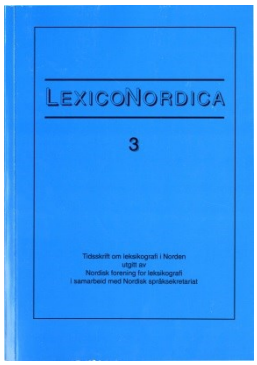
LexicoNordica

Titel:	Korpusbasert leksikonbygging
Forfatter:	Lars G. Johnsen og Torbjørn Nordgård
Kilde:	LexicoNordica 3, 1996, s. 69-79
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.

Lars G. Johnsen & Torbjørn Nordgård

Korpusbasert leksikonbygging

The article outlines the motivation for two research projects relevant for computational lexicography applied to Norwegian: a text corpus and a computational dictionary. The corpus must be composed with special attention to the rather special status for written Norwegian because Norwegian has two written standards (Bokmål and Nynorsk). It is argued that these standards contain substandards which are not described systematically. We argue that such substandards can be discovered by careful use of corpora, and the substandards can be represented as different electronic dictionaries to be used in for instance proof reading systems.

1. Innledning og sammendrag

I denne artikkelen blir det skissert hvordan korpus kan brukes i konstruksjon og bygging av et datamaskinelt leksikon innenfor det språkteknologiske samarbeidet *Norsk Infrastruktur For Språkteknologi* (NIFST). Dette samarbeidet startet våren 1995 mellom de språkteknologiske miljøene ved universitetene i Bergen (UiB), Oslo (UiO) og Trondeim (NTNU), og resulterte i følgende ønskeliste for ressurser og verktøy for norsk språkteknologi:

i. Et norsk tekstkorpus
ii. Et komputasjonelt ordleksikon
iii. En tagger for morfologisk analyse og merking av ordene i en vilkårlig tekst
iv. En grammatikk for automatisk syntaktisk analyse og generering av tekst

Norges Forskningsråd har bevilget støtte til utvikling av tagger[1] ved UiO og konstruksjon av komputasjonelt leksikon (NorKompLeks) ved NTNU. Gjennom NIFST sikres disse prosjektene en koordinert utvikling. Sentralt i koordineringsarbeidet er å utvikle et system for klassi-

[1] En tagger er et program som merker ord mht. ordklassetilhørighet, morfosyntaktiske egenskaper etc.

fisering av ord og fraser som brukes av både tagger, grammatikk og leksikon.[2]
I det følgende skal vi gå inn på korpus og leksikon. Koblingen mellom de to blir illustrert ved å se på hvordan hypoteser om stilformer kan valideres eller genereres fra korpuset.

2. Norsk Tekstkorpus

Humanistisk datasenter ved UiB har ansvaret for å gjennomføre oppbyggingen av et norsk tekstkorpus. Fra tidligere finnes ikke noe offisielt korpus til bruk i studier av norsk, og prosjektets formål er å gi forskere en felles empirisk ressurs for støtte til filologiske eller lingvistiske studier, spesielt med tanke på å danne et grunnlag for språkteknologisk utvikling. Prosjektets navn er Allmennspråklig Korpus (ASK), og det fokuserer på norsk skriftspråk i alle dets varianter med vekt på bokmål og nynorsk og delvarianter av disse.

Korpusets formål dikterer dets struktur og størrelse, forstått som det utvalg og den kvantitet av tekster som inngår. Gitt koblingen med det komputasjonelle leksikonet vil informasjonen om norsk som er beskrevet der, legge bestemte føringer på hvilke tekster eller tekstyper som skal inngå. Leksikonet vil, foruten kvantitativ informasjon, først og fremst inneholde kvalitativ informasjon om et ord og dets bøyningsformer. Et korpus balansert med hensyn til ytre faktorer vil ikke i seg selv gi noen garanti for at den ønskede informasjon er til stede.

Det er fire grammatiske moduler korpuset skal bevitne:

(1) Leksikonet som en liste av ord, stammer og affikser, idiomer og kollokasjoner
(2) De leksikalske regler som beskriver hvordan ord bygges opp til nye ord, og hvordan egenskaper til ord endrer seg i denne prosessen
(3) De syntaktiske regler som beskriver hvordan ord og fraser kan settes sammen til nye fraser
(4) Stilvalg som antyder hvilke systematiske delstandarder av språket som er i bruk

---

[2] Denne klassifiseringen er ikke klargjort ennå, men taggerprosjektet og leksikonprosjektet samarbeider om distinksjoner og grammatisk terminologi. Det må bl.a. avgjøres om den grammatiske beskrivelsen skal være i form av trekkstrukturer eller lister av kategorier.

Av disse modulene vil (1–2) og (4) kodes i NorKompLeks. Men det er spesielt (2) og (4) som legger en særlig føring på korpusets innhold. Med hensyn til (1) ville strategien for å sikre god dekning ganske enkelt være å samle flest mulig tekster med størst mulig variasjon. Problemet med å realisere (1), særlig for enkelte germanske språk og spesielt norsk, er at de har forholdsvis frie regler for ordsammensetninger slik at det vil være umulig å sikre at alle ord som finnes i språket, skal være representert i korpuset. Dette, sammen med de økonomiske og administrative føringer som ligger på selve størrelsen av korpuset, betyr at utvalget i større grad må bestemmes av representativiteten av de språklige egenskaper som *kan* kontrolleres.

I forbindelse med (2) kan vi se på de regler som lager substantiv av verb og som oppretter systematiske forbindelser mellom verbets syntaktiske kontekst og konteksten for det resulterende substantiv. For eksempel, regelen som legger affikset *ing* til enkelte verb, vil systematisk ta med seg noen av verbets kontekster, men ikke alle. For *spise* gjelder det at *spisingen av eplet* er helt i orden, mens *spisingen av opp eplet* ikke er mulig. Slike og andre eksempler er beskrevet i detalj i Johnsen *et.al.* (1989).

Dersom et korpus skal kunne bevitne slike leksikalske sammenhenger, må de ganske enkelt forekomme i korpuset, og tekstene må velges slik at det er en sjanse for å finne dem. Vi skulle forvente *a priori* at et utvalg av tekster som hadde høy hyppighet av enkelte ordstammer, vil ha større sjanse til å bevitne slike sammenhenger enn tekster der variasjonen i ordstammer er stor gitt samme størrelse på utvalget. Studiet av delspråk beskrevet i McEnery & Wilson (1996) støtter opp om dette. Dersom en begrenser tekstutvalget til å gjelde tekster som omhandler spesifikke områder[3], vil et bestemt leksikalsk element opptre i flere varianter enn i et generelt utvalg av tekster. For våre behov betyr det at i korpuset bør en substansiell del være viet et delspråk som inneholder forholdsvis få ordstammer.

For å kunne dekke stilarter må man ta hensyn til hvilke uoffisielle "stilnivå" tekstutvalgene er basert på. Stilnivå og stilarter vil vi ta for oss i seksjon 4, men for å kunne representere dem fordres det hypoteser om hvor man kan vente å finne variasjon i stil.

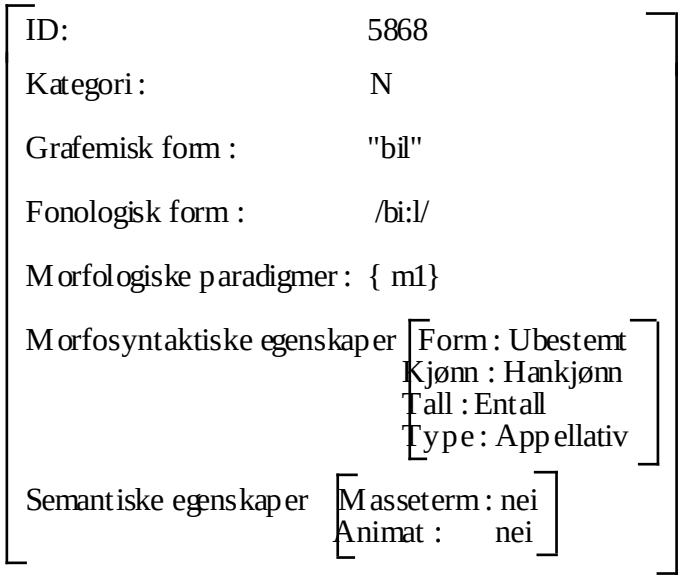
3. NorKompLeks

[3] I dette tilfelle IBM-manualer versus Canadian Hansard Corpus.

Prosjektet *Norsk komputasjonelt leksikon* (NorKompLeks) har som målsetting å realisere et språkteknologisk leksikon for norsk. Norsk blir forstått som de to skrifstspråksstandardene bokmål og nynorsk. I tillegg skal det utarbeides et omfattende fonologisk leksikon for norsk, basert på leksemene i bokmål og nynorsk.

3.1. Leksikalsk informasjon i NorKompLeks

I NorKompLeks vil oppslagsordene ha følgende typer informasjon knyttet til seg: *kategori, grafemisk form, fonologisk form, morfologiske paradigmer, morfosyntaktiske egenskaper, syntaktiske egenskaper* og *semantiske egenskaper*. Informasjonen er organisert som rekursive trekkstrukturer, jf. Shieber (1986), noe som gir et deklarativt og oversiktlig representasjonsformat som det er enkelt å modifisere, det være seg med mer informasjon, mindre informasjon eller annen informasjon. Et eksempel er angitt i figur 1, der leksemets grafemiske form er *bil*, ordklassekategorien er *N* (nomen), identifikasjonskoden er 5868, den fonologiske beskrivelsen er */bi:l/*, det morfologiske bøyningsparadigmet er *m1*, den morfosyntaktiske egenskapen *form* har verdien ubestemt, den morfosyntaktiske egenskapen *kjønn* har verdien hankjønn, den morfosyntaktiske egenskapen *tall* er entall, og den morfosyntaktiske egenskapen *type* er appellativ. Videre er de semantiske egenskapene *masseterm* og *animat* markert som negative.



Figur 1: Leksikonoppslaget **bil**

Vi skal i det følgende konsentrere oss om den morfologiske informasjonen.

3.2. Morfologiske beskrivelser i NorKompLeks

Morfologikoden *m1* skal fortolkes som en instruksjon om hvordan bøyningsparadigmene til leksemets grunnform skal produseres. I dette tilfellet er det enkelt: Ubestemt form entall er lik grunnformen, bestemt form entall dannes ved å legge til bokstavene *en*, ubestemt form flertall dannes ved tillegg av *er*, mens *ene* i tilleg til grunnformen gir bestemt form flertall. Andre instruksjoner kan erstatte en eller flere bokstaver i grunnformen med andre bokstaver, noe som er typisk for omlydsbøyninger. I NorKompLeks-prosjektet er det definert og implementert et system som ekspanderer morfologiske koder til fullstendige paradigmer; for mer detaljerte spesifikasjoner, se Nordgård (1996).

Norsk er spesielt mht. morfologisk valgfrihet og liberal, eller kanskje mer presist, ufullstendig normering. Dette har å gjøre med den norske språksituasjonen, noe vi kommer tilbake til nedenfor. I figur 2 finner man et utsnitt av beskrivelsen av verbet *stryke*.[4] Legg spesielt merke til at det finnes to sidestilte preteritumsvarianter: *strøk* og *strauk*. Det spesielle her er at vi finner to likestilte morfologiske realisasjoner av samme morfologiske kategori, i dette tilfellet preteritum av verbet *stryke*.

Imperativ : "stryk"

Infinitiv : "stryke"

Presens : "stryker"

Preteritum : {"strauk","strøk" }

Perfektum partisipp "strøket"

Presens partisipp : "strykende"

Perifrastisk passiv "strøket"

Inherent passiv : "strykes"

Figur 2: Bøyningsparadigmet til *stryke*

[4] Vi ser her bort fra de deriverte adjektivformene *strøket*, *strøkne* og *strykende*. Disse er med i den implementerte versjonen av leksikonet.

I bokmålsnormen er det faktisk slik at mer enn halvparten av alle verb har alternative bøyningsmønstre. Omkring 500 verb kan bøyes etter tre alternasjoner, f.eks. *anklage* som kan ha preteritumsformene *anklaget, anklaga* eller *anklagde*. Enkelte verb kan bøyes på fire måter, f.eks. *love*, der man kan bruke preteritumsformene *lovet*, *lova*, *lovde* og *lovte*.

Bokmålsnormen er altså liberal mht. hva som er tillatt i skriftlig norsk. Men det finnes en del stilnivåer som ikke bør brytes. Eksempelvis bør man ikke bruke formen *strauk* (av verbet *stryke*) og formen *anklaget* (av verbet *anklage*) i samme tekst; *strauk* og *anklaga* hører mer naturlig sammen, men det er ingen som kan referere til mer eller mindre offisielle normer i slike tilfeller. Før vi går videre i gjennomgangen av NorKompLeks og forholdet til korpusmateriale, er det nødvendig med en kort gjennomgang av noen aspekter ved den norske språksituasjonen.

3.3 Den norske språksituasjonen

Det er som kjent to skriftspråkstandarder for norsk. *Bokmål* har en historie som går tilbake til dansk, mens nynorsk ble etablert i forrige århundre på grunnlag av norske dialekter og gammelnorsk (norrønt). De to skriftspråkene er offisielt likeverdige i den forstand at man kan velge hvilket man vil bruke.

Det har alltid vært tendenser til polarisering i begge språkleirene. Den mest konservative bokmålssiden assosieres gjerne med *riksmål*. Forenklet kan vi si det slik at riksmålsforkjemperne helst unngår alle rettskrivningsreformer som får bokmål til å ligne nynorsk. Den radikale bokmålsfløyen vil derimot akseptere ordforråd og bøyningsformer som ligger nært opp til nynorsk. De fleste bokmålsbrukerne kan lokaliseres til et sted mellom riksmål og radikalt bokmål, men det er i dag ikke mulig å karakterisere flertallsnormen(e) på empirisk basis. I nynorskleiren finner vi også en konservativ fløy. Den er bl.a. kjennetegnet ved at den vil bevare et nynorsk som ligner mest mulig på norrønt og Ivar Aasens opprinnelige forslag, innenfor gjeldende valgfrihet. Den andre fløyen vil ha et nynorsk som både ligger nærmere opp til radikalt bokmål og observert språkbruk i dialektene.[5] Det finnes et slags kompromiss mellom (radikalt) bokmål og (moderat) nynorsk, og det kalles gjerne *samnorsk*. Samnorsk kan karakteriseres som unionen

[5] Denne leiren benevnes ofte som "moderat nynorsk", men "moderat" nynorsk må ikke forveksles med det vi her kaller konservativt nynorsk.

av (radikalt) bokmål og (moderat) nynorsk, gjerne med så lite normering som mulig.

På denne bakrunn kan vi systematisere skriftlig norsk som seks ulike varianter som vi skal kalle *stilnivåer*: riksmål, moderat bokmål, radikalt bokmål, samnorsk, moderat nynorsk og konservativt nynorsk. Dette er en grovklassifisering, og man kan sikkert argumentere for andre inndelinger. Likevel vil vi i det følgende holde oss til denne taksonomien.

3.4 Morfologiske stilnivåer

Stilnivåene bør ikke rotes sammen. Følgende eksempel blander sammen egenskaper ved alle stilnivåene:

(1)
Enhver offentleg tenestemann har høve til å leite efter sproglige inkonsistensar som me finn i sume tekster.

Vi vil nok aldri finne et så ekstremt tilfelle av stilsammenblanding som i eksempel (1). De aller fleste skrivekyndige nordmenn har en ubevisst eller innlært stilsans. Selv om ingen kan forby noen å sette sammen ord fra ulike skriftspråksvarianter som vist ovenfor, vil sammenblandinger fortone seg som komiske, og det er lett å la seg irritere av slikt. Gjeldende normering av bokmål tillater ikke ord som *sume*, *me* og *tenestemann*. Derimot er det tillatt å formulere seg slik i henhold til gjeldende bokmålsnorm:

(2)
Enhver offentlig tjenestemann har anledning til å leite etter språklige inkonsistenser som vi finner i somme tekster.

Det er grunn at til å anta at nordmenn flest vil reagere på ordvalget i en slik setning. Ord som *enhver* og *anledning* passer dårlig sammen med *leite* og *somme*. Men vi kan ikke referere til noen definerte normer som er brutt, bortsett fra at ord som *leite* og *somme* ikke er aksepterte riksmålsord.

3.5 Stavemåter og stilnivåer

Det kan også identifiseres systematiske valg av stavemåter som kan knyttes til de ovenfor nevnte stilnivåene. Verbene *støype* og *støpe* er sidestilte former i bokmål i den forstand at de er sidestilte staveformer av samme leksem. Videre er stavemåten *røyke* en hovedform, mens *røke* er sideform.[6] Dersom man velger å bruke alternativet *røke*, bør man også velge *støpe*, og man bør også holde seg til preteritumsformen *strøk* og ikke *strauk*, og man bør unngå *anklaga* og *lova*.

3.6 Leksemer og stilnivåer

Et leksem som substantivet *beriktigelse* bør ikke kobles sammen med radikalt bokmål, mens verbet *skeise* (å gå på skøyter) ikke hører naturlig sammen med riksmål og moderat bokmål.

4. Korpora og deduksjon av skriftstandarder

Vi har sett at bøyningsformer, stavemåter og leksemvalg kan, og etter vårt syn *bør*, systematiseres. Gitt den norske språksituasjonen kan man ved hjelp av korpusmateriale være interessert i å gjøre følgende:

- Beregne frekvensen til godkjente leksemer
- Beregne hvilke stavemåter som foretrekkes dersom det er valgfrihet (*støpe* eller *støype*)
- Beregne forholdet mellom hovedformer og sideformer (*røyke* eller *røke*)
- Identifisere nye leksemer
- Finne ut hvilke bøyningsalternasjoner som er mest frekvente der det er valgfrihet (*strøk/strauk*, *lovet/lova/lovde/lovte*)

Korpusmaterialet kan gjøre det mulig å etablere *empiriske* kriterier for hva som mest naturlig skal være leksikalske og morfologiske hovedformer. Resultatene vil naturligvis være avhengig av hvilket korpusmateriale som legges til grunn. Derfor vil spørsmål angående representativiteten til korpusmaterialet være meget viktige.

4.1 Frekvensen av tillatte leksemer

---

[6] At to former eller leksemer er sidestilte, også kalt jamstilte, betyr at begge er fullt ut aksepterte. En sideform er derimot marginalisert og skal ikke brukes i godkjenningspliktige skolebøker eller i offentlig forvaltning.

Ved hjelp av store taggede korpora er det mulig å beregne frekvensen til leksemer som er med i ordlisten. Dette forutsetter at taggeren tar med leksemkodene i merkingen av ordene i teksten. I NorKompLeks er alle lemmaer utstyrt med numeriske identifikasjonskoder, og det skal derfor være mulig å etablere en akseptabel frekvensordliste for norsk basert på leksemenes frekvens (til forskjell fra "naive" frekvensordlister som simpelthen teller opp frekvensen til bøyde ord i løpende tekst), naturligvis forutsatt at taggeren benytter NorKompLeks-ordlistene.

4.2 Frekvensen til sidestilte bøyningsformer

En mer avansert utnyttelse av korpuset er å foreta morfologiske analyser av de bøyde formene for på den måten å identifisere hvilke bøyningsalternativer som benyttes. Dermed kan man på *empirisk* grunnlag undersøke bruken av sidestilte (jamstilte) bøyningsalternativer, f.eks. alternasjonen *strøk/strauk*. Merk at man kan både beregne den totale bruken av en bestemt bøyningsklasse, og bruken av en bestemt klasse i relasjon til en annen i et bestemt leksem.

4.3 Sideformer og hovedformer

Gitt at man ved hjelp av korpus kan beregne frekvensen til sidestilte former, kan man også finne ut i hvilken grad en hovedform virkelig er mer utbredt enn sideformene.

4.4 Leksemer

Det er liten tvil om at bokmål tillater mange leksemer som hovedformer selv om de i praksis sjelden eller aldri er i bruk. Korpusmateriale setter oss i stand til å identifisere slike leksemer, ganske enkelt ved å inspisere frekvensopptellingen på leksemnivå, jf. 5.1 ovenfor.

4.5 Empirisk motiverte skriftstandarder

Kombinasjonen av korpora og maskinleselige ordbøker setter oss i stand til å identifisere skriftstandarder på empirisk grunnlag. Troverdigheten av slike standarder er direkte avhengig av tekstutvalget som bygger opp korpusmaterialet. Man kan se for seg et par hovedstrategier. Den ene er å samle sammen alt tilgjengelig elektronisk tekst-

materiale for bokmål og beregne frekvenser og foreta opptellinger på dette materialet. En annen strategi kan være å lage flere tekstsamlinger som tar hensyn til omtrentlige standarder slik de er definert ovenfor, eventuelt kombinasjoner av dem. Den første strategien er risikabel i den forstand at tekster som favoriserer en skriftstandard, kan bli overrepresentert. Store og balanserte korpora er et bedre alternativ, men det vil ta tid å lage slike. Strategi nummer to vil få problemer med å bestemme hvor ulike tekster skal kategoriseres.

4.6 Empirisk motiverte skriftstandarder

Vi har så vidt vært inne på at leksemvalg, valg av bøyningsvarianter, sideformer og hovedformer er med på å bestemme det vi har kalt stilnivåer. Det vil være svært vanskelig å definere slike stilnivåer på en måte som skaper tilstrekkelig oppslutning til at de tas i bruk. Derimot vil balanserte korpora kunne utnyttes til å utlede slike stilnivåer. Dette stiller naturligvis strenge krav til sammensetningen av tekstene, men som metodisk prinsipp er en slik empirisk tilnærming meget interessant.

5. Konklusjoner

Korpusbasert identifikasjon av skriftspråksstandarder vil være et nytt redskap for norsk skriftspråksnormering, forutsatt at det finnes et språkpolitisk klima som aksepterer at slik aktivitet. Resultatene vil gjøre det relativt enkelt å lage elektroniske ordlister for de stilnivåene man finner. Slike ordlister kan i sin tur inkorporeres i tekstbehandlingssystemer, og vi vil kunne få programmer for stavelseskontroll som er følsomme for ulike stilnivåer. Dette vil utvilsomt være et kjærkomment hjelpemiddel for folk som har problemer med å beherske skriftlig norsk, f.eks. utlendinger og folk med lese- og skriveproblemer.

Når det gjelder norsk leksikografisk arbeid, vil slike nye muligheter sette oss i stand til å produsere ordlister med nyttig informasjon til brukerne, f.eks. ved at stilistiske standarder tas med i fremtidige utgivelser av norske ordbøker, samt at frekvensinformasjon kan inkluderes.

Referanser

Johnsen, Lars/Anneliese Pitz/Lars Hellan 1989: *TROLL (The Trondheim Linguistic Lexicon Project).* Lingvistisk institutt, NTNU, Trondheim.

McEnery, Tony/Andrew Wilson 1996: *Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics, Edinburgh University Press, Edinburgh.

Nordgård, Torbjørn 1996: *NorKompLeks 1.1*. Manuskript, Lingvistisk institutt, NTNU, Trondheim

Shieber, Stuart 1986: *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes nr. 4, Stanford University.