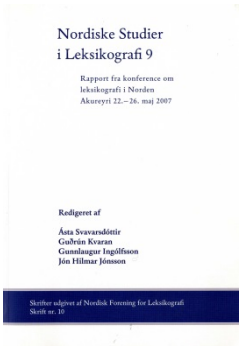
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Plan for leksikalsk dokumentasjon av moderne bokmål

Forfatter: Ruth Vatvedt Fjeld

Kilde: Nordiska Studier i Lexikografi 9, 2008, s. 131-142
Rapport fra Konference om leksikografi i Norden, Akureyri 22.-26. maj 2007

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Ruth Vatvedt Fjeld

Plan for leksikalsk dokumentasjon av moderne bokmål

The lexicon of the modern Norwegian bokmål standard needs a better description and documentation than what is the situation today. The article presents a plan for building a modern lexical database based on a balanced corpus of 40 million words of modern bokmål. This base should serve as a source for a traditional scientific dictionary as well as a dictionary for language technological applications.

1. Bakgrunn

Ordforrådet i moderne norsk bokmål er ikke godt dokumentert. Det nynorske ordforrådet blir godt dokumentert, både det historiske og det samtidige, i prosjektet *Norsk Ordbok: Ordbok over det norske folkemålet og det nynorske skriftmålet*, med Alf Hellevik som hovedredaktør for band I og II, fra band III Lars S. Vikør, fra band VI utvidet med Oddrun Grønvik og Laurits Killingbergtrø. Første band i prosjektet ble publisert i 1966, og planen er at siste band skal komme til det norske tohundreårsjubileumet i 2014. Med over tretti redaktører og god finansiering er det sannsynlig at det målet nås. Begrunnelsen for et så stort prosjekt for denne minoritetsvarieteten av norsk språk er blant annet at nynorsk må dokumenteres ut fra tekster og språkbruk der nynorsk er brukt (Skard 1932, Hageberg 1992). Tidligere er det gitt ut flere ordbøker over nynorsk, først og fremst Ivar Aasens ordbøker fra 1850 og 1873, senere utvidet og oppdatert med ordbøkene til Hans Ross (1895) og Steinar Schjøtt (1914), som var tillegg til Aasens verk. Disse ordbøkene danner grunnstammen i Norsk Ordbok, sammen med et stort innsamlet dialektmateriale og et moderne elektronisk korpus.

Ordbokstradisjonen for bokmålet er magrere, og det er synd, siden det brukes av ca. 85-90 prosent av alle nordmenn ifølge Språkrådets beregninger. Bakgrunnen er bl.a. at bokmål/riksmål lenge var omtrent det samme som dansk, og det danske ordforrådet var godt dokumentert i *Ordbog over det Danske Sprog* og andre ordbøker. En av de første ordbøkene som forsøker å dokumentere norsk bokmål, er skrevet av Knud Knudsen, han som også kalles bokmålets far. Ordboka *Unorsk og norsk, eller Fremmedords Afløsning* fra 1881 er imidlertid en svært spesiell ordbok, der puristen Knudsen selv lager mange nye norske ord som avløsere for internasjonale ord og også for tyske lån i norsk. Svært få av disse ordene er kommet i vanlig bruk, og ordboka er nærmest å regne som en kuriositet i dag.

Norsk riksmål og offisielt bokmål skilte lag i 1939 som følge av den offentlige

språkrøktens mål om å føre bokmål og nynorsk sammen i ett skriftspråk. Mange motsatte seg en slik styring av språket og lanserte *riksmål* som en uoffisiell variant ved siden av det offisielle bokmålet. Riksmål er en konservativ skriftnorm fastlagt av Det Norske Akademi for Sprog og litteratur, og det er den som leksikografisk sett er best dokumentert. Først med *Norsk Riksmålsordbok* av Trygve Knudsen, Alf Sommerfeldt og Harald Noreng m.fl., som kom ut i perioden 1937–57, fikk vi en egen norsk allmennordbok som ikke dokumenterte nynorsk. Ordboka dokumenterer først og fremst det som nå kalles riksmålstradisjonen slik den er anvendt hos de kjente norske dikterne som brukte riksmål. Men også en del avisspråk og annet skriftspråk ble dokumentert. Men det er helt klart en selektiv og normativ ordbok der store deler av det offisielle bokmålet ble utelatt. Den er utgitt av Riksmålsvernet. Nesten førti år seinere kom Harald Norengs to tilleggsbind, *Norsk riksmålsordbok, tilleggsbind I og II* (1995), som i seg selv er et storverk laget av én mann alene, og som dokumenterer det moderne riksmålet noe bredere og mindre normativt enn den tradisjonelle riksmålsstandarden. Men de ble gitt ut av det språkideologiske Norsk Akademi for Sprog og Litteratur, og lemma- og redigeringsspråket er derfor riksmål.

I mellomtiden hadde Universitetet i Oslo i 1986 gitt ut de to håndordbøkene *Bokmålsordboka* og *Nynorskordboka* i samarbeid med Norsk språkråd. De er gode dokumentasjoner av det sentrale ordforrådet i moderne bokmål og nynorsk, men dekker selvsagt ikke hele ordforrådet. De er klart normative.

2. Behov

Ordforrådet i et språk er i stadig utvikling, og det trengs kontinuerlig oppdatert utforsking og dokumentasjon av ordenes betydning, bøying, kombinasjonsegenskaper og annen bruk. En forskningsbasert ordboksbase kan både fylle behovet for ordkunnskap som understøtter kommunikasjon mellom brukerne, og dokumentere den faktiske utviklingen og bruken, slik at man har et faglig velfundert utgangspunkt for den normering og språkplanlegging som foretas av Språkrådet, samt gi grunnlag for annen språkforskning innen semantikk, morfologi og syntaks. Særlig er det interessant for et språk der normen gir stor valgfrihet, slik som i norsk, da man ikke vet noe systematisk om hvordan denne vide normen utnyttes. I tillegg til Norsk Ordbok trenger vi i Norge derfor en vitenskapelig basert dokumentasjon av hele ordforrådet i moderne bokmål, både slik det er normert av Språkrådet og vedtatt av Kulturdepartementet, og slik det faktisk brukes både i offentlig sammenheng og i privat språk som ikke er underlagt den offisielle normen, bygd på observasjon av faktisk språkbruk. Språkrådets melding "Norsk i hundre!" fra 2005 sier da også: "Norge mangler store ordbøker over samtidsspråket som tilsvarer de man har i de andre nordiske landene" (2005:145).

I tillegg er det et dokumentert behov for leksikalsk folkeopplysning om normert bokmål. Det vet vi blant annet fordi det i gjennomsnitt hver dag er ca. 25 000

søk i Bokmålsordboka på Internett. En enda bedre og mer utfyllende ordbok ville rimeligvis bli brukt av enda flere.

Videre er det et stort behov for å utvikle norsk datamaskinell leksikografi. For automatisk språkanalyse er det nødvendig med et veldefinert og automatisk lesbart leksikon. Det er blant annet nødvendig for automatisk syntaktisk analyse og automatisk ordklassemerking av løpende tekster. Maskinlesbare ordbøker brukes i utvikling av automatiske oversettingsprogram og flerspråklig informasjonssøking. Det er viktig å sikre norsk som nasjonalspråk ved å delta i den internasjonale utvikling av oversetting av mellomstatlig regelverk og andre språkfestede standarder. Men det krever veldefinerte, maskinlesbare leksikon, og det er nødvendig at en utforsking av norsk ordforråd tilpasses internasjonale språkteknologiske produkter. Systematisk gjennomgang av ordforrådet for en større ordbok gjør det mulig samtidig å trekke ut informasjon for en språkteknologisk leksikalsk beskrivelse uten store merkostnader. Arbeidet med de to ordboksproduktene vil gi en synergieffekt. EU-prosjektet SIMPLE (Semantic Information in Plurilingual Lexicons, jf. Lenci et. al. 2000) er et internasjonalt prosjekt for utvikling av et datamaskinlesbart leksikon over 12 europeiske språk til bruk i språkteknologi. En norsk variant av det danske SIMPLE-leksikonet er utviklet som et forsøksprosjekt, og et slikt leksikon kan videreutvikles og utvides i samsvar med den internasjonale forskningen innenfor LBD-prosjektet (Leksikografisk bokmålsdatabase). SIMPLE-leksikonet bygger på Pustejovskys semantiske modell for leksikalsk beskrivelse av nominale ledd ved hjelp av kvaliastruktur (Pustejovsky 1995) og Levins semantiske klassifisering av verb (Levin 1993). Disse teoriene er videreutviklet og tilpasset beskrivelse innenfor SPINN-nettverket (Pedersen, B, R.V.Fjeld, M.Toporowska Gronostaj 2002). Denne leksikalske beskrivelsen er imidlertid fortsatt på utprøvingsstadiet, da mer fullstendig dekning av det norske ordforrådet krever større ressurser en det som er tilgjengelig.

3. Ressurser

Det er utviklet en rekke bokmålsressurser for leksikografisk utnytting og videre forskning:

Ekserptsamling og elektronisk tekstsamling:

Store deler av den norske litterære arven fra perioden 1550–1950 er dokumentert i ekserptsamlingen for det ufullførte prosjektet *Det norske litterære ordboksverk*. Deler av dette er gjort elektronisk tilgjengelig ved at man i Dokumentasjonsprosjektet skannet inn mange av de samme verkene, i tillegg til ikke-ekserperte verk fra samme periode. Det elektroniske materialet utgjør 46 840 sider tekst, og er tilgjengelig på http://www.dokpro.uio.no/litteratur/.

Nyordsmaterialet

Nyordsmaterialet er en elektronisk base av ekserpter fra norske aviser for perioden 1972–2000. Ekserptene er kodet med ordklasse og med en rekke tilleggskoder for etymologiske, syntaktiske og semantiske egenskaper. Det består av over 500 000 ekserpter og er tilgjengelig på http://www.dokpro.uio.no/bokmaal/nyord/nyord_ramme.html.

Bokmålsordboka

Bokmålsordboka er den største norske definisjonsordboka som også dokumenterer hele den offisielle bokmålsmorfologien. Den blir stadig oppdatert med nye ord hentet fra moderne avistekster og annet materiale, og formverket er justert i forhold til de siste vedtak som er godkjent av Kirke- og kulturdepartementet. Den er tilgjengelig på http://www.dokpro.uio.no/ordboeker.html.

Nyord i norsk

Dette er en ordbok som følger opp *Nyord i norsk 1945–75* med nyord fra perioden 1975-2003. Den bygger særlig på innholdet i basen *Nyordsmaterialet*. Ordboka er under utgivelse og er et samarbeidsprosjekt med Språkrådet i Norge.

Leksikalsk bokmålskorpus (LBK)

LBK er det første balanserte tekstkorpus for moderne norsk og består av litterære tekster og sakprosa, samt normert og unormert litteratur fra perioden 1985 til i dag. Korpuset er søkbart i en database etter grunnord/lemma (leksikaliserte ord eller uttrykk), og gir opplysninger om forfatter, teksttype, kontekst og ordenes grammatiske egenskaper. Korpuset inneholder 40 mill. ord. Det er balansert slik:

- Periodika 20%
- Sakprosa 45 %
- Skjønnlitt 25%
- TV-tekst 5%
- Upublisert 5%

Disse kategoriene er inndelt i spesifiserte teksttyper som også holdes best mulig balansert. Beregningsgrunnlaget for balanseringen er hentet fra Norsk Mediebarometer, som angir folks lesevaner fordelt omtrent slik:

- Internett 35%
- Bøker 15%
- Avis 40%
- Tidsskrift 5%
- Tegneserie 5%

Korpuset skal dermed dekke omtrent det en gjennomsnittsnordmann møter av ord i vår tid.

Norsk ordbank

Norsk ordbank er utviklet i samarbeid med Språkrådet, Dokumentasjonsprosjektet/EDD og Tekstlaboratoriet ved Humanistisk Fakultet. Det er en leksikalsk database der lemmalisten med ortografi og morfologi fra Bokmålsordboka utgjorde grunnstammen og seinere ble utvidet, bl.a. med ordlister utarbeidet av IBM (jf. Engh 1992). Lemmaene er tillagt formalismer slik at ordbanken i dag framstår som en maskinleselig base over de aller fleste norske bokmålsordene. Alle ordene har formalisert beskrivelse av morfologiske og ortografiske egenskaper.

Norsk aviskorpus, bokmålsdelen

Norsk aviskorpus er det største tekstkorpuset for norsk. Det ble påbegynt i 1998 og består av tekstmateriale fra nettutgavene av utvalgte riks- og regionaviser. Det svært omfattende materialet har i 2008 passert 500 millioner ord og er det største i sitt slag i Norge. I prosjektet bygges det også opp en nyordsdatabase som vokser i omfang hver dag. Dagsaktuelle nyordslister klassifiseres automatisk og gjøres tilgjengelige for forskere og leksikografer. Hjemmeside: http://avis.uib.no/

NoTa- Oslo (Norsk talespråkskorpus, Oslo-delen)

Ved Tekstlaboratoriet ved Universitetet i Oslo er det utarbeidet et talespråkskorpus for flerbruk. Dette er landets største og best merkede korpus for moderne norsk talemål ca. 2005. Oslo-delen inneholder ca. 1 mill. ord som er ortografisk transkribert med lenker til video- og lydfiler. Ordene er grammatisk tagget. Dette korpuset er et verdifullt nytt grunnlag for nærmere utforsking av bokmålets leksikon. Korpuset ligger på: http://www.tekstlab.uio.no/nota/

Den Danske Ordbog (DDO)

Den Danske Ordbog er en ordbok i seks bind over det danske nåtidsspråket (utgitt 2003–05). Ordboken gir en grundig beskrivelse av ordforrådet i moderne dansk i perioden fra ca. 1950 frem til i dag og er en etterfølger til den store *Ordbog over det danske Sprog*. Den Danske Ordbog inneholder systematiserte opplysninger om oppslagsordene og bygger på et moderne dansk korpus. Ordboken regnes som den mest nyskapende og fullstendige beskrivelsen av et skandinavisk språk. Det norske og det danske språket har så mye felles at denne ordboken fungerer som et nyttig sammenlikningsgrunnlag og som en idéskaper for et prosjekt for moderne norsk bokmål.

Av ressurser som er under utvikling, er følgende særlig relevant for et prosjekt som LBD:

Datamaskinelt lesbart leksikon

På basis av det danske SIMPLE-leksikonet er det gjort forsøk med en norsk parallellkopi ved å legge inn norske ekvivalenter for de danske oppslagene, såkalte SemU-er, som tilsvarer delbetydninger, og legge til brukseksempler fra norske tekster. Alle oppslagsordene i SIMPLE er koblet til samme lemmas delbetydning i Bokmålsordboka i en database, slik at den norske morfologien og definisjonene der kan hentes inn for hvert enkelt oppslag. Ordene i leksikonet er markert med fire ontologiske strukturer (jf. Pedersen 2005 og Pedersen & al. 2002). Dette skal gjøre at en datamaskin med ganske stor nøyaktighet kan koble f.eks. et engelsk eller spansk ord til ekvivalenten i et av de skandinaviske språkene. Likeledes kan den koble rett ekvivalent mellom de skandinaviske språkene. Arbeidet har ført til en grundig ekvivalensanalyse mellom norsk og dansk for de ferdige postene, og viser falske venner mellom de to språkene som neppe har vært registrert tidligere.

Å skrive slike ordboksartikler kan være tidkrevende. Det forutsetter kunnskap om semantiske strukturer og plassering i ontologisk hierarki som ikke er utforsket før. Dessuten må alle definisjoner ha et stramt strukturert format som gjør det nødvendig å ta inn informasjon som for menneskelige lesere er "selvfølgeligheter", f.eks. at en rose er en blomst, og at en blomst er en plante. Fullføring av dette prosjektet forutsetter mer kunnskap om det norske ordforrådet og utvikling av et formalisert beskrivelsesapparat. Et språkteknologisk leksikon vil gi grunnlag for en systematisk og enhetlig beskrivelse av det norske bokmålets semantikk, som er en forutsetning for de fleste former for moderne språkteknologi.

Hittil har det norske leksikonet 13 434 innganger som er lenket til danske og svenske oppslag. I Danmark har man investert i en mer fullstendig språkteknologisk ordbok STO (http://cst.dk/cgi-bin/defisto/defisto), som kan vise seg å være enda bedre egnet som modell for en norsk språkteknologisk ordbok.

Norsk Ordbok 2014

Det leksikografiske fagmiljøet ved Universitetet i Oslo har fått et betydelig løft i og med særbevilgningene til Norsk Ordbok 2014. Mye av det arbeidet som legges ned i dette prosjektet, kan gjenbrukes i beskrivelsen av bokmålet, og et større bokmålsprosjekt kan gi synergieffekt tilbake til Norsk Ordbok. Samarbeidet med redaksjonen i Norsk Ordbok har hele tiden vært godt, og det er tjenlig å fortsette dette samarbeidet. Det er en klar fordel at ordforrådet i nynorsk og bokmål utforskes og dokumenteres ved samme enhet, slik at fagmiljøet i dokumenterende leksikografi er samlokalisert og kan samarbeide mest mulig rasjonelt.

Nationalencyklopedins ordbok (NEO) og Svensk ordbok utgiven av Svenska Akademien

Ved Språkdata i Göteborg ble Nationalencyklopedins ordbok utviklet i perioden

1995–96. Med utgangspunkt i den arbeides det videre for å utvikle Svensk ordbok utgiven av Svenska Akademien. Den skal bygge på et egenutviklet korpus, men den har også historisk informasjon med belegg helt tilbake til runetiden. Prosjektet skal ha mer systematisk angivelse av kollokasjoner og idiomer og fullstendige valensangivelser. Dette arbeidet vil ha stor relevans for utarbeiding av en omfattende ordbok for norsk bokmål.

4. Anvendelse

Den leksikalske databasen LBD skal danne grunnlag for forskjellige ordbøker, i første omgang er det planlagt en stor elektronisk ordbok som dokumenterer moderne bokmål, og en språkteknologisk ordbok.

Norsk elektronisk bokmålsordbok (NEBO)

NEBO skal gi utfyllende informasjon om norske ords ortografi, bøying og betydning slik det er normert, og en publisering av Språkrådets vedtak om den offisielle normen. Målgruppen er journalister, forfattere, oversettere, lærere og andre språkinteresserte. Ordboka skal være normativ og til en viss grad deskriptiv, slik at den gjør rede for mye brukte former som ligger utenfor normen, med henvisning til de normerte formene. NEBO skal også dokumentere faste uttrykk eller flerordslemmaer (leksikaliserte kollokasjoner), noe som mangler i norsk leksikografi, men som man har utforsket og dokumentert i de fleste andre språk.

Språkteknologisk bokmålsordbok (STEBO)

STEBO vil bli en enspråklig maskinlesbar ordbok med formalisert beskrivelse av de semantiske relasjonene mellom ordene, samt de syntaktiske egenskaper enkeltordene har. En språkteknologisk ordbok med god dekning av hele ordforrådet er hittil ikke utviklet for norsk. Som enspråklig base har den mange anvendelsesmuligheter:

- Til informasjonsfinning ved hjelp av søkemotorer er en ontologisk ordnet ordbase viktig.
- I dataprogrammer som genererer naturlig språk er en maskinlesbar leksikalsk komponent uunnværlig. Det gjelder særlig for taleteknologiske programmer der ordbøker med variantuttaler er maskinlesbart tilgjengelig.
- I dataprogrammer som lager skriveveiledning er det nødvendig med veldefinert leksikon.
- I dataprogrammer for T9-språk i mobiltelefoner er frekvensberegnet leksikon en forutsetning. De eksisterende leksikon har for dårlig kunnskap om frekvens i norsk, og kan forbedres betraktelig med mer nøyaktig kunnskap om bruken av norske ord.
- I rettskrivingsprogrammer (allmenne rettskrivingsprogrammer og spesielt

tilrettelagte for lese- og skrivesvake) trengs kunnskap om vanlige variantformer/feil i rettskriving og bøying for å gi veiledning om rett form.

Andre ordbøker

Ut fra en utfyllende og godt organisert leksikalsk database der alle informasjonstyper er sortert og merket, kan man lage en lang rekke mer spesialiserte ordbøker, f.eks. rettskrivingsordbok, fremmedordbok, lånordbok, skoleordbok, barneordbok, slangordbok m.v.

En leksikalsk database vil også være et viktig grunnlag for tospråklige ordbøker der bokmål er det ene språket. Databasen er dessuten viktig som grunnlag for annen leksikalsk forskning, både innen morfologi, orddanning, syntaks og semantikk. Det vil særlig gjelde revisjonen av *Norsk referansegrammatikk (NRG)*. Det vil gi en ny mulighet til innsikt i leksikalske og syntaktiske beskrivelser av moderne norsk språk.

5. Prosjektplan

Ontologi

Som et første skritt i selve ordboksarbeidet skal det genereres en ontologi ut fra Bokmålsordboka. En ontologi er et begrepssystem der alle ordene er ordnet etter hvordan det som ordene betegner (referentene), forholder seg til hverandre. Innenfor SIMPLE-prosjektet er ordene allerede ordnet ontologisk, og til en viss grad kan det bygges videre på den. Det er dessuten gjort forsøk på automatisk generering av en ontologi for norsk bokmål (DOXA, Nygaard 2006) som et ordnett ekstrahert ut fra ordklassetagging av definisjonstekstene i Bokmålsordboka. Dette programmet skal prøves ut i full skala for hele lemmautvalget i Bokmålsordboka. En ontologi som er produsert automatisk, må etterprøves manuelt, og det gjelder også jamføringen med SIMPLE-ontologien, men dette er et nødvendig arbeid for at ordboksarbeidet skal kunne gjennomføres systematisk og semasiologisk og ikke i den tilfeldige og alfabetiserte onomasiologiske rekkefølgen. En videre utbygging av ontologien skal også jamføres med det danske OntoQuery-prosjektet (http://www.ontoquery.dk/index.php), slik at den norske beskrivelsen harmonerer med internasjonale beskrivelser.

Ordboksbasis: Leksikografisk bokmålskorpus LBK

LBK er et balansert korpus over bokmålstekster fra 1985 til nåtiden. Den inneholder 40 mill. ord, og hvert ord er merket med ordklasse av en automatisk ordklassetagger. Tekstene er manuelt merket med teksttype, sjanger, domenene og forfatter, samt sosiologiske variabler knyttet til forfatter. Korpuset vil således gi en god oversikt over ordbruken i forskjellige sammenhenger og gi grunnlag for en utfyllende dokumentasjon av hvordan normen anvendes eller brytes. Dermed

vil den leksikalske databasen også være godt grunnlag for normeringsarbeid og for språksosiologisk forskning.

I tillegg til LBK-korpuset skal talespråkskorpuset NoTa, Norsk aviskorpus AVIT, Norsk riksmålsordbok og Norsk Ordbok brukes som ordboksbasis for dokumentasjon av bruken av norske ord og uttrykk.

Lemmautvalg

Grunnlaget for de ordene som skal defineres og beskrives i prosjektet, hentes fra Bokmålsordboka. Den har en ordboksbase som er stadig utvidet og oppdatert fra førsteutgaven i 1986, og er således den mest utfyllende beskrivelse vi har av det standardiserte bokmålet. Dette ordforrådet skal suppleres med de ordene som fins i LBK-korpuset og i andre kilder og korpus som er tilgjengelig, og som ikke er med i Bokmålsordboka. Ut fra korpuset lages det en konkordans (alfabetisert liste over alle ord som fins i tekstsamlingen). Ordlista i Bokmålsordboka kjøres mot konkordansen fra LBK, slik at vi får en liste over de leksikalske enhetene som fins i korpuset, men ikke i Bokmålsordboka, og omvendt. Det samme vil vi gjøre med andre tilgjengelige korpus, som NoTa og Norsk Aviskorpus. Disse lemmakandidatene blir undersøkt i forhold til sekundærkildene og eventuelt lagt inn i databasen. Utvelgelse for videre behandling og presentasjon i NEBO er bl.a. avhengig av bruksfrekvens og plass i ontologien. Alle lemmakandidatene skal i første omgang lemmatiseres. Etter nøyere overveielser kan grenseverdier og hapax legomenon strykes, men i prinsippet skal alt ordmateriale som forekommer i LBK-korpuset dokumenteres i den leksikalske databasen.

Med denne metoden får vi også se hvilke lemmaer som fins i Bokmålsordboka men ikke i korpus, og således kartlagt eventuelle lakuner i korpuset. I tillegg til lemmautvalget fra LBK skal det vurderes om det trengs et tilleggsutvalg fra Sekundærkildene, NoTa og Norsk Aviskorpus pga. åpenbare lakuner i korpuset.

Lemmabeskrivelse og informasjonstyper

Hvert enkelt lemma i lemmautvalget beskrives med utgangspunkt i definisjonen i Bokmålsordboka, sammenliknet med den i DDO og tilgjengelige ordbøker for norsk. I tillegg søkes eksempler fra LBK og fra suppleringsmaterialet i ordboksbasen.

Lemmabeskrivelsen skal bestå av artikkelhode, artikkelkropp og artikkelfot.

Artikkelhodet angir lemmategnform og id-nummer for SIMPLE-basen. Deretter følger forskjellige normerte former, samt unormerte, belagte former med kildehenvisninger. Det samme gjelder forskjellige bøyningsformer av den lemmatiserte formen, både normerte og unormerte.

Hver delbetydning skal lemmatiseres som eget oppslag, men slik at sammenhengen mellom delbetydningene går klart fram. Dette er nødvendig for den semantiske beskrivelsen og for at kopling til SIMPLE-beskrivelsene skal bli mulig.

Artikkelkroppen angir definisjonen fra Bokmålsordboka, som da blir delt opp som egne lemmaer for hver deldefinisjon, såkalte Semantic Units (SemU). Den første betydningen skal være grunnbetydningen, det vi kaller den bokstavelige eller konkrete betydningen, deretter følger de forskjellige metaforiske eller avledede betydningene etter et fast mønster. Delbetydningene fra DDO legges inn der det er ekvivalens mellom dansk og norsk for å sikre konvergens med SIMPLE-definisjonene. Alle SemU-ene plasseres i et system som svarer til den ontologiske beskrivelsen med samling av kohyponymer under felles hyperonym. I tillegg angis antonym, synonym og meronym. Videre skal argumentstruktur og kvaliastruktur angis, noe som i praksis ofte vil bli en formalisert gjengivelse av den tradisjonelle delen av artikkelkroppen. For flerordslemmaer angis en henvisning til ettordslemma og kollokasjonell betydning, for øvrig gis de samme informasjonstyper som ved ettordslemmaene.

I artikkelfoten angis lemmaets avledningsmuligheter, leksikaliserte sammensetninger det kan inngå i, og etymologisk informasjon.

Redigeringsverktøy:

Det fins flere redigeringsverktøy det kan være aktuelt å bruke.

Prosjektet Norsk Ordbok 2014 har sammen med EDD utarbeidet et redigeringsprogram som har vist seg effektivt for det prosjektet. Det er spesielt tilrettelagt for det dokumentasjonsarbeidet som gjøres med nynorsk skrift- og talemål, men det er åpenbart at mye av dette kan gjenbrukes i et bokmålsprosjekt.

I 2004 ble det utarbeidet et redigeringsverktøy og et program for lenking av forskjellige SemU-er i SIMPLE-leksikonet, dvs. en editor som gjør det forholdsvis enkelt å koble de skandinaviske språkene sammen. Det vil da vise seg hvilke delbetydninger/SemU-er for hvert flertydig lemma som er felles for de tre språkene, og det vil gi et godt utgangspunkt for videre utbygging for et mer anvendbart maskinlesbart semantisk leksikon for skandinaviske språk. I denne editoren er også Bokmålsordboka lagt inn for direkte utnyttelse i redigeringsarbeidet. Derved er sammenlikning og avdekking av falske venner mellom de nordiske språkene forholdsvis lett å få med. Den norskutviklede editoren har vakt interesse både i Norden og internasjonalt (Fjeld et. al. 2004).

6. Kvalifikasjoner

Leksikografimiljøet ved Institutt for lingvistiske og nordiske studier ved Universitetet i Oslo er det største i Norge og har gode faglige referanser. Miljøet er det eneste i landet som arbeider med teoretisk leksikografi og gir undervisning i emnet. Det er derfor her det er rimelig å forvente teoretisk innsikt og vilje til å gjennomføre et så stort leksikografisk prosjekt. Miljøet har et internasjonalt faglig nettverk og deltar i den generelle fagutviklingen. Ved samme institutt er

Tekstlaboratoriet, som har et godt datateknologisk fagmiljø for språkbehandling og vil være en viktig samarbeidspartner.

7. Publisering

LBD skal ligge gratis tilgjengelig på Internett for søk i ordbasen og i korpuset, jf. slik man har valgt for det danske *Ordnet.dk*. Det må utarbeides forskjellige grensesnitt for forskjellige brukergrupper og forskjellige behov.

Arbeidskraft til redigering og datamaskinell ekspertise

Omfanget av databasen er beregnet til ca. 65 000 ordartikler. Om man regner et snitt på tre arbeidstimer pr. oppslagsord, vil det bli 195 000 timeverk. Tidsbruken settes så lavt da man regner med at mye kan hentes fra definisjonene i Bokmålsordboka, DDO og andre ordbøker. Hovedarbeidet vil bestå i å samordne og systematisere tidligere beskrivelser og komplettere og oppdatere dem ut fra funn i korpuset. Den datamaskinelle beskrivelsen vil sannsynligvis gå raskt når man får rutine i å systematisere de nødvendige informasjonene ut fra de tradisjonelle definisjonene. Det trengs datamaskinell ekspertise for redigering og utvikling av databasen og for drift av databasene.

8. Litteratur

Engh, Jan 1992: Leksikografi i IBM Norge. I: Fjeld, R.V. (red.) *Nordiske studier i leksikografi.* Oslo: Nordisk forening for leksikografi, Skrift nr. 1, 409–422 .

Fjeld, Ruth Vatvedt, Wik, Preben and Nygaard, Lars 2004: Managing complex and multilingual lexical data with the Simple-editor. I: *The Eleventh EURALEX 2004Proceedings,* Bretagne, 693-69.

Hageberg, Arnbjørg 1992: Norsk Ordbok – i 1950 eit nybrotsverk, i 1991 ein anakronisme? I: R.V. Fjeld (red.). *Nordiske studier i leksikografi.* Oslo: Nordisk forening for leksikografi, Skrift nr. 1, 235–243.

Lenci, Alessi et. al. 2000: SIMPLE – A General Framework for the Development of Multilingual Lexicons. I: T. Fontenelle (ed.): *International Journal of Lexicography.* Vol 13, 249–263.

Levin, B. 1993: *English Verb Classes and Alternations. A Preliminary Invistigation.* Chicago

Norsk i hundre! Norsk som nasjonalspråk i globaliseringens tidsalder. Et forslag til en strategi 2005. Språkrådet.

Nygaard, Lars (2006): *Frå ordbok til ordnett. Cand.philol.-oppgåve, Institutt for lingvistiske og nordiske studium.* (http://wo.uio.no/as/WebObjects/theses.woa/wa/these?WORKID=42189)

Pedersen, Bolette, V.Fjeld, Ruth, Toporowska Gronostaj, Maria 2002: Harmonisering og sammenkædning af sprogteknologiske ordbaser med særligt henblik på informationssøgning – en rapport fra SPINN-netværket I: *Nordisk Sprogteknologi*. København: 2002, 233–256.

Pedersen, Bolette 2005: Datamatisk leksikografi i Norden – status og visioner. I: *Nordiske studiar i leksikografi 7,* Volda: Nordisk foreining for leksikografi, 302–314.
Pustejovsky, James 1995: *The Generative Lexicon.* Massachusetts: 1-298
Skard, Sigmund 1932: *Norsk Ordbok. Historie – Plan – Arbeidsskipnad.* Oslo

9. Omtalte ordbøker og ordboksprosjekt

Bokmålsordboka (BOB). 2006. Oslo
Den Danske Ordbog (DDO) 2003-2004. København
Knudsen, Knud. 1881: *Unorsk og norsk eller fremmedords avløsning.* Kristiania
Nationalencyklopedins ordbok (NEO). 1995-96. Göteborg
Norsk Ordbok (NO). 1966- . Oslo
Norsk Riksmålsordbok (NRO). 1937-1957. Oslo
Norsk Riksmålsordbok (NRO). 1995, Tilleggsbind I og II. Oslo
Ordbog over det danske Sprog (ODS). 1919-1954. København
Ross, Hans. 1895: *Norsk Ordbog.* 1895-1913. Christiania
Schjøtt, Steinar. 1914: *Norsk Ordbok.* Oslo
Aasen, Ivar 1850: *Ordbog over det norske folkesprog.* Christiania
Aasen, Ivar. 1873: *Norsk Ordbog med dansk forklaring.* Christiania

Ruth Vatvedt Fjeld
professor, f. 1948
ILN
Universitetet i Oslo
Boks 1001 Blindern
N-0315 Oslo
r.e.v.fjeld@iln.uio.no