere") eller automatisk syntaksanalyse. Etablering av slike ressurser er krevende, både økonomisk og faglig. Universitetene i Oslo, Bergen og Trondheim er i ferd med å bygge opp datalingvistiske fagmiljøer, men miljøene er små. Disse to forholdene, behovet for ressurser og den svake bemanningen, tilsier et samarbeid mellom universitetene. Dette samarbeidet har funnet sin foreløpige form i den uformelle sammenslutningen *Nasjonal infrastruktur for språkteknologiske ressurser*, forkortet NIFST. I tillegg til universitetene er Telenor Forskning og Utvikling med i dette samarbeidet. NIFST-samarbeidet er konkret manifestert i to relativt store prosjekter: *Norsk Komputasjonelt Leksikon*, forkortet NorKompLeks, og "Tagger-prosjektet" ved Universitetet i Oslo.

NORKOMPLEKS

NorKompLeks er hovedsakelig finansiert av Norges forskningsråd, men med delfinansiering fra Telenor Forskning og Utvikling. Prosjektet har følgende målsettinger:

1. Lage fullstendige og formaliserte beskrivelser av bøyningsmorfologien i bokmål og nynorsk slik at fullformsordlister (lister bestående av bøyningsparadigmer) med morfosyntaktisk informasjon (tall, form,

[1] Takk til Jostein Ven som har kommentert en tidligere versjon av artikkelen.

kjønn, grad, tempus, ...) kan genereres automatisk. Ordforrådet er definert av Bokmålsordboka og Nynorskordboka.
2. Gi fonologiske beskrivelser av oppslagsordene i Bokmålsordboka og Nynorskordboka.
3. Definere fonologiske bøyningsnøkler som kan brukes til å produsere "uttaleparadigmer" med basis i de fonologiske beskrivelsene av grunnformene.
4. Beskrive syntaktisk og semantisk argumentstruktur til verbene i bokmål og nynorsk. Argumentstruktur forstås som valensrammer med informasjon om tematiske roller (agent, recipient, ...), syntaktisk funksjon (subjekt, objekt, ...) og formell realisasjon (nominalfrase, preposisjonsfrase, ...)

Følgende milepæler er definert:

Delprosjekt	Periode	Status
Bokmålsmorfologi	1996	Ferdig i første versjon
Nynorsk morfologi	1997 - August 1998	Ikke ferdig
Bokmålsfonologi	1996 - 1998	Grunnformer ferdig
Nynorsk fonologi	1997 - 1998	Ikke ferdig
Argumentstruktur Bokmål	1997 - Mars 1998	Ferdig
Argumentstruktur Nynorsk	1998	Påbegynt. Forventet ferdig i oktober 1998

La oss se nærmere på hva slags informasjon man i dag kan trekke ut av NorKompLeks-materialet. Vi illustrerer med noen eksempler for substantiver og verb. Leksemet *bil* er representert omtrent slik i den maskinleselige versjonen av Bokmålsordboka (jf. Landrø og Wangensteen 1994):

Kodefelt	**Kodefor-klaring**	**Informasjon**
NB001	Leksem	bil
NB001a	Bøyningskode	M1
ARTNR	Artikkelnummer	5868
TR007		
..OPP	Oppslagsform	**bil** *m1*
..ETY	Etymologi	(fork. av automobil)
..DEF	Definisjon	motordrevet transportvogn på fire (el. flere) hjul
..UTR	Brukseksempel	*lasteb-, personb-, ruteb-, vareb-*
..UTR	Brukseksempel	*ha, kjøpe, kjøre b-*
..UTR	Brukseksempel	*det blir flere og flere b-er på veiene*
..UTR	Brukseksempel	*b-en går som et skudd*
..UTR	Brukseksempel	*ta en b-*
..FOR	Forklaring	drosje

I den morfologiske delen av NorKompLeks er samme leksem beskrevet slik:

```
bil,[m1],5868
```

Ved hjelp av NorKompLeks-koden m1 kan et genereringsprogram (utviklet ved Lingvistisk institutt, NTNU) produsere et bøyningsparadigme med morfosyntaktisk informasjon:

```
bil:    subst,masc,sg,ubest,app
bilen:  subst,masc,sg,best,app
bilene: subst,masc,pl,best,app
biler:  subst,masc,pl,ubest,app
```

Hvor mye informasjon som skal være med og hvilke benevnelser som skal benyttes, kan brukerne selv definere med utgangspunkt i de distinksjonene som NorKompLeks-koden m1 er opphav til, nemlig `sg_ind_m`, `sg_def_m`, `pl_ind_m` og `pl_def_m`:

Morfologisk kode m1:	**Bøyningsbeskrivelse**	**Bøyningsoperasjon**
	`sg_ind_m`	`0`
	`sg_def_m`	`en`
	`pl_ind_m`	`er`
	`pl_def_m`	`ene`

Symbolene i bøyningskodene skal fortolkes slik:

sg = entall
ind = ubestemt form
m = hankjønn
pl = flertall
def = bestemt form

Bøyningsoperasjonene forteller hva som skal gjøres med grunnformen. "0" betyr at intet skal gjøres med grunnformen, "en" betyr at suffikset **-en** skal legges til grunnformen, "er" betyr at suffikset **-er** skal legges til, osv.

Anvendt på grunnformen bil får vi

bil : sg_ind_m
bilen : sg_def_m
biler : pl_ind_m
bilene : pl_def_m

Bøyningsbeskrivelsen *`sg_ind_m`* blir i den nåværende versjonen av NorKompLeks fortolket slik:

```
sg_ind_m  =  subst,masc,sg,ubest,app
```

Men man kan fortolke det på andre måter, f.eks. som

```
sg_ind_m  =  nomen,hankjønn,entall,ubestemt
```

dersom man ønsker norske termer. En variant som ligger nær opp til nokså etablerte internasjonale standarder innen datamaskinell leksikografi, jf. f.eks. Langendoen & Simons (1995), kan være

```
sg_ind_m  =  [category  =  noun,
              proper  =  -,
              agreement  =  [gender  =  masculine,
                             number  =  sg,
                             form  =  indefinite]
```

Bøyningsbeskrivelsene kan man altså fortolke som man måtte ønske, men det viktigste fra et leksikografisk synspunkt må være at distinksjonene er foretatt.

I tillegg til morfologiske koder er det utarbeidet fonemiske beskrivelser for oppslagsordene, og vi henter vårt eksempel fra beskrivelsen av *bil*:

```
bil   "bi:l   m1   5868
```

Legg merke til at NorKompLeks bruker SAMPA istedenfor IPA ved koding av fonologisk form. Dette skyldes tekniske begrensninger ved hjelpeprogrammene som brukes i tilknytning til kontroll av de fonemiske beskrivelsene. SAMPA er ikke så detaljert som IPA, men SAMPA kan brukes uten at man trenger å definere egne tegnsett, noe som gjør at SAMPA-beskrivelser kan fortolkes uavhengig av maskintyper og operativsystem. Det er en enkel sak å foreta automatisk oversettelse av SAMPA-beskrivelser til IPA dersom definisjonene for det aktuelle IPA-tegnsettet er tilgjengelig.

Fonologisk informasjon kan legges til den morfologiske beskrivelsen via identifikasjonsnøkkelen `5868`. Denne informasjonen

legges til oppslagene i Bokmålsordboka, f.eks. ved å legge inn et nytt felt *Uttale*:

Kodefelt	**Kodefor-klaring**	**Informasjon**
NB001	Leksem	bil
NB001a	Bøyningskode	M1
ARTNR	Artikkelnummer	5868
TR007		
..OPP	Oppslagsform	**bil** *m1*
..UTT	Uttale	"bi:l
...	...	...

Dette betyr at i fremtidige utgivelser av ordbøkene til Seksjon for leksikografi kan alle oppslagsordene utstyres med uttaleinformasjon, om det er ønskelig.

Fonologiske bøyningsparadigmer er under utarbeidelse, og vi kan skissere hva som etter hvert skal komme:

"bi:l	**sg_ind_m**
"bi:l%n, "bi:l@n	**sg_def_m**
""bi:l@r	**pl_ind_m**
""bi:l%n@,""bi:len@	**pl_def_m**

SAMPA-symbolet " betyr tonem 1, "" står for tonem 2, %n symboliserer stavelsesbærende n og @ er IPA-symbolet ə (redusert e).

La oss se nærmere på et verb, f.eks. *kverulere* (v2 i feltet NB001a er den relevante morfologiske bøyningskoden i NorKompLeks, altså en annen kode enn i det originale materialet ved Seksjon for leksikografi ved UiO):

NB001	**kverulere**
NB001a	**v2**
ARTNR	**33171**
TR007	

```
..OPP   kverule>re  v2
..ETY   (fra  mlat  'klage')
..Def   stadig  klage,  si  imot
```

Det minimale morfologiske oppslagsordet:

```
kverulere,[v2],  33171
```

Den fonologiske informasjonen for dette verbet er representert slik (symbolet { er SAMPA-beskrivelsen av æ, mens } står for norsk u:

```
kverulere    kv{r}"le:r@      v2       33171
```

Alle verbene i NorKompLeks er i tillegg utstyrt med informasjon om hvilke syntaktiske og semantiske argumenter de kan opptre sammen med. Verbet *kverulere* brukes intransitivt, altså med subjekt. Dette er leksikografisk sett viktig informasjon, men den er ikke systematisk tilgjengelig i de eksisterende ordbøkene ved Seksjon for leksikografi. I argumentstrukturlistene i NorKompLeks er denne informasjonen slik for verbet *kverulere*:

```
kverulere,33171,[intrans1]
```

intrans1 betyr "verb som tar ett syntaktisk argument, og dette argumentet har funksjonen subjekt, tematisk rolle agent og formell realisering som nominalfrase". Dette bør fortolkes som en konstruksjonstype, altså en setning som har et obligatorisk setningsledd, nemlig et referensielt subjekt som er en nominalfrase med rollen agent. Den formelle beskrivelsen er slik:

Konstruksjonstype	**Argument**
`intrans1`	`arg1:su,ag,np`

Vi tar også med et eksempel på et standard transitivt verb:

```
kvele,33136,[trans1]
```

Koden *trans1* betyr "verb som tar to syntaktiske argumenter der det ene har funksjonen subjekt, tematisk rolle agent og formell realisering som nominalfrase, mens det andre har funksjonen objekt, tematisk rolle "theme" og formell realisering som nominalfrase". Formell beskrivelse:

Konstruksjonstype	**Argument**
trans1	arg1:su,ag,np
	arg2:obj,th,np

Slike beskrivelser kan enten ekspanderes slik at de kan brukes i nyere teoretiske rammeverk, f.eks. Leksikalsk-funksjonell grammatikk, Styrings- og bindingsteori, osv., jf. Nordgård (1998). De kan også reduseres hvis man bare er interessert i opposisjonen intransitiv - transitiv.

Dersom oppslagsord i Bokmålsordboka utstyres med fonologisk informasjon og argumentstrukturinformasjon, kan vi få oppslag som

```
NB001   kvele
NB001a  v2,v134
UTT     ""kve:l@,""kv{:rL@
ARG     trans1
ARTNR   33136
TR007
```

rL er SAMPA-versjonen av "tjukk l". Kodene `v2,v134` er NorKompLeks-koder.

POTENSIAL FOR LEKSIKOGRAFISK ARBEID

Norske ordbøker inneholder i varierende grad uttaleinformasjon, og transitivitetsegenskapene til verb er ikke systematisk beskrevet. NorKompLeks-basen gjør det mulig å inkludere denne informasjonen i kommende versjoner av ordbøkene ved Seksjon for leksikografi. Vi kan se for oss oppslag som dette:

bil m1 (fork. av *automobil*) (utt. *'bi:l*) motordrevet transportvogn på fire (el. flere) hjul *lasteb-, personb-, ruteb-, vareb- / ha, kjøpe, kjøre b- / det blir flere og flere b-er på veiene / b-en går som et skudd / ta en b-* drosje

kvele v2 el. *kvalte, kvalt*, trans1 (norr kvelja 'pine, plage') (utt: *''kve:lə* el *''kvæ:ɽə*) **1** drepe ved å stanse lufttilførselen til lungene *den drepte var kvalt med et skjerf / k-s av røyk /* adj i pr pt: *luften var k-nde tykk,* tung å puste i */ det var k-nde varmt* **2** slokke, dempe, hindre *k- ilden med et teppe / k- et gjesp / forsøket ble kvalt i fødselen / k- motoren* gi den så lite gass at den stanser

Å legge inn uttaleinformasjon er kanskje kontroversielt fordi det ikke finnes noen offisiell uttalenorm for norsk. Dermed kan uttaleinformasjonen fra NorKompLeks oppfattes som en normeringstilsnikelse siden denne informasjonen er basert på sørøstlandsk tale. Dette temaet skal ikke diskuteres ytterligere her.

Elektroniske versjoner av ordbøkene kan utstyres med mer informasjon enn de foreliggende versjonene, f.eks. slik:

kvele (norr kvelja 'pine, plage') (utt: *''kve:lə* el *''kvæ:ɽə*)
1 drepe ved å stanse lufttilførselen til lungene *den drepte var kvalt med et skjerf / k-s av røyk /* adj i pr pt: *luften var k-nde tykk,* tung å puste i */ det var k-nde varmt*
Bøyning
Syntaks
2 slokke, dempe, hindre *k- ilden med et teppe / k- et gjesp / forsøket ble kvalt i fødselen / k- motoren* gi den så lite gass at den stanser
Bøyning
Syntaks

Brukere kan klikke på *Bøyning* for å få frem bøyningsparadigmer, og *Syntaks* vil gi informasjon om hvilke krav verbet stiller til sine omgivelser (subjekt, objekt, osv). Det er også mulig å lage lenker til en

konkordansbase slik at brukerne kan finne eksempler på leksemet i bruk, men da må en slik base først etableres.

Ordmerkingsprogrammet som er under utvikling ved Tekstlaboratoriet ved Universitetet i Oslo, kan integreres med den elektroniske versjonen av Bokmålsordboka ved at programmets lemmaliste settes lik leksemene i Bokmålsordboka. Denne versjonen av ordmerkingsprogrammet settes så til å arbeide med nyere maskinleselige tekster, og man kan få ut lister over ord som ikke finnes i ordboka sammen med brukskontekster. Dette vil være til hjelp for leksikografenes arbeid, siden man da raskt kan identifisere nyord og deretter produsere oppslagsord med brukseksempler.

KORT OM DATABASESTRUKTUREN

Den maskinleselige versjonen av Bokmålsordboka har en unik nøkkel for hvert oppslag (et heltall). Dette er en stor fordel når forskjellige leksikalske komponenter skal settes sammen, særlig fordi komponentene utvikles av ulike fagmiljøer. Så lenge man sørger for å holde seg til et vedtatt nøkkelsystem, kan Bokmålsordboka og Nynorskordboka fungere som ryggraden i en nasjonal leksikografisk ressurs der ulike fagmiljøer lar sine resultater bli tilgjengelige. Oppslagsordet *kvele* kan presenteres på denne måten:

Oppslag	Offisiell BMO	BMO bøyning	NorKomp Leks bøyning	NorKomp Leks fonologi	NorKomp Leks arg.str	Annet
kvele	**kvele**, (norr *kvelja* 'pine, plage') drepe ved å stanse lufttilførselen til lungene *den drepte var kvalt med et skjerf / k-s av røyk* ...	v2 el. *kvalte, kvalt*	v2, v134	"kve:lə el "kvæ:ʈə	trans1	

VEDLIKEHOLDSANSVAR

Det er viktig at det etableres et organ som har ansvar for vedlikehold og oppdatering av den nasjonale leksikografiske databasen. Det nasjonale samarbeidet for språkteknologiske ressurser (NIFST) bør, sammen med leksikografene i Oslo, ha ansvar for at materialet til enhver tid er oppdatert og holder akseptabel faglig standard. I skrivende stund har ingen påtatt seg ansvaret for vedlikehold av databasen, men Dokumentasjonsprosjektet ved Universitetet i Oslo ville være et godt valg fordi Dokumentasjonsprosjektet har relevant programmerings-kompetanse og erfaring fra arbeid med leksikografisk materiale.

REFERANSER

Landrø, Marit Ingebjørg og Boye Wangensteen (1994). *Bokmålsordboka. Definisjons- og rettskrivningsordbok.* Universitetsforlaget, Oslo.

Langendoen, D. Terence og Gary F. Simons (1995). "A Rationale for the TEI Recommendations for Feature-Structure Markup". I *Computers and the Humanities* 29: 191 - 209.

Nordgård, Torbjørn (1998). "Norwegian Computational Lexicon (NorKompLeks)". I Proceedings of the 11th Nordic Conference of Computational Linguistics NODALIDA 98. Center for Sprogteknologi, Köbenhavns Universitet.

Noen tanker om parallellkorpus og leksikografi

STIG JOHANSSON
INST. FOR BRITISKE OG AMERIKANSKE STUDIER, UIO

1. Å HATE OG ELSKE PÅ ENGELSK OG NORSK

For noen dager siden så jeg følgende i en tekst på Aftenpostens lederside:

(1)

> Paul Jefferson er imidlertid en tidligere minerydder fra den britiske hær, som har ryddet miner i Angola, Afganistan [...], inntil han mistet et ben og ble blind av en landmine i Kuwait. Han sier: "Jeg *hater å bruke* [min utheving] mine egne skader for å skape troverdighet [...] (*Aftenposten* 9. januar 1998, s. 14)

Jeg tenkte: dette ser ut som en anglisisme. Den samme dagen fant jeg et liknende eksempel i Østlandets Blad:

(2)

> Jeg *hater å bringe* sladderen videre. Men hva skal man ellers gjøre med den? (oppgitt å være et sitat fra Shirley MacLaine; *Østlandets Blad* 9. januar 1998, s. 15)

Jeg spurte et par kolleger om den uthevede teksten var idiomatisk norsk. De var i tvil om det og sa at de oppfattet uttrykksmåten som engelsk. Jeg gikk til vårt engelsk-norske parallellkorpus, som jeg skal introdusere senere. Hva sier korpuset om saken?

Hvor vanlige er de norske og engelske verbene?

For det første: engelsk *hate* og norsk *hate* har forskjellig frekvens. Det engelske verbet er omtrent tre ganger så hyppig som det norske, og det

samme gjelder *love* sammenlignet med *elske* (se den venstre kolonnen under; tallene er basert på de skjønnlitterære tekstene i korpuset):

	Originaltekst	Oversettelse
N *hate*	23	34
N *elske*	36	90
E *hate*	67	25
E *love*	100	62

Går vi derimot til oversatt tekst, blir bildet annerledes (se den høyre kolonnen). *Hate* og *elske* er faktisk noe vanligere i norsk tekst som er oversatt fra engelsk enn *hate* og *love* i engelsk tekst som er oversatt fra norsk. I begge tilfellene går tallene i retningen av frekvensen i det andre språket. Dette er et mønster som vi ofte finner når vi sammenligner tekster i forskjellige språk, dvs. forskjellene er tydeligst i originalteksten, og det er tendens til at de viskes ut i oversettelsene. Men bare delvis.

Oversatt tekst sammenlignet med originaltekst

Det er ikke ukjent at oversatt tekst kan være forskjellig fra originaltekst på det samme språket. Martin Gellerstam har skrevet om emnet og har også vært i Oslo og presentert resultater fra sin forskning om forholdet mellom svensk originaltekst og svenske oversettelser fra engelsk.[1] Tage Danielsson har en parodi på dårlig oversettelse fra engelsk til svensk. Jeg siterer begynnelsen av teksten:

(3)

- Är du verkligen gående iväg ifrån mig, hon suckade.
- Darling, så som varande gift tidigare ser jag ingen väg ut, han svarade och flyttade mot dörren. En kärleksaffär som denna kan inte vara för evigt, du vet det. (*Tage Danielssons paket*, Wahlström & Widstrand, s. 97)

Oversettelsen av verbene hate/elske og hate/love

Heldigvis er oversettelsene i vårt korpus på et helt annet nivå. Hvis vi nå ser på oversettelsen av de aktuelle verbene, finner vi et interessant mønster:

N *hate*	IKKE	E *hate*	2 (av 23)
N *elske*	IKKE	E *love*	4 (av 36)
E *hate*	IKKE	N *hate*	31 (av 67)
E *love*	IKKE	N *elske*	37 (av 100)

Det betyr at norsk *hate* og *elske* blir oversatt med *hate* og *love* unntatt i et fåtall tilfeller, mens de engelske verbene ofte blir oversatt med noe annet enn *hate* og *elske*. Vi kan tolke dette slik at de engelske verbene har et videre betydningsområde enn de norske verbene. Kanskje kan vi også uttrykke saken slik at de norske verbene uttrykker en sterkere følelse. Her er noen eksempler på andre oversettelser enn *hate* for engelsk *hate*:

(4)
avsky, forakte, ha en ingrodd motvilje mot
ikke kunne fordra/utstå
ikke like/tåle, ikke ha lyst til

(5)

I hate it	det er helt fælt
	jeg synes det var så flaut ...
I hate the country	jeg liker meg ikke på landet
... had hated leaving home	... hadde vært knust over ...
I hate to have to keep reminding you, but ...	det er ikke for å mase, men...

Vi ser en klar forskjell mellom engelsk og norsk hvis vi sammenligner hvilke typer objekt verbene tar:

	Originaltekst		Oversettelse	
	Person	Ikke person	Person	Ikke person
N *hate*	65 %	35 %	35 %	65 %
N *elske*	61 %	39 %	36 %	64 %
E *hate*	27 %	73 %	56 %	44 %
E *love*	46 %	54 %	65 %	35 %

Det vil si: i originaltekst tar de norske verbene et personlig objekt i de fleste tilfellene, mens de engelske verbene har en overvekt av ikke-personlige objekt. På engelsk finner vi ofte eksempler som: ... *hate that finicky style, travel, being crammed together, starvation, a job*, osv.; ... *love soap operas, shoes, fish, committees, doing things*, osv. Oversettelsene viser igjen et mønster som avspeiler originaltekstene, med hyppig bruk av ikke-personlige objekt i oversatt norsk og av personlige objekt i oversatt engelsk.

Konklusjonen blir: korpuset bekrefter den intuitive vurderingen av eksempel (1) og (2). De norske og de engelske verbene har forskjellig distribusjon, men oversettere ser ikke ut til å være tilstrekkelig oppmerksomme på forskjellene. Parallellkorpuset gir et mye rikere bilde enn tospråklige ordbøker. Jeg skal senere gi flere eksempler på hvordan vi kan bruke et parallellkorpus. Men først noen generelle bemerkninger.

2 HVA ER ET PARALLELLKORPUS?

Man pleier å skille mellom to hovedtyper:

- Et korpus med originaltekster på ett språk og oversettelser til ett eller flere andre språk.

- Et korpus med sammenlignbare originaltekster på to eller flere språk.

Mange avviser den første typen, på grunn av at oversatt tekst kan skille seg fra originaltekst (som vi har sett), og de understreker sterkt at man må bruke parallelle originaltekster. Men det er også problemer med dette: 1) Hvordan skal man matche tekster i de forskjellige språkene

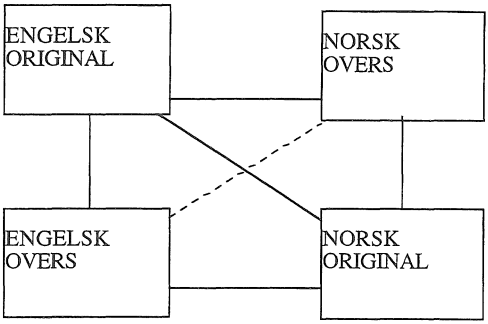
under korpusbyggingen? 2) Hvordan kan man vite hva man skal sammenligne etter at korpuset er ferdig?

Jeg kan ikke nå gå inn på en nærmere diskusjon av argumentene for og imot de to typene. Hovedsaken er at det ikke er et spørsmål om enten-eller. Vi kan få begge typene i det samme korpuset, slik vi har har gjort det i vårt eget engelsk-norske parallellkorpus. Og vi får da fordelene med begge, og kan bruke den ene typen for å kontrollere den andre.

3 ENGELSK-NORSK PARALLELLKORPUS[2]

Engelsk-norsk parallellkorpus består av 200 tekster, totalt ca. 2.6 millioner ord. Det inneholder 50 norske originaltekster (30 skjønnlitteratur + 20 sakprosa) + engelsk oversettelse, og et tilsvarende antall engelske originaltekster + oversettelse til norsk. Hver tekst består av et utdrag på ti til femten tusen ord tatt fra begynnelsen av det aktuelle verket. Tekstene er publisert på anerkjente forlag etter 1980 og er oversatt av profesjonelle oversettere. Original og oversettelse er parallellstilt på setningsnivå, slik at det er mulig å søke etter ord eller uttrykk i ett språk og få fram tilsvarende tekst i det andre språket. Se videre beskrivelsen av korpuset på vår hjemmeside: http://www.hf.uio.no/iba/prosjekt/

Figur 1 viser strukturen i korpuset. Det finnes seks måter å sammenligne på. Vi kan sammenligne originaltekster og oversettelser i begge retningene. Vi kan også sammenligne parallell originaltekst (se den heltrukne diagonale linjen). Vi kan sammenligne originaltekst og oversatt tekst både på engelsk og norsk (se de vertikale linjene). Og vi kan sammenligne oversatt tekst på begge språkene (se den prikkede diagonale linjen). For å lette sammenligningen har vi laget alle hoveddelene omtrent like store. Dette er grunnen til at jeg kunne bruke absolutte tall i sammenligningen tidligere.

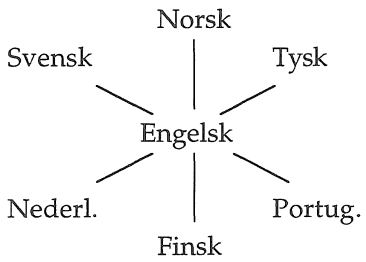


Figur 1: Strukturen i ENPC

4. Å NAVIGERE I KORPUSET

Å bruke korpuset er som å *navigere*, og som all annen navigasjon, krever dette dyktighet og innsikt. En annen metafor er å *samtale* med korpuset. Det kommer ikke noe automatisk fra korpuset. Man må stille spørsmål. Og svarene kan føre til nye spørsmål. Korpusarbeid blir ofte oppfattet som noe dødt og kjedelig. Det er en misforståelse som vi ennå finner blant mange språkforskere. Godt korpusarbeid krever fantasi og innsikt.

Før jeg gir noen flere eksempler på korpusnavigasjon, skal jeg kort kommentere en ny utvikling i korpusarbeidet vårt. Figur 2 viser at vi har trukket inn flere språk i prosjektet. Fordi vi har et nordisk samarbeid, kan vi ta med svensk og finsk. I Oslo har vi også tatt med oversettelser av mange av våre engelske originaltekster til tysk, nederlandsk og portugisisk. Som figuren viser, kan vi altså med utgangspunkt i engelsk originaltekst sammenligne med seks andre språk.



Figur 2: Et flerspråklig korpus

Det er en ting som jeg vil understreke spesielt før vi går over til eksemplene. Når vi bruker oversettelser, er det ikke primært fordi vi ønsker å avsløre feil. Vi ser på oversetteren som en ressurs som vi bruker for å få innsikt i forholdet mellom språkene, og for å få innsikt i oversettelsesproblemer og karakteristiske trekk ved oversatt tekst. Og vi har bygd inn en kontrollfunksjon gjennom strukturen i korpuset.

Modalpartikkelen nok

Diskurspartikler er et stort problemområde for oversettere. Hva kan vi se i korpuset? Tabell 1 viser de viktigste typene korrespondanser for den norske modalpartikkelen *nok* (basert på de skjønnlitterære tekstene i korpuset). Her kan vi blant annet se:

• *Nok* er mye vanligere i originaltekst enn i oversatt tekst: 141 mot 79. Vi har et stort antall forskjellige typer korrespondanser: adverb, modale hjelpverb, kommentarsetninger osv. Korrespondansene er forskjellige avhengig av om vi går fra norsk til engelsk eller omvendt. Merk spesielt den store forskjellen i nullkorrespondanser, dvs. eksempler hvor vi ikke kan spore en tilsvarende form i det andre språket. To tredjedeler av *nok* i oversatt tekst ser ut til å komme ingensteds fra.

• Hva skjer om vi ser på hovedkorrespondansen for *nok* ved oversettelse fra norsk til engelsk, dvs. *probably*? Se tabell 2. Her får vi hovedsakelig norske adverb og veldig få nullkorrespondanser. En sammenligning av tabell 1 og 2 har en del å si om forskjellen mellom en diskurspartikkel og et vanlig adverb som *probably*.

I Johansson & Løken (1997) er det en mer detaljert diskusjon av engelske korrespondanser til modalpartikkelen *nok*. Igjen får vi et mye mer nyansert bilde enn i en tospråklig ordbok. Men her er det også ny innsikt å hente om de enkelte språkene. Oversettelse forutsetter en tolkning av originalteksten. Den holder opp et speil for originalspråket. Gjennom å studere speilbildet får vi ny innsikt i den norske diskurspartikkelen.

Tabell 1: Korrespondanser til den norske modalpartikkelen *nok*, uttrykt i prosent for hver kolonne

Korrespondanse	E oversettelse (N = 141)	E original (N = 79)
probably	25	6
annet adverb	21	4
verbkonstr	11	10
setning	9	10
andre former	3	5
null	31	65

Tabell 2: Korrespondanser til det engelske adverbet *probably*, uttrykt i prosent for hver kolonne

Korrespondanse	N oversettelse (N = 94)	N original (N = 141)
nok	3	25
vel	6	28
antagelig(vis)	21	3
kanskje	3	9
sannsynligvis	37	16
sikkert	11	9
trolig	3	1
andre former	13	6
null	2	4

Substantivet mind

Substantivet *mind* i engelsk er et vanskelig ord, både fra et enspråklig og et to- eller flerspråklig synspunkt. Her er noen substantiver som er brukt i norske oversettelser:

tanke/tanker, sinn, sjel, hjerne, hode, indre, bevissthet, fantasi, forstand, ånd

Tanker og de bestemte formene *tanken/tankene* er vanligst, etterfulgt av *sinn*, som vanligvis blir satt opp først i tospråklige ordbøker, og deretter *sjel* og *hjerne*. Frasen *in X's mind* ble gjengitt på mange forskjellige måter:

i tankene, i hennes sinn, i hennes indre, i hennes sjel, i fantasien
din, i mine ører, i Xs øyne, [ser ham] for meg, [gjentatt] for seg selv

Det som er mest slående, er avhengigheten av konteksten. *Mind* inngår ofte i mer eller mindre faste uttrykk og svarer da gjerne til et norsk verb som betegner en mental prosess og tar et personlig subjekt:

(6)

adjust one's mind	tilpasse seg
bear in mind	huske
change one's mind	ombestemme seg
have a (good) mind to	være fristet til
have in mind	tenke seg/på
keep the mind off	la være å tenke på
make up one's mind	bestemme seg
put out of one's mind	slå fra seg
take one's mind off	gjennomgå
turn over in one's mind	fundere på

I slike tilfeller trengs det ikke noe substantiv som svarer til *mind*. Det samme mønsteret finner vi ofte hvor engelsk har et ikke-personlig subjekt eller hvor subjektet er *X's mind*:

(7)

in my mind it was as if	jeg følte det som
there was no doubt in her mind	hun var ikke i tvil
it crossed his mind	han lurte på
her mind was so nimble	hun var rask i oppfattelsen
his mind began to drift	han drev bort

Den norske tendensen til å begrense seg til å bruke et mentalt verb og en referanse til en person, heller enn til en persons sinn, er spesielt slående i norske originaltekster:

(8)

sånn jeg tenker	the way my mind works
hun kjente at det raste en storm i henne	her mind was in a turmoil
ellers var hun jo helt klar	her mind was as clear as anyone's
bilder kom for ham	images came to his mind
enda en gang skjøt spørsmålet gjennom ham	again the question ran through his mind

Der hvor den norske originalteksten har et substantiv som svarer til *mind* i den engelske oversettelsen, finner vi enda større variasjon enn ved oversettelse fra engelsk til norsk.

Konklusjonen fra dette eksemplet blir at de to språkene gjerne refererer til mentale prosesser på forskjellige måter, og at korrespondansene varierer avhengig av konteksten. Det finnes ingen klar korrespondanse i norsk til engelsk *mind*, og i omtrent halvparten av tilfellene har den norske teksten en form uten et tilsvarende substantiv.

Mangelen på en klar korrespondanse gjør også at *mind* ikke er så hyppig brukt i oversettelser som i engelsk originaltekst (96 mot 138 forekomster i de skjønnlitterære tekstene av korpuset). Forholdet er akkurat det samme som ved den norske modalpartikkelen *nok*, som også er underrepresentert i oversatt tekst (se ovenfor).

Norsk hende *og engelsk* happen

Med eksemplet *mind* ville jeg spesielt vise avhengigheten mellom ord og kontekst. I mitt siste eksempel vil jeg fokusere på det nære forholdet mellom leksikon og syntaks. Norsk *hende* og engelsk *happen* ser i utgangspunktet ut til å være uproblematiske fra et tospråklig synspunkt. Hva sier korpuset?

På norsk brukes ofte uttrykket *det hender/hendte at* ... (tilsvarende uttrykk finnes også på svensk). Vi har tre hovedmønstre i oversettelsene, alle uten *happen*:

A. Adverbial

Den engelske teksten har ofte et adverbial, vanligvis *sometimes*:

(9)
Det hender forresten at jeg tar med en ting eller to ... (KF1)
= Incidentally, I *sometimes* help myself to a thing or two ..

Andre former i den engelske teksten er *occasionally* og *once in a while.*

B. Modale hjelpeverb

I eksemplet under finner vi ett tilfelle med *sometimes* og ett med et modalt hjelpeverb:

(10)
...det hender jo at gamle mennesker ikke er fattige, at de til og med er rike ... (KA1)
= ... old people are *sometimes* anything but poor, *might* even be rich ...

Foruten *might* finner vi også det modale hjelpeverbet *would.*

C. Adverbial pluss et modalt hjelpeverb

Den norske *hende*-setningen kan svare til et adverbial pluss et modalt hjelpeverb:

(11)
Det hendte at vi møttes i døråpningen ... (GS1)
= *Sometimes* we *would* bump into each other in the doorway ...

Det hender (!) også at norsk har dobbel markering, dvs. en *hende*-setning og et modalt hjelpeverb i leddsetningen:

(12)
... *det hendte* at Sverre *kunne* få øya fulle av tårer ... (JM1)
= ... *sometimes* his eyes *would* fill with tears ...

Ved siden av det formelle subjektet *det* og verbet *hende* har den norske teksten noen ganger et modalt hjelpeverb (vanligvis *kan*) eller et adverbial:

(13)
Det er ikke noe farlig, men *det kan hende* du mister litt av håret ditt, Herman. (LSC1)
= "It isn't anything serious, but you *might* lose a little of your hair, Herman."

(14)
Men *det hendte aldri* at jeg hilste først. (EHA1)
= But I *never* greeted them first.

Igjen har den engelske teksten en enkel setning med et modalt hjelpeverb (13) eller et adverbial (14).

Unntaksvis har den engelske oversetteren valgt en setning med *happen,* her i kombinasjon med modale hjelpeverb:

(15)
Det hendte at Maria forsøkte seg på en sigarett ... (BV1)
= *It might happen* that Maria *would* try a cigarette ...

Slike eksempler forekommer ikke i de engelske originaltekstene, hvor *it* regelmessig er referensielt i forbindelse med *happen* og forteller om noe som skjer:

(16)
It happens. (ABR1)
= *Slikt* skjer.

Men det finnes spesielle tilfeller hvor det formelle subjektet *it* brukes sammen med *happen*:

(17)
"I don't know if I mentioned that *it so happens I train dogs*." (AT1)
= "Jeg husker ikke om jeg nevnte det tidligere at *jeg dresserer hunder*?"

(18)
... *it happens* that the rent on my apartment was due the next day. (SG1)
= ... husleien min forfalt *tilfeldigvis* neste dag.

Her har ikke norsk *det* pluss *hende*. Merk nullkorrespondansen i (17) og adverbialet i (18), hvor mønstret er akkurat det motsatte av det vi finner i eksempel (9) ovenfor.

Forholdet mellom engelsk *happen* og norsk *hende* er avgjort ikke et enkelt tilfelle av leksikalsk korrespondanse. For å avdekke slike komplekse korrespondanser, hvor det er både syntaktiske og leksikalske forskjeller, er det viktig å kunne ha tilgang til et parallellkorpus.

5. TEKSTLEKSIKOGRAMMATIKK

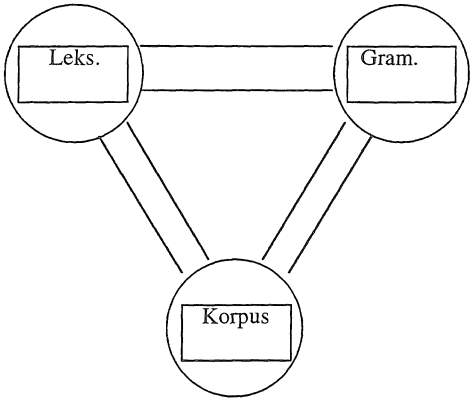
Jeg skal nå prøve å gi en kort oppsummering. Jeg har ønsket å vise:

- korrespondanser mellom språk er avhengig av konteksten (kanskje best illustrert gjennom eksemplet *mind*);

- leksikalske og syntaktiske forskjeller er knyttet nært sammen (kanskje best illustrert gjennom eksemplet *det hender*).

Dette er sikkert ikke noe nytt for leksikografer. Det nye er kanskje bare hvilken bruk man kan gjøre av parallellkorpus for å avdekke og beskrive slike korrespondanser.

Men parallellkorpus er ikke bare viktig i arbeidet med å skrive ordbøker. Jeg har noen ganger fremført tanken om at vi trenger en ny type språkbeskrivelse som kobler sammen leksikon, grammatikk og korpus. Se figur 3.



Figur 3: Leksikon, grammatikk og korpus: en integrert modell

Tradisjonelle ordbøker og grammatikkbøker er begrenset i omfang. Det er umulig å vise språk i kontekst annet enn i meget begrenset utstrekning. Vi kan nå tenke oss en elektronisk ordbok og en elektronisk grammatikkbeskrivelse som begge er koblet til et korpus. På denne måten kan vi vise forholdet mellom ord og konstruksjon og tekst (f.eks. for *mind* og *det hender*).

Like viktig er det å koble sammen leksikon og grammatikk. Leksikografer og grammatikere har altfor ofte arbeidet uavhengig av hverandre og med forskjellig begrepsapparat. Men det er ingen skarp grense. Det viser mitt eksempel med *det hender*. Michael Halliday ser på *lexis* og *grammar* som "different ends of the same continuum – they are the same phenomenon as seen from opposite perspectives".[3] Det er også helt klart at ordbøker vanligvis inneholder grammatisk informasjon, og at grammatikkbeskrivelser inneholder leksikalsk informasjon, f.eks. lister med verb som tar spesielle typer komplement.

Fremtidens ordbok, enten det er snakk om ett, to eller flere språk, er ikke en konvensjonell ordbok og heller ikke en leksikalsk database, men en kombinasjon slik som den som er illustrert i figuren. Dette er det målet vi bør arbeide for. Halliday pleier å snakke om *lexico-grammar* – kanskje kan vi kalle den nye språkbeskrivelsen *tekstleksikogrammatikk*.[4]

NOTER

1 Se Martin Gellerstam, Translationese in Swedish novels translated from English, i Lars Wollin & Hans Lindquist (utg.), *Translation Studies in Scandinavia*, Lund: CWK Gleerup, 1986, s. 88-95.

2 Engelsk-norsk parallellkorpus er blitt finansiert gjennom midler fra Det historisk-filosofiske fakultet og Institutt for britiske og amerikanske studier, Universitetet i Oslo. Jeg er takknemlig for støtten, og vil også spesielt takke mine medarbeidere i prosjektet: Knut Hofland, Bergen, som har skrevet vårt program for automatisk parallellstilling av tekst i forskjellige språk; Jarle Ebeling, Oslo, som har utarbeidet et søkeprogram for parallellkorpus og som, sammen med Signe Oksefjell, Oslo, har utført det meste av arbeidet i forbindelse med korpusbyggingen. For utviklingen av prosjektet har det også vært av avgjørende betydning at vi kunne etablere det nordiske forskernettverket "Språk i kontrast", finansiert av Nordisk Forsker-utdanningsakademi, og at vi fikk sette opp en forskergruppe ved Senter ved høyere studier ved Det Norske Videnskaps-Akademi 1996-1997.

3 See M.A.K. Halliday, *Introduction to Functional Grammar*, 2nd ed., London: Arnold, 1994, s. 15.

4 Hilde Hasselgård har pyntet på min norsk. Jeg er takknemlig for det og også for at hun har vært en god samtalepartner i forbindelse med korpusarbeidet.

REFERANSER

A. Korpustekster (tekstkode, forfatter, oversetter, originalets tittel, tittelen på oversettelsen)

ABR1 Brink, André (Malde, Per) *The Wall of the Plague / Pestens mur.*

AT1 Tyler, Anne (Roald, Bodil) *The Accidental Tourist / Tilfeldig turist.*

SG1 Grafton, Sue (Rogde, Isak) *"D" for Deadbeat / "D" for druknet.*

BV1 Vik, Bjørg (McDuff, David) *En håndfull lengsel / Out of Seasons and Other Stories.*

EHA1 Haslund, Ebba (Wilson, Barbara) *Det hendte ingenting / Nothing Happened.*

GS1 Staalesen, Gunnar (McDuff, David) *I mørkret er alle ulver grå / At Night All Wolves Are Grey.*

JM1 Michelet, Jon (Nations, Ellen) *Orions belte / Orion's Belt.*

KA1 Askildsen, Kjell (Lyngstad, Sverre) *En plutselig frigjørende tanke / A Sudden Liberating Thought.*

KF1 Faldbakken, Knut (Lyngstad, Sverre) *Adams dagbok / Adam's Diary.*

LSC1 Christensen, Lars Saabye (Nordby, Steven Michael) *Herman / Herman.*

B. Om engelsk-norsk parallellkorpus

Johansson, Stig & Knut Hofland. 1994. Towards an English-Norwegian parallel corpus. I Udo Fries, Gunnel Tottie & Peter Schneider (utg.), *Creating and Using English Language Corpora*. Amsterdam & Atlanta, GA: Rodopi. 25-37.

Johansson, Stig, Knut Hofland & Jarle Ebeling. 1996. Coding and aligning the English-Norwegian Parallel Corpus. I Karin Aijmer, Bengt Altenberg & Mats Johansson (utg.), *Languages in Contrast. Papers from a Symposium on Text-based Contrastive Studies in Lund, 4-5 March 1994*. Lund: Lund University Press. 87-112.

Johansson, Stig & Berit H. Løken. 1997. Some Norwegian discourse particles and their English correspondences. I Carl Bache & Alex Klinge (utg.), *Sounds, Structures and Senses. Essays Presented to Niels Davidsen-Nielsen on the Occasion of His Sixtieth Birthday*. Odense: Odense University Press. 149-170.

Johansson, Stig & Signe Oksefjell (utg.). Under trykking. *Corpora in Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam & Atlanta, GA: Rodopi.

Se videre referansene på prosjektets hjemmeside: http://www.hf.uio.no/iba/prosjekt/

Behovet av och marknaden för ordböcker mellan finska och övriga nordiska språk

ILSE CANTELL
WSOY, HELSINGFORS

Denna uppsats är ett försök att placera in ordböckerna och i detta fall uttryckligen tvåspråkiga ordböcker mellan finska och övriga nordiska språk i ett sociologiskt sammanhang. Jag har plockat fram ett antal sifferuppgifter som belyser situationen inom affärslivet, det nordiska samarbetet på politisk och administrativ nivå, undervisningen, flyttningsrörelsen och turismen. Jag utgår ifrån att detta är de sektorer inom samhällslivet som kan anses vara centrala då språk kommer i kontakt med varandra. Några direkta slutsatser kan man naturligtvis inte dra av de siffror jag fått fram, men siffrorna visar i alla fall vad som är centralt ur finländsk synvinkel och var finskan kan tänkas vara intressant.

Tvåspråkiga ordböcker skiljer sig ju från enspråkiga ordböcker i det att de i mycket högre grad än de enspråkiga är kommersiella produkter. De har sitt existensberättigande i ett behov och en marknad.

SVENSKAN OCH SVENSKA ORDBÖCKER – ETT SÄRFALL I FINLAND

Finland avviker sig från de övriga nordiska länderna på det sättet att vi har ett nordiskt språk, dvs. svenskan som i sin användning som främmande språk och också på ordboksmarknaden kan ta upp konkurrensen med engelskan. Svenska språk är ett obligatoriskt undervisningsämne i alla skolor men dessutom är det 295 000 finländares modersmål.

Finska medborgare med svenska som modersmål har vissa rättigheter som garanteras i grundlagen. Dessa rättigheter preciseras i språklagen från 1922. Finska medborgare skall ha rätt att använda

någotdera av landets nationalspråk, finska eller svenska, vid domstolar, övriga statsmyndigheter samt i kommuner, kommunalförbund och andra självstyrelseområden[1].

STOR MARKNAD I FINLAND

Detta innebär att finsk-svenska och förstås också svensk-finska ordböcker har en stor marknad i Finland. Svenskan är ett viktigt språk för alla myndigheter. Dessutom behöver också de svenskspråkiga medborgarna översättningshjälp mellan sitt modersmål och majoritetsspråket. – Detta har förstås följder också för innehållet i ordböckerna. Den officiella och juridiska terminologin är ett oumbärligt inslag i våra ordböcker.

Stora finsk-svenska ordboken var länge väntad då den äntligen kom ut i maj 1997. Trots det var det en överraskning för förlaget att första upplagan på några tusen exemplar såldes slut på några månader. Ingen annan storordbok har i Finland någonsin sålt så bra på så kort tid. Andra upplagan togs i november 1997.

De nu befintliga allmänspråkliga ordböckerna mellan finska och svenska är följande:

Cannelin, Knut & Cannelin, Aulis & Hirvensalo, Lauri & Hedlund, Nils 1986 (1976): *Suomi-ruotsi suursanakirja. Finsk-svensk storordbok.* Helsingfors: WSOY.

Cantell, Ilse & Martola, Nina & Romppanen, Birgitta & Sundström, Mats-Peter 1995: *Suomi-ruotsi opiskelusanakirja. (Finsk-svensk studieordbok.)* Helsingfors: WSOY.

Cantell, Ilse & Martola, Nina & Romppanen, Birgitta & Sundström, Mats-Peter 1997: *Suuri suomi–ruotsi-sanakirja. Stora finsk-svenska ordboken.* Helsingfors: WSOY

[1]Språklag 1.6.1922/148 1§ (reviderad 10.1.1975/10)

Collin, Anders & Streng, Tauno 1989: *Uusi ruotsi–suomi sanakirja. (Nya svensk-finska ordboken.)* Helsingfors: Kustannusosakeyhtiö Otava.

Iso ruotsalais-suomalainen sanakirja. Stora svensk-finska ordboken (I–III) 1982–87. Huvudredaktör: Göran Karlsson. Suomalaisen Kirjallisuuden Seuran toimituksia 358. Helsingfors: Suomalaisen Kirjallisuuden Seura.

Sarantola, Anja & Sarantola, Tauno 1995: *Ruotsi-suomi opiskelusanakirja. (Svensk-finsk studieordbok.)* Helsingfors: WSOY.

Köykkä, Lea & Saanila, Marianne & Saari, Marianne & Tirkkonen, Kirsti & Viljanen, Kari 1991: *Suomi/ ruotsi / suomi -sanakirja. (Finsk-svensk-finsk ordbok.)* Jyväskylä: Gummerus.

FINSKAN I NORDEN OCH NORDEN I FINLAND

I detta avsnitt går jag över till att betrakta finskans kontakter med de övriga nordiska språken. Här behandlas svenskan inte längre som vårt andra nationalspråk, utan som ett främmande språk bland andra främmande språk. Det faktum att svenskan också i detta sammanhang visar sig vara ett särfall är en sak för sig men bidrar naturligtvis också till att svenskan har en så betydande ställning på den finländska ordboksmarknaden.

NYETABLERINGAR TACK VARE EU

Inom affärslivet etablerar sig som bäst ett stort antal företag i Finland. Som ett exempel på nyetablerade företag i Finland under de senaste åren kan jag nämna några svenska och danska banker. I Helsingfors finns Handelsbanken och SE-banken, Den danske bank med fem anställda danskar och Unibank likaså med fem danskar.

En nyhet inom bankvärlden är samgåendet av den finländska banken Merita och svenska Nordbanken i slutet av 1997. Den nya

banken heter MeritaNordbanken och har två hemorter, Helsingfors och Stockholm. Inom denna bank kommer svenskan att vara arbetsspråket, något som har traditioner inom Merita. (Merita är resultatet av ett tidigare samgående av två finländska banker, den finspråkiga Kansallis-Osake-Pankki och Föreningsbanken i Finland, tidigare Nordiska föreningsbanken, som mycket länge uppfattades som en svenskspråkig bank.) Men de svenska direktörerna inom Nordbanken har också uttryckt sitt intresse för att lära sig finska!

MIGRATIONEN

Den nordiska arbetsmarknaden har varit öppen för nordiska medborgare sedan 1954. Finland har under denna tid varit och är fortfarande ett land med dominerande utvandring. Det land som finländarna i de allra flesta fallen flyttar till är Sverige. Det är också mest svenskar bland de inflyttade nordborna i Finland. Tabellen nedan visar dock att Sveriges ställning som invandringsland för finländare brutits något under 90-talet och att Norge blivit allt mera lockande för finnar.

1997: Finska medborgare bosatta i Norden[2]:

Danmark	2 102	
Island	92	
Norge	3 884	(1991: 3 051)
Sverige	103 091	(1990: 123 867)

1997: Nordiska medborgare bosatta i Finland och i Sverige[3]:

	i Finland	i Sverige
DK	482	25 983
FIN		103 091
IS	107	4 709
N	513	31 669
S	7 291	

[2] Källa: Nordisk statistisk årsbok 1997
[3] Källa: Nordisk statistisk årsbok 1997

TURISMEN

Tabellen nedan visar samma tendens som tabellerna ovan: kontakterna med Sverige och svenskarna är dominerande. Då det är fråga om "vanlig" turism kommer först Norge långt efter Sverige, först sedan Danmark och Island.

Nordbor i Finland[4]:

	1995	1996
DK	85 000	82 000
IS	7 000	7 000
N	141 000	138 000
S	485 000	541 000

Finländare i Norden[5]:

	1995	1996
Da	80 700	381 000
Isl	–[6]	23 000
No	89 865	506 000
Se	332 229	1 192 000

Som mål för arbetsresor är Danmark viktigare än Norge. Detta har säkert delvis sin förklaring i att Danmark är ett EU-land, delvis i att man via Danmark kan resa vidare till Europa.

Arbetsresor från Finland[7]:

Da	25 000
No	17 000
Se	151 000

[4] Källa: Nordisk statistisk årsbok 1997
[5] Källa: Nordisk statistisk årsbok 1997
[6] Siffran saknas.
[7] Källa: Nordisk statistisk årsbok 1997

UNIVERSITETSSTUDIER

Alla nordiska språk kan studeras vid finländska universitet. Här är förstås också svenskan dominerande, men siffrorna nedan visar att de andra nordiska språken inte är så främmande och exotiska för finländska språkstuderande som man kunde tro. En stor del tar kurser i danska, isländska och norska bara som ett tillägg till studierna i svenska, men för en del blir något av dessa huvudspråket.

Både från Helsingfors universitet och de två universiteten i Åbo rapporterades att intresset för norskan är växande. Speciellt efterlyses praktiska kurser. Den norska arbetsmarknaden drar fortfarande.

Nordiska språk vid finländska universitet
Kursdeltagare / år:

	da	no	isl
Helsingfors[8]	60	48	70
Tammerfors[9]	26	26	26
Åbo Akademi[10]			
Åbo universitet[11]	80	80	30[12]
Joensuu	30	30	<15[13]
Uleåborg	<30	<40	<30
Vasa[14]	35	35	10
Jyväskylä[15]	20	40	

BEFINTLIGA ORDBÖCKER OCH EVENTUELLT BEHOV AV NYA

Ovan gavs en förteckning över de ordböcker som nu finns på marknaden mellan finska och svenska. Dessa ordböcker har en dubbel funktion. Dels fungerar de inom vårt land vid kommunikationen mellan våra två nationalspråk, dels på internationell nivå vid

[8] Lektorer i danska, norska och isländska
[9] Norska och danska kan läsas som huvudspråk
[10] Svenskspråkigt universitet, lektor i danska
[11] Finskspråkigt universitet, lektor i norska
[12] Kurs ordnas ibland
[13] Fornisländska
[14] Lektor i norska
[15] Lektor i danska

kommunikationen mellan finländare och övriga nordbor, framför allt svenskar. I nordiskt samarbete är svenskan vårt främsta lingua franca, bara i några få fall har norskan eller danskan denna funktion. Detta och det faktum att de ovan beskrivna kontakterna mellan Finland och Norden i så hög grad domineras av kontakterna mellan Finland och Sverige återspeglas naturligtvis i den marknad de övriga ordböckerna kan få i Finland.

Nedan ges en förteckning över de ordböcker som nu finns på marknaden mellan finska och norska respektive finska och danska:

Farbregd, Turid & Seppinen, Hannele (1993): *Finsk-norsk ordbok*. Oslo: Universitetsforlaget.

Farbregd, Turid & Kämäräinen, Aili (1996): *Suomi–norja–suomi-taskusanakirja. (Finsk-norsk-finsk fickordbok.)* Helsingfors: WSOY.

Kroman, Sirkka (1991): *Suomi–tanska–suomi-taskusanakirja. (Finsk-dansk-finsk fickordbok.)* Helsingfors: WSOY.

Nuutinen, Olli (1991): *Tanskalais-suomalainen sanakirja. Dansk-finsk ordbog.* Loimaa: Oy Finn-Lectura Ab.

Azeem, Mirja (1993): *Tanska–suomi-sanakirja*. Helsingfors: Otava. (Dansk upplaga 1993: *Dansk-finsk ordbog*. København: Munksgaard.)

Pajuoja, Reijo (1993): *Suomi–tanska-sanakirja. (Dansk-finsk ordbok.)* Anjalankoski. (Eget förlag.)

Av dessa har den norska fickordbokens första upplaga sålts slut på två år. En andra reviderad upplaga kommer ut under våren 1998. En mellanstor finsk-norsk och norsk-finsk ordbok är under arbete. Av den danska fickordboken togs en andra upplaga under hösten 1997.

Det finns tyvärr tills vidare ingen ordbok mellan finska och isländska. Detta har betraktats som en liten kulturskandal. Till all lycka

är en sådan under arbete. Marknaden för den kommer inte att vara stor, men utgivningen av den kan betraktas som en kulturbragd.

Ordböcker mellan våra små nordiska språk är fortfarande också ett arbete för idealister. Här ges ordböcker till och med ut på eget förlag!

LITTERATUR

Finlands lag. (1997) Helsingfors: Juristförbundets förlag.

Nordisk statistisk årsbok 1997. (1997) Köpenhamn: Nordiska ministerrådet.