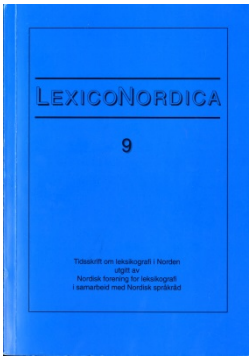
# LexicoNordica

Titel: Normering i klemme mellom språkteknologiske og pedagogiske ordbøker

Forfatter: Ruth Vatvedt Fjeld

Kilde: LexicoNordica 9, 2002, s. 131-148

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






**Søgbarhed**
Artiklerne i de ældre LexicoNordica (1-16) 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*

# Normering i klemme mellom språkteknologiske og pedagogiske ordbøker

With the development of dictionaries for language technology lexicography has to cope with a new dictionary typology. Descriptivity or prescriptivity has long been discussed in human-oriented dictionary making. This article tries to point out how language technology evokes a new discussion of this question, on the one side requiring a wider norm, on the other a tighter norm than in dictionaries written for pedagogical situations. This makes the reuse of existing dictionaries as the lexicon part in language technology programs nearly impossible.

## 1. Innledning

Språk er store og kompliserte tegnsystemer, og det er bare mindre deler som kan normeres. Siden språk også hele tiden er i forandring, vil enhver norm bare være midlertidig. Det er alltid en stor og bevegelig gråsone mellom det normerte og det unormerte i et språk, der noe diffunderes ut fra det normerte og over i det upleide eller unormerte, og selvsagt tas stadig noe nytt inn og blir normert. Man kan forestille seg dette som et stykke natur der noe er dyrket innmark mens mesteparten er udyrket utmark. Hele tiden brytes nytt land som legges under plogen og dyrkes og stelles, men samtidig er det områder som gror igjen og blir til utmark fordi de ikke er interessante lenger i den aktuelle kulturen. For leksikografer kan dette duge som et bilde på normeringsarbeid. Når man driver med språknormering og språkrådgivning, er det viktig å være klar over om en gir oppskrift for pleie av utmarka eller innmarka.

Tidligere ble ordbøker laget utelukkende for menneskelige brukere, slik at de kunne få informasjon om ordforråds form og innhold eller andre egenskaper ved det leksikalske inventaret i språk. Det var nyttig både for morsmål og fremmede språk. En viktig oppgave for mange ordbøker er å gjøre språkbrukerne kjent med hva som regnes som korrekt og hva som regnes som feil bruk av ord og uttrykk. Selve ordboksredigeringen ble regnet som et praktisk håndverk uten spesielle krav til teoretisk lingvistisk kompetanse. Det gjaldt å skrive ned det vi alle var enige om.

I leksikografien er det nå en ny situasjon. Med utviklingen av språkteknologien har leksikografi og leksikalsk beskrivelse blitt et sentralt felt i lingvistikken, de fleste språkteknologiske nyvinninger

forutsetter en systematisk og veldefinert leksikalsk komponent. Ny innsikt, særlig i morfologi og semantikk, har ført til utvikling av språkteknologiske ordbøker som viser at ordforrådet ikke er mer ubeskrivelig enn andre deler av språkvitenskapen. Det har både interne strukturer og relasjoner og oppviser regelmessigheter som ikke skiller seg mye fra reglene man finner i morfologi, syntaks og semantikk.

Etterspørselen etter språkteknologiske leksikalske beskrivelser kan imidlertid føre til løsninger som gir nye utfordringer for leksikografene, ikke minst overskrides lett grensene mellom det normerte og det unormerte fordi maskiner har andre lesemåter og kommunikasjonsbehov enn menneskelige ordboksbrukere.

## 2. Ordbokstypologi og normering

Normering i leksikografi har forskjellig innhold avhengig av hva slags leksikografisk arbeid det dreier seg om. Jeg vil derfor først se normering i forhold til forskjellige ordbokstyper.

### *2.1 Deskriptive ordbøker*

De første ordbøkene ble laget med et rent deskriptivt siktemål, de dokumenterte den ordbruken som faktisk forekom i et språk. Først på 1700-tallet fikk leksikografene for seg at de skulle vise eller bestemme hva som var godt og hva som var dårlig språk ved at noen former ble anbefalt, andre ble frarådd.

En deskriptiv ordbok har som mål å registrere og beskrive alle de leksikalske enhetene i et språk, systematisert og samlet under ett sett av lemmaer, der det angis alle de varianter dette kan ha med hensyn til fonologisk og morfologisk inventar. Videre angis de forskjellige semantiske og syntaktiske egenskaper det kan ha. Slike ordbøker prøver å dokumentere alt mennesker kan finne på å uttrykke i et språk.

Det fins få eller ingen fullstendig deskriptive ordbøker, ikke minst fordi det er en umulig oppgave å registrere alt det som befinner seg i språkets utmarker. Leksikografer som hevder å være rent deskriptive, lurer sannsynligvis seg selv. Men det fins ordbøker med forskjellig grad av deskriptivitet. Ofte er hensikten med deskriptive ordbøker å gi et så godt grunnlag som mulig for å utarbeide en strammere norm, dvs. streke opp hagegangene og luke vekk ugresset i innmarka. Det er vanskelig å se at det er noe poeng i seg selv å registrere alt mulig som kan forekomme i et språk, hvis det ikke skal ha et eller annet pedagogisk

eller normativt siktemål til sjuende og sist, da den språklige kreativiteten er ubegrenset. Om ikke annet er hensikten å vise hva som er forbilledlig eller godt og hva som ikke er fullt så godt. Et mål kan imidlertid være å dokumentere det store språklige mangfoldet for å vise at mangfold i seg selv er et gode.

### *2.2 Preskriptive ordbøker*

En preskriptiv ordbok er først og fremst et pedagogisk og normstyrt oppslagsverk som skal være en hjelp for språkbrukere til å orientere seg om de reglene som gjelder i et språksamfunn. Slike ordbøker kan være en viktig del av arbeidet med implementeringen av vedtatte normer. Ordbøkene dokumenterer den vedtatte normen, slik at brukerne gjennom dem kan finne ut hva som er innenfor normen. Det som oppgis i slike ordbøker, er oftest vedtatt av et normeringsorgan, som har valgt ut noen av alle registrerte former i språket. Utvalgskriterier er alltid diskuterbare, på et eller annet grunnlag må normeringsinstansen anse de formene som velges som bedre enn de som ikke velges, og derved forbys. Det kan være både språkvitenskapelige, ideologiske og politiske begrunnelser. De aller fleste ordbøker som gis ut i dag, både i papirversjon og elektronisk versjon, har et preskriptivt mål.

### *2.3 Språkteknologiske ordbøker*

Det er en utbredt misforståelse at enhver elektronisk ordbok er en språkteknologisk ordbok. Selve publiseringsmåten er ikke avgjørende for det, ofte er helt vanlige normative ordbøker elektronisk leselige, men de er ikke derved nødvendigvis brukbare i språkteknologisk sammenheng. Den største fordelen med det er at de er automatisk søkbare, slik at brukerne slipper å bla seg fram til de enkelte informasjonstypene, og det er mulig å sortere informasjonen på helt nye måter, men ellers er de som tradisjonelle ordbøker.

Språkteknologiske ordbøker er derimot i utgangspunktet skrevet for maskiner og ikke for mennesker. Det betyr at de må ha en spesiell tilgangsstruktur og at lemmaene har en formalisert og systematisk beskrivelse som er maskinelt tolkbar ved hjelp av dataprogrammer. Også språkteknologiske ordbøker kan være normative eller deskriptive, men behovet for deskriptivitet er vanligvis større enn i tradisjonelle ordbøker.

Språkteknologiske ordbøker har forskjellige genuine formål, og forskjellige formål krever forskjellige leksikalske beskrivelser også innen språkteknologien. Det viktigste formålet med språkteknologiske ordbøker er å kunne ”lese” og analysere autentiske tekster for å kunne gjøre visse operasjoner med dem, som oversettelse eller informasjonssøking. De har ikke noe pedagogisk eller normerende siktemål, men er rent tekniske eller praktiske hjelpemidler. Dermed er det behov for at alle former som kan tenkes, korrekte og feilaktige, skal kunne kjennes igjen av leksikondelen. Ordboksdelen i en automatisk oversetter eller i et korrekturlesingsprogram må være normativ i produksjonsdelen, ordboksdelen i programmer for tale- eller tekstgjenkjenning eller i grammatikkontroller bør være mest mulig deskriptiv.

Ore (1998:41) sier at en leksikalsk database er datateknikkens svar på en ordbok, men det er en forenklet og til dels misvisende påstand. Dataene i en språkteknologisk ordbok kan gjerne være ordnet i en database, men det er ikke avgjørende. Det viktigste er at informasjonen er formulert slik at den kan leses og tolkes maskinelt, slik at den kan brukes i språkteknologiske programmer som morfologiske analysatorer, syntaksanalysatorer, oversettelsesstøttesystemer og informasjonssøkingssystemer.

### *2.4 Ordbøker for språkproduksjon*

Det normative i leksikografien er tydeligst mht. ortografi (ordet selv sier jo at dette er normativt). Det er på dette området ordbøker kan påvirke språkbruken mest. Noah Webster gjorde det bevisst, og det samme prøvde de norske språkarkitektene Knud Knudsen og Ivar Aasen seg på – alle med en god porsjon suksess.

Ordboksdelen i en stavekontroll har samme bruksområde som en typisk norsk normativ ordliste. Men i en stavekontroll er det ikke nok å liste opp alle de formene som er tillatt innenfor normen, slik som i en tradisjonell ordliste. Hvis programmet skal fungere godt, er det nødvendig at det inneholder lister over frekvente feilstavinger og liknende feil. En rød korreksjonsstrek kommer i prinsippet opp når et ord *ikke* er oppført i den normerte lista. For også å kunne foreslå korrekt staving må programmet ha en (heuristisk) algoritme for å gjette hvilket ord som er det rette (ved å se hvilke ord i ordlista som ligner mest). Denne algoritmen kan forbedres ved at det legges inn informasjon om vanlige feilstavinger av ord, eller annen relevant informasjon som hvilke taster som ligger nær hverandre på tastaturet eller hva det skulle

være. Ordlistene må derfor bygge på kunnskaper om fonologiske og morfologiske egenskaper i det aktuelle språket.

Stavekontroll bare rettet mot ortografi trenger altså utelukkende å ha kunnskap om det som markeres grafisk i språket. I dag er de vanlige stavekontrollene svært gode, teknisk sett, og brukes av mange. Normering som ikke kommer inn i stavekontrollene, har sannsynligvis liten sjanse til å bli tatt opp i usus. Ordlister på papir vil nok bli uinteressante etter som de fleste venner seg til å bruke stavekontroller. Det er derfor et problem at de er kommersielle produkter og ikke underlagt kontroll fra de offisielle normeringsorganene. Dersom det ikke kommer krav om godkjenningsordning for stavekontroller, slik det er for ordlister til burk i skoleverket, vil mange viktige språkkamper ha vært kjempet forgjeves. Det er derfor viktig å få til et nordisk samarbeid for å få gjennomslag for en slik godkjenningsordning.

En stavekontroll alene kan imidlertid ikke påpeke feil i setninger som **Jeg har enn hest*, det trengs kobling til syntaks for at kontrollen skal kunne rette til *Jeg har en hest*. Derfor må det ligge både en deskriptiv og en normativ komponent inne i en god stavekontroll. Den deskriptive trengs for at stavekontrollen skal kunne registrere og tolke feilene, den preskriptive angir hva som er tillatt innenfor normen. Tidligere var den deskriptive komponenten altfor dårlig, slik at det var mer et heft enn en hjelp å bruke stavekontroll.

Språkteknologiske ordbøker er en nødvendig modul i programmer for simulert eller maskinell språkproduksjon. Den leksikalske komponenten der bør være slik at den både har semantisk kunnskap og er koblet til en ontologi, og har en normativ komponent slik at det bare kommer ut korrekte ord og ordformer. Det gjenstår mye forsknings- og utviklingsarbeid innen leksikografi, semantikk og datalingvistikk før slik språkproduksjon er god nok, og det dreier seg her om leksikalske komponenter som går videre enn i tradisjonell leksikografi. Blant annet må stilmarkeringene være omfattende og gjennomført eksplisitt. Slik eksplisittering kan gi normeringsproblemer som ennå ikke har vært drøftet.

## *2.5 Ordbøker for språktilegnelse*

En annen type språkteknologiske ordbøker trengs for simulert språktilegnelse, der man forsøker å få en maskin til å lese og tolke en tekst. Spørsmålet om deskriptivitet eller normativitet er da irrelevant. Maskinell språktilegnelse er nå blitt et svært populært forskningsfelt. Den semantiske beskrivelsen må gjøre det mulig å tolke absolutt alt som

kan forekomme av tegnsekvenser i en tekst, eventuelt språklyd i tale, både det som er innenfor og det som er utenfor normen, og videre kunne koble det til lemmaene i en ordbok. Det må også være mulig å kjenne igjen åpenbare feil, slik stavekontroller også bør være i stand til. Videre må det være mulig å sortere ord i synonymer, hyperonymer og hyponymer med kohyponymer, paronymer m.m. Det forutsetter at alle lemmaene er systematisert i en gjennomstrukturert ontologi. Slik ordning krever normering av betydningsbeskrivelse som hittil bare i en viss grad har vært gjort av terminologer.

*2.6 Ordbøker for mennesker eller maskiner?*

Det er viktig for normeringsmåten, både i menneskelesbare og maskinlesbare ordbøker, om ordbøkene er beregnet på produksjon eller resepsjon. Menneskelesbare ordbøker er som regel multifunksjonelle ordbøker, mens de språkteknologiske må være helt målspesifikke.

Når en maskin skal produsere eller tilegne seg språk, er ordboksbehovet annerledes enn om det er et menneske. I leksikondelen i informasjonssøkeprogram er det nødvendig med maskintolkbar semantisk beskrivelse som er koblet sammen etter ontologiske eller tesauriske prinsipper, og også her er det en fordel om programmet kjenner igjen alle formvarianter og klarer å koble dem til rett lemma. I språkteknologiske ordbøker der produksjon ikke er målet, er det ikke alltid nødvendig med noen normativ komponent, men det er en fordel når en f.eks. skal koble en form til en definisjon. Inngangen til definisjonen vil være en normert form.

## 3. Krav til ordbøker

Det genuine formålet med pedagogiske ordbøker og språkteknologiske ordbøker er altså forskjellige. De språkteknologiske skal helst være så omfattende og entydige som mulig, mens de pedagogiske skal være innsnevrende i forhold til det mulige. Pedagogiske ordbøker presenterer det aktuelle ordforrådet og formverket som skal følges, språkteknologiske må ha med både det aktuelle og det potensielle ordforrådet.

En språkteknologisk ordbok skal blant annet kunne kjenne igjen alle ord i tekster som forekommer. Det innebærer at det også skal gå an å analysere det som er mulig og kanskje brukbart i visse kontekster, men som ligger utenfor den pedagogiske normen.

### *3.1 Enkelhet og mangfold*

Språkteknologiske ordbøker bør være så enkle som mulig. Det er derfor lettere å lage språkteknologiske ordbøker for strengt normerte språk. Jo flere valgmuligheter og jo større språklig mangfold et språk har, jo mer kompliserte algoritmer må det til for å lage en dekkende leksikalsk beskrivelse. Resepsjonsordbøker kan ta inn så mye som mulig av mangfoldet i språket uten at det får store konsekvenser, for produksjonsordbøker er det annerledes. De som har utarbeidet stavekontroller for norsk, har problemer med alle de mulige tillatte formene innenfor normen. Selv enkle syntagmer får fort flere hundre kombinasjonsmuligheter. Victoria Rosén (2000:216) har funnet at følgende setning har hele 165.888 mulige stavemåter på norsk bokmål:

> De lavtlønte sykehjemsansatte ble helt utmattet og slukket tørsten med den surnete fløtemelken.

### *3.2 Fullstendighet*

Idiosynkratiske trekk i morfologien gir store problemer i språkteknologiske ordbøker. I tillegg til ufullstendige paradigmer er det i mange språk suppletivismer, som gir dobbelte paradigmer eller uregelmessig avledede former. Ufullstendige paradigmer oppstår fordi det er en del former vi ikke har behov for, fordi de ikke tilsvarer referenter eller begreper i virkeligheten. Men i fagspråk oppstår ofte behovet for fulle paradigmer, og derfor er det greit å ha standardisert hele det mulige forminventaret, slik man gjerne gjør innen terminologi. Det betyr ikke at man regner de deriverbare, men ikke brukte formene som korrekte, de er uproblematiske, sovende former i resepsjonsordbøker. Slike former kan imidlertid gi problemer i de pedagogiske ordbøkene.

Hvilke substantiv som kan flertallsbøyes, er ikke noe helt enkelt språkvitenskapelig spørsmål. Ord som vi tradisjonelt bare har brukt i entall, forekommer nå stadig oftere med flertallsbøying. Ifølge Norsk referansegramatikk (Faarlund et al. 1997:143f) flertallsbøyes ikke abstrakter og masseord. Det er en begrensningsregel som er semantisk fundert. Men denne regelen er ikke fast og endelig. Ved spesielle kommunikasjonsbehov er det både mulig og akseptabelt å flertallsbøye også slike ord:

> Han har få *ærgjerrigheter* ut over å slenge bannord i trynet på borgerskapet.
> De har fått mange nye og gode *viner* på Vinmonopolet i høst.

Likevel fins det en grense for hva en kan flertallsbøye. Ingen vil komme på å snakke om *flere varmer* eller *flere kulder*. Og selv om slike former forekommer en sjelden gang som en stilistisk variant eller i spesielle kommunikasjonssituasjoner, er det ikke selvfølgelig at formene skal føres opp i en normerende ordbok ment for tekstproduksjon. Problemet blir likevel hvor man skal sette grensen. Det er mange former som er opplagte, slik som *kulder, molybdener*, men hva med *ærlighet, bly, glass?* Man kan tvile ved flertallsformen *ærligheter*, men det synes som om *uærligheter* er helt ok. Slikt er det svært vanskelig å gi velbegrunna regler for, det kan være både semantiske og pragmatiske forhold som avgjør mulighetene for bøying. Ordbøker i nordiske språk er særlig vanskelig å beskrive pga. utstrakt evne til å lage sammensetninger. Når sammensetninger lages automatisk ut fra grunnord med kode for bøying, følger koden for det usammensatte lemmaet med til sammensetningen. Men ofte er det slik at et grunnord kan ha fullstendig bøyningsparadigme, mens sammensetningen er blokkert for det. Vi har fullstendig paradigme for en *galskap* flere *galskaper*, men vi kan ikke bøye sammensetningen i flertall til *stormannsgalskaper*, det samme gjelder *heim* og **tåkeheimer*. Grunnen til denne blokkeringen er sannsynligvis at sammensetningene betegner så spesielle eller spesifiserte referenter at det sjelden vil forekomme flere eksemplarer i samme kontekst. I andre tilfeller er det snakk om individualiserte fenomener som har de egenskapene forleddet uttrykker. Dersom det blir flertallsbøyd, er det fordi betegnelsen *terminologiseres* som fellesbetegnelse på varianter av fenomenet, f.eks. fins det flere typer *salater*: *rucculosalat, issalat, bladsalat* etc.

### *3.3 Sterke regler vs. usus*

Orddanningsreglene i språk som de nordiske, er som vi vet svært sterke. Ved avledning kan vi danne mange flere ord enn de som faktisk er i bruk. Det kan lages helt nye ord som uttrykk for nye begreper, for eksempel: *behyggende, inspirativ, replikere, tidslig* i stedet for *hyggelig, inspirerende, replisere, tidsriktig.* Hva som gjør noen former akseptable og andre ikke, er umulig å avgjøre før usus viser det. Alle de først nevnte formene blir merket med stavekontrollen, ingen i den sist

nevnte. Det sier at stavekontrollen ikke bare er basert på regler, men også har en ordliste å kontrollere feil mot.

På samme måte kan vi føye avledningsendelsen *-het* til et adjektiv og få et substantiv: god+het = *godhet*, stor+het = *storhet*. Men regelen gir også former som *sivilhet, gørrhet, klaverspillendehet.* Slike former er ikke akseptable i norsk, ikke fordi de bryter med regelsystemet, men fordi vi sannsynligvis ikke har bruk for dem. De hører ikke med i en pedagogisk produksjonsordbok, men vil dukke opp i en regelgenerert ordbok. Nettopp fordi reglene i grammatikken gir flere former enn det vi trenger for å dekke våre kommunikasjonsbehov, trenger vi ordbøker. Ordbøkene skal kun vise hvilke former som faktisk er i bruk. Men for en språkteknologisk ordbok som bare skal brukes i tekstanalyse, er det viktig både at vanlige former som må regnes til usus, og tilfeldige former som avviker fra disse, kan gjenkjennes. Det er derfor helt greit at en språkteknologisk ordbok inneholder alle genererbare former.

## 4. Gjenbruk av pedagogiske ordbøker i språkteknologiske programmer

Mange har forsøkt å tilpasse tradisjonelle ordbøker og bruke dem som ordboksdel i språkteknologiske produkter. Det er mye som kan innspares av arbeid ved utvikling av språkteknologiske ordbøker på grunnlag av eksisterende, pedagogiske ordbøker, men det er også mange fallgruver. Man må balansere mellom kravene til enkelhet og til mangfold.

At ordbøker med forskjellige genuine formål ikke uten videre fungerer for andre formål, blir tydelig illustrert i nettversjonene av Bokmålsordboka og Nynorskordboka. Bokmålsordboka, både som nettversjon og papirversjon, er først og fremst et pedagogisk verktøy for normert formverk og noenlunde standardisert betydningsbeskrivelse. Det er utarbeidet en morfologisk base for bokmål som er koblet til internettversjonen av Bokmålsordboka. Den morfologiske basen er utviklet ut fra språkteknologiske behov om gjenkjenning og tolkning av flest mulige ord og ordformer, bl.a. hos IBM Norge (jf. Engh 1993).

Da Bokmålsordboka og Nynorskordboka i sin tid ble laget, ble det brukt kodenøkler for bøyningsparadigmene, bl.a. for å forenkle framstillingen og for å spare plass. I stedet for å skrive

**stol – stolen – stoler – stolene**

står det bare

**stol** m1

Ut fra en kodenøkkel kan en så under m1 finne de enkelte bøyningsformene. Et slikt kodesystem er rasjonelt og praktisk, selv om det gir brukerne noe ekstra bryderi med å bla fram til innsiden av permen for å finne nøkkelen med et fullt utskrevet bøyningsparadigme.

## *4.1 Overgenererte former*

De mange ufullstendige paradigmene i naturlig språk gjør det likevel ikke umulig å benytte et kodesystem for bøying i pedagogiske ordbøker. Selv om kodene strengt tatt oppgir former som aldri har vært realisert eller som kan regnes som korrekte, blir disse formene sannsynligvis ikke lagt merke til, siden ingen vil komme på å lage dem eller slå opp på dem. Paradigmene er jo ufullstendige nettopp fordi reglene gir mulige former vi ikke trenger. Det er altså et skisma mellom potensielle og realiserte former som kan feies under teppet i en menneskerettet pedagogisk ordbok, men som kommer fram i dagen i en teknologisk ordbok.

I 1989 sammenliknet man ved IBMs språkavdeling det ordforrådet som lå inne i en elektronisk versjon av Bokmålsordboka med de formene som faktisk var i bruk i tekster. Resultatet av denne undersøkelsen var nyttig også ved revidering av Bokmålsordboka, fordi den avslørte mange feil og inkonsekvenser. IBM arbeidet imidlertid med et språkteknologisk mål for øye, ikke et språkpedagogisk, og utarbeidet deskriptive lister av ordforrådet i løpende tekster. IBM-listene ble seinere utgangspunktet for de morfologiske basene som Tekstlaboratoriet og Dokumentasjonsprosjektet ved Universitetet i Oslo har utviklet til, det som nå kalles Norsk ordbank. En slik orddatabase var nødvendig for å få laget en automatisk ordklassetagger. Orddatabasen for bokmål har i dag 1,2 mill. fullformer og er en slags språkteknologisk ordbok for den formelle siden av ordforrådet.

## *4.2 Menneskelig fleksibilitet vs. maskinell rigiditet*

I en pedagogisk ordbok er det mulig å variere den grammatiske informasjonen, f.eks. slik at det vanlige kodesystemet brukes ved substantiv som har fulle paradigmer, mens en annen kode gis ved lemmaer med ufullstendige paradigmer. Der paradigmene er ufullstendige, settes

de vanlige kodene ikke på. I Bokmålsordboka (1986) er problemet løst nettopp slik for å markere ufullstendigheten:

> **hull** n1 (= et hull, hullet, flere hull, hulla/hullene)

men bare

> **gull** -et

Brukerne skal altså ut fra dette forstå at *hull* kan bøyes i alle former, mens *gull* bare har bestemt form entall som bøyningsmulighet, dvs. at det ikke kan stå i flertall. En slik notasjon gir på en ganske implisitt måte informasjon om bøyningsmulighetene, og klarer slik stort sett å redegjøre for normen. Det er bare i få tilfeller at det vil være problematisk, og det er nettopp i gråsonen mellom det normerte og det unormerte, det vil si der normene kanskje er i forandring. Vanligvis gir ikke det store problemer ved bruk av en pedagogisk ordbok. Brukerne forstår at former som er mye i bruk, kan godtas selv om de ikke er nevnt i ordboka. En menneskelig bruker kan legge inn vurdering i tillegg til det som er nevnt i ordboka. Men datamaskiner kan ikke hanskes med slik fleksibilitet. I nettversjonen av Bokmålsordboka får brukerne følgende informasjon om de søker på former som *gullene* eller *kulder*:

| Tilslagsord: | Bøyingsform | Oppslagsord |
|---|---|---|
| gullene | substantiv intetkjønn, flertall, bestemt form | gull |

| Tilslagsord: | Bøyingsform | Oppslagsord | **kulde** *-a* el. *-en* (norr *kuldi*) |
|---|---|---|---|
| kulder | substantiv hunnkjønn, flertall, ubestemt form | kulde | **1** lav temperatur, kaldt vær *frost og k- / 20 graders k-* |
| kulder | substantiv hannkjønn, flertall, ubestemt form | kulde | **2** uvennlighet, avvisende holdning *alle nykommere ble møtt med k- og mistenksomhet* |

Det er ikke uten videre klart at formen *gullene, kulder* ikke er tillatte former i norsk bokmål, noe som må virke både irriterende og villedende for en person som vil orientere seg om hva som regnes som korrekt eller ukorrekt bøyningsmåte.

I en språkteknologisk ordbok som skal brukes i simulert tekst-resepsjon, trenger man ikke legge slike begrensninger på beskrivelsen, da er det først og fremst gjelder at alle tenkelige former kan gjen-kjennes. Automatisk genererte paradigmer tvinger fram former som ingen vil komme til å velge. Men en pedagogisk ordbok må renses for uaktuelle former. En pedagogisk ordbok med utgangspunkt i automatisk genererte fullformer stiller redaktøren overfor en lang rekke normeringsspørsmål som ingen tidligere har tenkt på.

Da Bokmålsordboka og Nynorskordboka ble lagt ut på Internett som gratisvare, ble samtidig alle kodene gjort om til fulle former, slik at brukerne skulle slippe å finne ut hvordan kodenøklene skulle tolkes. Alle ord fikk fulle former, og man så bort fra at det fantes ufullstendige paradigmer, jf:

| | |
|---|---|
| **revolusjonær** | **I ~æ´r** m1 person som kjemper for revolusjon, opprørstilhenger |

| Tilslagsord: | Bøyingsform | Oppslagsord |
|---|---|---|
| revolusjonærene | substantiv hannkjønn, flertall, bestemt form | revolusjonær |

I de morfologiske basene, som fullformene genereres ut fra, er det lagt inn en lang rekke former som er mulige ut fra reglene, men som ikke er tillatt innenfor den offisielle rettskrivingen. Det gjelder feilformer eller unormerte former som *juge – jugde – jugd*. Og selv om det heter *et stadion*, er det behov for å tolke ”feilen” *en stadion* som substantiv med samme betydning som *et stadion*, eller bøying som *ei maskin – maskina*. Dette er former som ligger utenfor normen, men som er ganske vanlige i tekster likevel. Disse formene hentes imidlertid ikke inn ved søk i de elektroniske ordbøkene. Det hadde vært mulig å tenke seg at de gjorde det, men foreslo de korrekte formene i stedet, slik en del stavekontroller gjør.

Søk i de ordbøkene som ligger på nettet, blir loggført. Det gjelder også Bokmålsordboka og Nynorskordboka. Analyse av slike logger gir nyttig informasjon om brukerbehovene. Loggene viser at de fleste faktisk søker på bøyde former av lemmaene, det er derfor viktig at de er tilgjengelige for brukerne. Men denne koblingen av de morfologiske basene og Bokmålsordboka har dermed gitt en del problemer med de formene som ligger i gråsoneområdet mellom det normerte og det unormerte. Det gjelder særlig flertallsbøying av substantiv og gradbøying av adjektiv. Men det er også mange andre former som må kontrolleres for at en språkteknologisk utviklet base samtidig kan fungere i et pedagogisk verktøy. Det ble dermed nødvendig å gå

gjennom alle formene og vurdere om de var aktuelle eller bare potensielle former i norsk.

Den manuelle gjennomgangen av basen har ført til diskusjoner om bøying og rettskriving av sjeldne ord og ord som ikke egner seg for normering, bl.a. en del interjeksjoner og andre typiske talemålsuttrykk. Mange restriksjoner har med forholdet til verden å gjøre, vi bøyer ikke alt i flertall, f.eks. fordi ordet har et utellelig innhold: *gull, karies, varme, kulde*. Vi gradbøyer ikke adjektiv for fenomener som er absolutte, som *gift, avsluttet, sivil.*

Men nettversjonen av Bokmålsordboka gir treff på *sivilere, logiskest.*

Koblingen har tvunget oss til å normere det unormerbare. Det vil alltid være en stor gråsone mellom det normerte i språket og det som vi ikke trenger normere. Disse formene er av to typer:

1. ordformer som regnes som feil, men som mange bruker
2. ordformer som ingen noen gang vil komme på å lage, men som er resultat av automatisk overgenerering.

Opprettinga har vist seg å være et ganske stort arbeid. Det er sannsynligvis ikke mulig å finne regler som skiller mellom disse to typene feil, og det er heller ingen klar grense mellom dem, for vi har jo mulighetene til å tøye grensene. Det bør derfor vurderes om det er fornuftig å bruke så mye arbeid på å plukke ut aktuelle ordformer fra automatisk genererte lister over potensielle ord i pedagogiske ordbøker. Det må tas stilling til en lang rekke normeringsteoretiske spørsmål som aldri har vært aktuelle, og som sannsynligvis er helt marginale i språkbrukssammenheng. Ingen vil komme på å bøye *karies, kulde* eller *kusma* i flertall.

### *4.3 Relevans*

Pedagogisk leksikografi har tatt utgangspunkt i usus og hva som ut fra den er blitt konvensjonalisert i språket. Med automatisk genererte regler tar man utgangspunkt i et rigid system for at det skal tilpasses menneskelige brukere. Det gir utrolig mange uavklarte normeringsspørsmål, og det må vurderes om disse spørsmålene er verdt å bruke kreftene på. Det er ikke uten videre åpenbart at de spørsmålene som dukker opp via automatisk genererte lister, er relevante å diskutere for praktisk språkrådgivning. Det er forholdsvis enkelt å avgjøre at visse

adjektiv aldri skal gradbøyes (jf. Fjeld 1998), men det vil likevel være mange som er tvil om former som

*falleferdigere – falleferdigst*
*fingerferdigere – fingerferdigst*
*all rightere – all rightest.*

Det er usikker normering her, og det er faktisk få (ingen?) norske ordbøker som angir regler for gradbøying av adjektiv. Dialektal og sosial variasjon er kanskje grunnen til at gradbøying har vært unndratt språknormeringsinstansenes kontroll.

Det leksikografiske problemet er at man forsøker å kombinere flere ordboksfunksjoner i en og samme ordbok. De pedagogiske ordbøkene blir derved fulle av feil, og de morfologiske basene som grunnlag i teknologiske produkter vil bli ufullstendige og mangelfulle om de unormerte formene fjernes. Spørsmålet er om det er regningssvarende å forsøke å kartlegge gråsonen mellom det normerte og det unormerte i et språks ordinventar.

## 5. Hva er viktig informasjon i en pedagogisk ordbok?

Brukerstudier viser hvordan allmennspråklige ordbøker brukes. Béjoint (2000:151) oppsummerer en lang rekke brukerundersøkelser i grunnleggende brukerbehov og forutsetninger for god leksikografi:

1. Kompetente brukere har avanserte behov.
2. Enspråklige ordbøker brukes mest for avkoding av skriftspråk.
3. Enspråklige ordbøker konsulteres pga. definisjonene, særlig for sjeldne ord. Mange har likevel problemer med å forstå det som står i definisjonene.
4. Informasjon for språkproduksjon blir lite brukt.
5. Generelt viser ordboksbrukere lite fantasi. De leter sjelden i flere ordbøker, og tar det første som passer sånn noenlunde som den hele og fulle sannheten.

Béjoint konkluderer derfor med noe han kaller leksikografenes paradoks:

> Thus lexicographers are faced with a paradox: dictionaries are used almost exclusively as collections of definitions, and the information that takes most time to prepare (the entries for frequent, polysemous words and the information for encoding) is hardly ever used.
>
> (Béjoint 2000:153)

Dette er viktige funn fordi koding av syntaktisk informasjon tar mye tid. Mange mener at leksikografer bruker masse tid på unødvendig arbeid, man skal bare ta med det som folk slår opp. Jackson 1988 (ref. Béjoint ibid.) mener at brukerne heller ikke ville savnet grammatisk informasjon som ordklasse, etymologi og kanskje også uttale. Også Della Summers peker på at det er betydning som folk flest søker opp, annen informasjon blir oppsøkt nesten bare i opplæringssituasjoner. Det folk leter etter, er ”hard words, new words, obsolete words”. Men da ser man bort fra den pedagogiske normerende oppgaver ordbøker skal ha. Jeg mener imidlertid at ordboksbehovene kan endre seg med moderne teknologi i leksikografien. Siden en stavekontroll kan kobles direkte opp mot en ordliste for valg av rett ortografi, burde det være mulig å gå rett fra tekst og inn i en ordbok også, der man kan finne fullt sett av alle tillatte bøyningsformer.

## 6. Normering vs. registrering

Normer kan være kvalitative eller kvantitative (Rey 1972). Kvantitative normer baseres på språkbruk og ut fra konsensus mellom språk-brukerne. Kvantitativ norm tillater en hvilken som helst form, bare den er brukt av mange nok personer. Her er korpusstudier en viktig metode for normeringsarbeidet. Kvalitative normer nedfelles derimot ut fra den tanke at noen språkbrukere er bedre enn andre, som berømte forfatterne eller andre med ledende posisjon i forhold til skriftspråket. Normeringen blir da gjerne begrunnet ut fra studier av tidligere produsert språk. Det blir gjerne også referert til etymologi og logikk som viktige argumenter, men dette er svært vanskelige kriterier å anvende på en konsekvent måte i praktisk språkrøkt. Mange ordbøker med utgangspunkt i kvalitative normer kan klassifiseres som moraliserende leksikografi. Stabilitet regnes da som viktig, dessuten at visse ord eller former eller betydninger unngås. Det vil si at noen ord eller former blir regnet som gode, andre ikke.

Men mange ordboksfolk liker dårlig å arbeide ut fra kvalitative normer, og forsøker å finne mellomløsninger. En måte å være både deskriptiv og normativ på, kan være å oppgi bruksmarkering som

opplyser at et ord f.eks. er tabu eller utenfor normen. Pussig nok kan det både fungere som en måte å unngå å være normerende på, og som det omvendte. Imidlertid blir slik nyansering helt umulig i språkteknologiske ordbøker, i hvert fall kan det bare virke i tilfelle man har avanserte stilfiltere som gjør at visse ord velges ut fra visse kontekstkrav i forhold til et angitt stilnivå som blir innstilt på forhånd

Pedagogiske ordbøker er rettet mot den språklige innmarka. Ideologien er å beskytte språket mot forfall. Metaforen som ligger under antyder en forestilling om at språk oppstår og vokser opp til sitt maksimum i styrke og kraft, det er gjerne det stadium det hadde da en selv gikk på skolen, og derfra er det utsatt for all slags farer og angrep som vi må beskytte det mot. En merkelig tankegang! Som språkkonsulent i NRK får jeg mange henvendelser om alle de forferdelige feilene som gjør språket så fattig. Under et slikt spørsmål ligger selvsagt tankegangen om at noe språk er rikt, og annet språk er fattig, og at all utvikling går i retning av deprivasjon. Enhver endring anses som et tap, jeg hører aldri at folk gleder seg over at noe nytt er kommet inn, f.eks. når mange i dag sier: *Jeg hadde tenkt til å gjøre det*, i stedet for det ”korrekte” *jeg hadde tenkt å gjøre det*. Man kunne jo tenke seg at disse to uttrykksmåtene ville få hver sin betydningsnyanse etter noen år, og dermed var språket blitt rikere, ikke fattigere. Det er vanskelig å forstå at språket blir fattigere av å legge til et ord ekstra i en frase.

Den normering som har foregått gjennom de senere års vedtak i Norsk språkråd, har blitt kalt oppryddingsvedtak, fordi man har forsøkt å gjøre reglene for bøying og rettskriving enklere. Men dermed innskrenkes også utviklingen av det språklige mangfoldet. Det bør i det minste kunne diskuteres om det uomtvistelig er bra**.** Har man i så fall ikke en språkteknokratisk ideologi som styrende prinsipp? Når nettversjonene for Bokmålsordboka og Nynorskordboka har fått lagt til en komponent fra de språkteknologiske ordbøkene, blir det tydeliggjort mange ”regelfeil”, fordi reglene er altfor sterke. Det språklige mangfoldet eller den såkalte lingvistiske uregelmessigheten blir derved tydeliggjort, og normeringstiltak settes inn for å ”rydde opp”. Spørsmålet blir da om det er hensiktsmessig at språkteknologiske ordbaser danner grunnlag for språkrådgiving.

Overgenererte former gjør ikke noen skade i analysedelen i en språkteknologisk ordbok, men de er ikke akseptable i et pedagogisk og retningsgivende oppslagsverk. Det viser seg at det i praksis er vanskelig å gjenbruke tradisjonelle, eksisterende ordbøker i språkteknologiske programmer. Derimot er det i dag viktig å ha begge mål for øye når en lager en ny ordbok, slik det forsøkes gjort i prosjektet LDB (Leksikalsk database for moderne bokmål) ved Institutt for nordistikk

litteraturvitenskap ved Universitetet i Oslo. Basen skal både fungere som ordboksdel i språkteknologiske programmer og samtidig gi grunnlag for revisjon av Bokmålsordboka. Med utgangspunkt i Bokmålsordboka og det danske SIMPLE-leksikonet, som er en språkteknologisk ordbok for det sentrale ordforrådet i dansk, lages en semantisk og morfologisk beskrivelse som både er formalisert og strukturert og har fridefinisjoner som kan benyttes i pedagogisk øyemed, men som innholdsmessig svarer til den formaliserte definisjonen.

## 7. Avslutning

For å knytte tilbake til sammenlikningen mellom språk og natur i innledningen, kan man se det slik at de språkteknologiske ordbøkene må forsøke å harve seg gjennom hele den språklige utmarka. Det er dermed klart at språkteknologer må bygge på en helt annen normforståelse enn de som lager pedagogiske ordbøker. Et ufravikelig krav for språkteknologiske ordbøker er enkelhet, noe som forutsetter streng normering. Målet for god pedagogisk leksikografi må være å registrere det språklige mangfoldet, og derved dokumentere hva som er mulig og fungerer kommnikativt i det menneskelige språket. Dette er hensyn som ikke uten videre kan forenes.

## 8. Litteratur

Béjoint, Henri 2000: *Modern lexicography: an introduction*. Oxford: Oxford University Press.

Engh, Jan 1993: Linguistic Normalisation in Language Industry. I: *Hermes* 10.

Fjeld, Ruth Vatvedt 1998: *Rimelig ut fra sakens art. Om tolkning av adjektiv i regelgivende språk*. Dr.avhandling, Universitetet i Oslo.

Faarlund. J. et. al 1997: *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Landrø, Marit & Boye Wangensteen 1993: *Bokmålsordboka*, Oslo: Universitetsforlaget.

Omdal, Helge 1999: Språknormsikkerhet i bokmål og nynorsk. I: Helge Omdal (red.): *Språkbrukeren – fri til å velge? Artikler om homogen og heterogen språknorm*. Kristiansand: Høgskolen i Agder

Ore, Chr.-Emil 1998: Hvordan lage databaser for språk- og kulturfag. I: Knut Aukrust og Bjarne Hodne (red.): *Fra skuff til skjerm. Om*

*universitetenes databaser for språk og kultur*. Oslo: Universitetsforlaget.
Rey, J. 1972: *Le mot el l'idee: révision vivante du vocabulaire anglais.* Paris: Ophrys.
Rosén, Victoria og Koenraad de Smedt 2000: Er korrekturlesingsevnen di god? Resultater fra SCARRIE. I: *Nordlyd* No. 28. Universitetet i Tromsø.