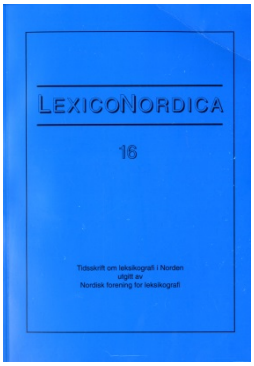
LexicoNordica

Titel: Leksikografisk dokumentasjon av flerordsenheter i norsk

Forfatter: Ruth Vatvedt Fjeld

Kilde: LexicoNordica 16, 2009, s. 103-118

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

Leksikografisk dokumentasjon av flerordsenheter i norsk

Phraseological units are inadequately described in Norwegian lexicography. There exists no dictionary of constructions for Norwegian, and in monolingual standard dictionaries phrases and collocations are occasionally registered as sublemmas or examples of one of the words in the multiword unit. This article presents an automatic analyzer for finding phraseological units in a POS-tagged corpus, called DeepDictLexifier, through comparing some of its results with the description of multi word units in traditional Norwegian dictionaries. The comparison indicates that DeepDictLexifier is a valuable tool in mapping out collocations and other phraseological units in a language.

1. Innledning

Hvilke ord som kan kombineres med andre ord til større språklige enheter, er regulert av seleksjonsregler på tre nivåer. De har forskjellig relevans i leksikografisk arbeid:

a) Universelle ensyklopediske restriksjoner er begrunnet i hvordan verden faktisk er. For eksempel kan man ikke si **steinen smiler* i normal kommunikasjon. De ensyklopediske restriksjonene gjelder for alle språk, og hører neppe hjemme i leksikografisk dokumentasjon, selv om den selvsagt kommer fram indirekte gjennom brukseksemplene.

b) Universelle grammatiske restriksjoner regulerer syntagmatiske egenskaper ved leksemene og er begrunnet i kommunikative behov, for eksempel at en setning normalt må ha markering av subjekt og verbal. Disse restriksjonene trenger man ikke dokumentere leksikografisk.

c) Språkspesifikke restriksjoner av grammatisk eller tilfeldig konvensjonalisert art. Disse restriksjonene gir hel- eller halvfaste fraser og utgjør språks idiomatikk. De språkspesifikke grammatiske restriksjonene blir dels redegjort for i grammatikker, de leksikalsk-konvensjonaliserte er typisk en leksikografisk oppgave. De beskrives dels i tospråklige ordbøker og i noen grad i dokumenterende enspråklige ordbøker, eller i egne konstruksjonsordbøker.

Reglene for orddanning er svært sterke. De gir mange flere uttrykksmuligheter enn vi trenger eller finner det hensiktsmessig å bruke. Dermed blir en del potensielle ordformer lite eller ikke brukt i det hele tatt, mens andre blir mye brukt. Det er leksikografiens oppgave å redegjøre for de vanlig brukte ordtypene, ikke bare de mulige. Leksikografer legger stort arbeid i å observere språkbruk og kartlegge bruken av enkeltord.

De konvensjonaliserte flerordsenhetene, de som utgjør et språks idiomatikk og som vi gjerne kaller fraseologi, har det vært mindre vanlig å redegjøre for, i hvert fall i norsk leksikografi. Med moderne databehandling som metode er det nå lettere å registrere og kartlegge flerordsenheter i språket, og denne artikkelen skal redegjøre for et forsøk med automatisk kartlegging av en type flerordsenheter i moderne norsk.

2. Hva er flerordsenheter?

Mye brukte sammenstillinger av flere ord gir det man i dagligtale kaller faste fraser eller stående vendinger. Det er mange typer av dem. I leksikografi kalles de gjerne med en fellesterm for flerordsenheter. De fungerer som en slags prefabrikkerte elementer i språket, da språk er svært komplekse med uendelig mange kombinasjonsmuligheter, vil konvensjonalisert avgrenset ordvalg gjøre det lettere å ordlegge seg raskt. Skal man forklare hvorfor et flerordsuttrykk er nettopp slik det er, har vi ofte ikke annet argument enn: ”Det er slik vi ”pleier å si det””! Det er f.eks. vanskelig å forklare andrespråksinnlærere hvorfor én kombinasjon er ”riktigere” enn en annen i et flerordsuttrykk, for som oftest finner vi ikke noen regel eller semantisk eller leksikalsk forklaring. Flerordsenheter har begrenset ordinventar der det hadde vært mulig å bruke andre ord uten å bryte ensyklopediske, semantiske eller grammatiske restriksjoner. Ofte får konvensjonaliserte flerordsenhetene en betydning som ikke er tolkbar ut fra betydningen av enkeltordene i enheten.

Flerordsenheter er faste og halvfaste fraser som har utviklet seg gjennom et språks historie og de er dermed språkspesifikke. De er forbindelser der det er en viss affinitet mellom ordene på semantisk, morfologisk eller syntaktisk grunnlag. Det er ikke bare på ordnivå denne affiniteten forekommer, språk kan ha visse morfologiske preferanser der flere er mulig. Reglene trenger ikke engang være forskjellige bare for forskjellige språk, det kan være egne restriksjoner for forskjellige varieteter innen et og samme språk, som mellom bokmål og nynorsk, eller mellom forskjellige stillag.

Flerordsenheter er annerledes enn frie kombinasjoner, der leddene kan byttes ut og enkeltordene i frasen kan defineres helt uavhengig av hverandre. Men både leksikalske og grammatiske restriksjoner kan endres over tid. Cruse (1991:29) nevner at de er stabile i et gjennomsnittlig menneskeliv (i vår tid i Norge ca. 80 år). Rutineformler er kanskje de mest faste flerordsenhetene vi har, men også med dem kan det skje endringer. For eksempel var *gledelig jul* en temmelig uvariabel rutineformel for rundt hundre år siden, men i dag regnes den som avleggs; det normale eller nøytrale uttrykket nå er *god jul*. I det senere er det også dukket opp en ny variant, nemlig *fredelig jul*, sannsynligvis en konsekvens av allmenn sekularisering i det moderne Norge. Kanskje vil den fortrenge *god jul* om noen tiår. Hilseformelen *god dag* høres også stadig sjeldnere, i dag sier de fleste *hei* når de treffer hverandre, og *morna* som avskjedshilsen er nesten helt fortrengt av *ha det* (også skrevet *hade(t)* i unormert skriftspråk, noe som antyder at formen er blitt leksikalisert).

Endringer i språkbrukernes vaner reflekterer allmenne kulturelle endringer. I leksikografien er det viktig å følge disse endringene nøye og dokumentere dem på samme måte som man dokumenterer enkeltords utvikling. De grammatiske endringene i f.eks. kasus, valens og argumentstruktur blir dokumentert i indre språkhistorie og grammatikker, mens å dokumentere de leksikalske og idiomatiske endringene er leksikografiens oppgave.

3. Hvordan er flerordsenheter behandlet i norsk leksikografi?

Det fins ikke noen norsk konstruksjonsordbok eller fraseologisk ordbok, slik man har for de fleste andre europeiske språk, bl.a. svensk og dansk. Det vi har av ordbøker over flerordsenheter i norsk, er ordspråksordbøker og sitatordbøker, som *Bevingede ord* (Evensberget og Gundersen 2006), og en lang rekke ordbøker over idiomatiserte, faste uttrykk, som *Katta i sekken og andre uttrykk* (Vannebo 2006). Denne typen ordbøker er imidlertid helt spesielle kulturhistoriske eller populærfaglige beskrivelser og ikke egnet for forståelse av hvordan ordene spiller sammen i større leksikalske enheter i språket. Det mangler en beskrivelse av systematisk utvalgte flerordsenheter i norsk.

Det er imidlertid tatt med et stort antall flerordsenheter i de større vanlige norske ordbøkene, både enspråklige og tospråklige. Men det kan synes tilfeldig hvilke som er tatt med, om de er ført opp som brukseksempler på ett av ordene i enheten eller som egne sublemmaer. Det er

heller ikke lett å se begrunnelsen for hvilket lemma de er oppført under eller forklart med.

Olsen (2001) har gjort en studie av flerordsenheter i noen norske ordbøker, og konkluderer med at de er konsentrert om idiomatiske uttrykk. Hun fant at flerordsenheter ofte var markert diasystematisk med *overf*. eller *i uttr*. om betydningen, uten at hun fant noe entydig system for bruken av de forskjellige markeringene. Hun kritiserer ordbøkene for at de heller tar med idiomer enn kollokasjoner, selv om idiomene utgjør en forsvinnende liten del av det leksikalske inventaret i forhold til andre flerordsenheter. Dette begrunner hun slik:

> Språket kan [dermed] snarere betegnes som et system av flerordsenheter enn som et system av enkeltord. Denne erkjennelsen bør få konsekvenser for hvordan fraser behandles i ordbøker. Frasen som flerordsenhet bør skilles tydelig fra leksikografiske eksempler. Dette gjøres i noen ordbøker, men ikke i alle. Når frasen inngår i et eksempel, kan dette gjøres ved at orda som frasen består av, utheves med fete typer. (Eksempler står vanligvis i kursiv.) Frasens betydning bør forklares, enten der den står oppført, eller med henvisning til forklaring et annet sted, dersom hele eller deler av frasen er metaforisert. Dette er viktig for å unngå misforståelser om betydning og bruk. (Olsen 2001)

Olsen fant altså at det ikke er markert forskjell på om et uttrykk er oppført som eget sublemma, eller om det er satt som et eksempel på bruken av det lemmaet det er oppført under, og at markeringen ”overf.” skjuler forskjellen mellom eksempel og flerordslemma. Det gjenspeiles i at det sjelden er forklart hvordan markeringene er brukt, og også i hvilken grad man har vært klar over at flerordsenheter er selvstendige lemmaer som krever en egen definisjon.

I *Norsk riksmålsordbok* (band 1, 1937), som fortsatt er den mest omfattende dokumentasjonen av ordforrådet i moderne norsk, forklares flerordsenheter i en introduserende veiledning om bruk av ordboken, slik:

> Faste ordforbindelser, vendinger og ordtak er forklart under det ord som naturlig fremtrer som hovedordet, f.eks. *komme i skade for* under *skade.* (s. XV)

Hvordan man vet at noe er naturlig hovedord, er ikke forklart. Det blir heller ikke klargjort hvilke kriterier som er fulgt ved utvelgelse av de ordforbindelsene som er blitt lemmatisert, og heller ikke hva det vil si at et av leddene ”naturlig framtrer som hovedord”. Flerordsenhetene har fått markeringen ”i uttr.”, og det antyder likevel at man er klar over at

det ikke er bare enkeltordene som er viktige, men helheten. Og videre i veiledningen står det om betydningsdeling:

> Efter definisjonen i hvert betydningsnummer kommer, trykt med kursiv, eksempler og citater, først de som viser ordet i fri bruk, derefter faste forbindelser og stående vendinger hvor ordet forekommer. Når der er mange av den slags faste forbindelser, har vi så langt det praktisk lot sig gjøre ordnet dem alfabetisk og uthevet hver ny forbindelse med sperret trykk; en slik fremgangsmåte faller særlig naturlig ved en lengere rekke faste forbindelser av verbum + adverbium eller komplement (f.eks. under verbet *legge: legge an, legge av* o.s.v.) (s. XV)

Dette viser at redaksjonen i NRO har vært oppmerksom på forskjellen mellom frie og faste kombinasjoner, men det er likevel ikke klart når flerordsenheter står som eksempler på det ettordslemmaet frasen er oppført under, eller når de er selvstendige flerordslemmaer, det de kaller ”stående vendinger”. Det framgår av sitatet over at redaksjonen tenkte strukturelt og systematisk i forhold til kombinasjonen verb +adverbial/ komplement. Andre ”stående vendinger” var tilfeldig lemmatisert eller beskrevet.

I tospråksordbøker er det lagt mye arbeid i å beskrive flerordsenheter, nettopp fordi inventaret i dem er språkspesifikk. Det gjelder særlig de større tospråksordbøkene, og mye av dette kunne ha vært utnyttet i enspråklige ordbøker, der tilfeldighetene har rådet, både i handordbøkene og i de større, dokumenterende enspråklige ordbøkene. Grunnen til det har vært manglende teori og metode for kartlegging og utvelgelse av mer eller mindre frie ordforbindelser og flerorsenheter.

4. Kategorisering av flerordsenheter

Flerordsenheter kan avgrenses eller defineres på forskjellige måter, ut fra syntaktiske og semantiske prinsipper, eller ganske enkelt etter bruksfrekvens. Det er terminologisk forvirring og uklare grenser mellom typer av flerordsenheter i leksikografisk faglitteratur. Jeg vil her følge Svensén (2004:206), som skiller mellom faste og ikke-faste ordforbindelser, og kaller de faste forbindelsene idiomer. De ikke-faste skiller han i frie kombinasjoner og kollokasjoner.

Frie kombinasjoner er flerordsforbindelser der det ikke er restriksjoner mellom leddene ut over de generelle seleksjonsreglene. Ord som opptrer i kollokasjoner, har begrenset kombinerbarhet med andre ord, det råder en slags affinitet mellom ordene i en slik flerordsenhet. Hoved-

ordet i en kollokasjon kalles base, og ord som basen gjerne tiltrekker seg, kalles kollokator. Enkeltordene i en slik enhet har en betydning som er begrenset av de ordene de står sammen med, det Cruise kaller ”collocational uniqueness” (Cruse 1991:29).

I den seinere tid har kollokasjoner vært svært aktuelle i leksikografisk forskning, også i nordisk sammenheng. Blant annet fins en oversikt hos Svensén (2004:212f) over grammatisk definerte kollokasjoner for svensk. Om man antar at mønstrene er omtrent de samme i norsk, kan man regne med følgende hovedtyper:

Kollokasjonstype	***Eksempel***	***(i motsetning til)***
1. Verb + substantiv (obj.)	begå mord	*gjøre mord
	betro/fortelle en hemmelighet	*gi en hemmelighet
2. Substantiv (subj.) + verb	dagen gryr	*dagen lysner
3. Adjektiv + substantiv	høy mann	*lang mann
	dyp sorg	*høy sorg
4. Verb + adverb	betvile sterkt	*betvile grundig
5. Adverb+adjektiv	overstadig beruset	*overstadig glad

Hvordan disse kollokasjonsmønstrene passer for moderne norsk bokmål, skal undersøkes systematisk i Leksikografisk bokmålskorpus (Fjeld 2008). Korpuset er nå på 40 mill. ord fra perioden 1985–2008, som er balansert med hensyn til teksttyper og sosiologiske variabler hos forfatterne.

5. Semiautomatisk metode for kartlegging av kollokasjoner

Det er i dag utviklet semiautomatiske metoder for å undersøke kollokasjoner, den mest kjente er WordSketchEngine (Kilgarriff & Tugwell 2000), men det fins også andre. Spesielt tilpasset de skandinaviske språkene er DeepDictLexifier (DD), utviklet av Eckhard Bick ved Syddansk universitet (Bick 2009), som er tilpasset norsk av Lars Nygaard. Det er en kolloksasjonsanalysator som er utviklet som et komplement til oversettingsprogrammet GramTrans. GramTrans er spesielt utviklet verktøy for oversettelse mellom de skandinaviske språkene. DeepDictLexifier gjør det mulig å bygge komplekse ordboksoppføringer med kontekstinformasjon ut fra søk etter statistisk beregnbare relasjoner mellom ordene i et korpus. Ordrelasjonene dokumenterer ikke bare at ordene opptrer sammen rent statistisk, ved hjelp av constraint grammaranalyse og grammatiske funksjoner som ordene har i kontekst, finner det

sannsynlige kollokasjoner automatisk. Hvert ord blir bestemt ut fra absolutt og relativ frekvens i de analyserte tekstene. En demoversjon basert på tekster fra Wikipedia (http://gramtrans.com/deepdict/) ligger fritt tilgjengelig på Internett.

Når man søker på et ord i DD, får man statistiske opplysninger over hvilke ord som opptrer sammen med søkeordet, f.eks. om man søker opp ordet *hus,* får man bl.a. tallene 6,24:5 foran kollokatoren *leie.* Det betyr at et vanlig verb foran ordet *hus* er *å leie.* Tallet foran kolon er et uttrykk for hvor sterk relativ binding det er mellom leddene i kollokasjonen i det analyserte materialet (her mellom *leie* og *hus*). Tallene angir en relasjon utregnet etter forholdet mellom hvor frekvente ordene er totalt sett i korpuset, og hvor ofte de to ordene opptrer sammen. Jo høyere tall, jo sterkere binding. Tallet etter kolon angir et logaritmisk tall for hvor høy forekomsten av kollokasjonen er i det analyserte korpuset, omformet til forekomstklasser. Dersom forekomsten er klassifisert til over 3, blir den uthevet med halvfeit stil for å gi en antydning om at dette er en sannsynlig kollokasjon.

DeepDictLexifier gir automatisk oversikt over alle kollokasjonstypene som Svensén har satt opp for svensk. Leksikografisk bokmålskorpus er analysert automatisk ved hjelp av DD-programmet, og her presenteres noen eksempler på resultatene det har gitt. Resultatene kontrolleres så mot beskrivelse av de samme kollokasjonene i *Bokmålsordboka* (BOB) og *Norsk Riksmålsordbok* (NRO). Fordi korpuset er forholdsvis lite, kontrolleres resultatet i noen tilfeller også mot søk på Google og i Norsk Aviskorpus, et stort norsk aviskorpus på 700 mill. ord (http://avis.uib.no/).

5.1. Verb + substantiv (objekt)

For substantiv er det angitt hvilke (post- eller pre-)modifikatorer det kan ha, og hvilke verb som brukes sammen med det. Her er valgt substantivet **lit** som første eksempel:

lit:
man kan
6,93:4 **feste** · 7,72:3 romme · 1,17:6 **sette** · 4,12:2 drikke · 4,86:1 puste
16,51:1 kry på · 4,49:1 yte som · 3,56:1 notere på

DD-resultatet viser for det første at noen av tilslagene er typisk støy. Det skyldes blant annet at korpuset som er analysert, er ordklassetagget automatisk, f.eks. har ordformen *liter* blitt tolket som flertallsform av *lit*.

Men det er ikke flere feil enn at det er raskt gjort å fjerne eller overse slik støy fra tilslagene. Ellers ser vi at vi de to mest frekvente verbene som brukes om *lit* i korpuset, er *feste* og *sette.*

En fin tilleggsfunksjon i selve analysatoren er at man kan klikke på resultattallet og få opp en konkordans over tilslagene i korpuset. Det gjør det greit å vurdere verdien av søkeresultatet:

bok-SK01MiMa01-2046095	Han **festet** ikke lenger **lit** til sin egen stemme .
bok-AV04TU9606-423261	Han **fester lit** til utsagnet om at Kjell Inge Røkke tenker langsiktig når det gjelder Aker . -

Vi tar med bare to eksempler her, men alle tilslagene kommer opp automatisk; i dette tilfellet 30 belegg, som det er raskt å gå gjennom manuelt.

Resultatet av den automatiske analysen samsvarer med fraseologien som er angitt i BOB: *sette sin l- til noe*(*n*) stole på noe(n) / *feste l- til* stole på . NRO har ført opp flere verb som kollokator til substantivet **lit**: *vekke* lit, *ha* lit (til), *sette* lit (til), *feste* lit (til), *slå* lit (til).

I NRO er det altså dokumentert flere verb brukt som kollokator som vi ikke finner med DD-analysen i det moderne korpuset. Ifølge NRO er frasen *slå lit til* fra et Ibsen-sitat. Sannsynligvis hadde det ikke hjulpet med et større korpus av moderne norsk for å få tilslag på den. Uttrykket er ute av bruk nå, og hører ikke hjemme i en moderne produksjonsordbok, men i en historiserende resepsjonsordbok som NRO hører den selvsagt med. Kollokasjonsegenskapene til substantivet *lit* har endret seg fra Ibsens dager til vår tid. Dokumentasjon av den historiske utviklingen av kollokasjoner i norsk er ikke gjort før, men kan ved hjelp av analysator som DeepDict lettere gjennomføres. Det er ingen treff på *slå lit til* på Google eller i Norsk Aviskorpus. Men NRO har også med kollokasjonen *ha lit*, og det er kanskje det vanligste verbet som kollokerer med substantivet *lit.* Dette viser at det analyserte vårt korpus er for lite til at alle aktuelle kollokasjoner kommer med. Et søk i norske tekster på Google viser 100 treff på frasen *ha lit til*, og 47 på frasen *sett lit til* i imperativ. Det viser at en kollokasjonsanalysator som DeepDict krever et ganske stort korpus. Vi er derfor i ferd med å utvide bokmålskorpuset fra 40 til 100 millioner ord.

Alle kollokasjonene er oppført som sitater eller eksempler i artikkelen i NRO. Hvilket verb som er vanlig eller umarkert verbkollokator for substantivet *lit*, må man dermed selv trekke ut av disse eksemplene, mens oppføringen i BOB med døde eksempler, kanskje implisitt antyder

hva som er vanlige faste vendinger. Det blir enda tydeligere i resultatene fra DD-analysen, der det er statistisk dokumentasjon over bruken. Den viser at det vanligste er at man *fester (sin) lit til noe(n),* og mer sjelden *setter man (sin) lit til noe(n)*. En fraseologisk ordbok er gjerne en aktiv ordbok eller produksjonsordbok, og da er det svært nyttig å få vite hva som er vanlig og hva som er stilmarkert eller sjelden i bruk når man lager en konstruksjon. *Slå lit til* er rett nok markert med *foreld., arkais*. i NRO, men om man i stedet skal velge *feste* eller *sette* som kollokator, og hvilke tolkningsendringer valget kan medføre, får man ingen informasjon om.

5.2. Substantiv (subjekt) + verb

Substantivet *lit* kan knapt stå som subjekt i en setning, i hvert fall er det ikke belegg på det i korpuset eller i de to ordbøkene. Jeg velger derfor å undersøke et nærsynonym, nemlig *tro* med DD-analysatoren.

tro tro kan ...
8,13:1 svinne · 7,07:2 bekjenne
2,07:2 svekke...s

Vi ser igjen at den automatiske analysen har feil, her er feilen syntaktisk, da *tro* har blitt tolket som subjekt i følgende belegg:

bok-SA09Wiki01-631967 Den katolske kirke **bekjenner** seg til den kristne **tro** slik den er uttrykt i den nikenske trosbekjennelse.

Men selvsagt kan *tro* både bekjennes/bli bekjent og svekkes/bli svekket, så med verbet i passiv stemmer det, og beleggene for *svekke* står da også i passiv:

bok-SK03StTh02-1330675 Vår **tro** på å kunne forstå vårt eget eller framtidige samfunn, er ikke blitt **svekket** med årene.

BOB har ikke med denne kollokasjonstypen for *tro*, altså hvilke verb som kan ha *tro* som subjekt. Det eneste reelle belegget i korpuset er *svinne*, og det er ikke uventet at en så sjelden konstruksjon ikke er med i ei handordbok som BOB.

Men NRO dokumenterer flere verb som kollokerer med *tro*:

troen går fritt; gid min tro står sterkt som grunnen, tro kan flytte bjerge

Dette er kollokasjoner som enten er foreldet og/eller bare brukes i faste og metaforiske uttrykk. Det mer normale uttrykket *troen svinner* fra korpuset, blir ikke dokumentert, heller ikke under verbet *svinne*. Beleggene i korpuset er:

bok-SK01SmKi01-1663372	Men **troen** på de allierte **svant** visst en smule hen etter det «redningstoktet».
bok-SA02LeSt01-1184506	For 28-åringen, som blir hengende etter, **svinner troen** på at det er mulig å bli popstjerne, realitystjerne, gründer, forfatter med opplagssuksess.

Det understøttes av mange hundre treff på Google og flere i Norsk Aviskorpus. DeepDict-analysatoren viser dermed at ganske vanlige kollokasjoner mangler i norske ordbøker.

5.3. Adjektiv + substantiv

For denne kollokasjonstypen kan vi også bruke *tro* som søkeord i DD. Det gir følgende kollokasjoner, og ti av dem er statistisk beregnet til å være sterke kollokasjoner:

> 5,54:6 **kristen** · 7,54:3 usvikelig · 4,88:5 **religiøs** · 2,66:6 **god** · 2,23:6 **stor** · 2,11:5 **liten** · 3,08:4 **rett** · 4,03:3 overdreven · 3,84:3 PROP-hum's · 4,53:2 urokkelig · 3,4:3 vane · 2,25:4 **sterk** · 2,11:4 **katolsk** · 2,99:3 muslimsk · 3,19:2 optimistisk · 0,67:4 **ny** · 0.57:4 **gammel** · 2.35:2 blind · 3.33:1 inderlig · 1.27:3 jødisk · 2.09:2 ortodoks · 1.02:3 mangle · 1,63:2 luthersk · 2,58:1 enfoldig

BOB har disse adjektivene ved *tro*:

> det er min faste t- at det er slik / ha god, dårlig t- på noe(n)

Her er det altså belagt kollokasjonene *fast tro, god tro, dårlig tro. Fast tro* er ikke med i korpuset, DD-analysen viser sterk kollokasjon mellom *tro* og *religiøs,* samt spesifiseringene *kristen* og *katolsk.* Ellers har korpuset kollokasjonene *stor/liten, sterk* og *rett tro,* samt *ny/gammel*. Disse kunne gjerne vært med i BOB. NRO har følgende kollokasjoner med adjektiv +tro:

> fast og uryggelig en tro; god tro, ond tro, hellig tro, kristen tro, rett tro, arvet tro, gammel tro.

De fleste beleggene refererer til religiøs tro, og ikke til andre religioner enn kristendommen. Man må lete gjennom en spalte som går over vel en side og spredte sitater fra litterære kilder for å finne disse beleggene, noe som viser tydelig at NRO ikke er en produksjonsordbok.

5.4. Verb + adverb

Særlig ved deleksikaliserte, betydningsomfattende verb opptar kollokasjonstypen stor plass i ordbøker. En gransking av artiklene for verbet *komme* i *Norsk Ordbok, Norsk Riksmålsordbok* og *Norsk-engelsk stor ordbok* viser store forskjeller i valg av kollokasjoner som er tatt med. En analyse ved hjelp av DD viser at mange vanlige adverb-konstruksjoner med *komme* i moderne norsk likevel ikke er dokumentert i disse ordbøkene i det hele tatt (Fjeld, Nygaard og Bick 2009). Jeg vil her illustrere kollokasjonstypen med to litt mer leksikaliserte verb, nemlig *brøle* og *skrike*. Analyse med DD gir følgende kollokasjoner for *brøle* + adverb:

brøle

5,4:3 høy · 4,69:1 taktfast · 2,67:2 nå · 1,26:2 inn · 1,87:1 litt · 1,83:1 alltid · 1,37:1 nesten · 1,25:1 tilbake · 1,22:1 vid · 0,48:1 derfor · 0,25:1 bare

Verken BOB eller NRO har eksempler på denne kollokasjonstypen ved verbet *brøle*. DD-analysen markerer ikke noen av kollokasjonene med høy frekvens, men beleggene viser at *brøle høyt* har 9 treff. Et større korpus ville sannsynligvis ha gitt et tydeligere resultat, og det er den naturlige kollokasjonen som gjerne kunne vært med i en ordbok. Det er ikke idiomatisk norsk å si *brøle sterkt, brøle hardt*, det heter *brøle høyt*.

For verbet *skrike* gir DD følgende resultat:

skrike

7,95:5 **høy** · 8,03:2 formelig · 7,62:2 fæl · 7,03:2 stygg · 5,93:3 plutselig · 5,48:2 sånn · 3,14:3 igjen · 2,93:3 slik · 2,88:3 tilbake · 2,84:3 bare · 4,32:1 innvendig · 4,19:1 i søvne · 3,08:2 litt · 4,05:1 rasende · 2,82:2 hvorfor · 2,54:2 ute · 3,5:1 vond · 3,38:1 der oppe · 3,38:1 for eksempel · 2,23:2 kanskje · 3,21:1 lav · 2,06:2 like · 1,92:2 alltid · 1,67:2 nå · 2,66:1 engang

Det er 30 treff på kombinasjonen *skrike+høy*. At verbet *skrike* har flere treff enn *brøle* i kollokasjon med *høy*, er ikke rart, da det allerede ligger i semantikken til *brøle* at det er med stor styrke, og dermed er det ikke nødvendig å legge til et adverb som markerer det. Bokmålsordboka har bare med en adverbkollokator til *skrike*: s- høyt av *smerte*, og det sam-

svarer altså med resultatet fra DD. NRO har også med *skrike høit av smerte,* samt *skrike håst.* DD-analysen ga ikke noe nytt i forhold til ordbøkene her, og den viser at begge ordbøkene har dokumentert den sterkeste kollokasjonen ved verbet *skrike*, noe som underbygger antakelsen om at verktøyet viser de beste kollokasjonene i det analyserte materialet.

5.5 Adverb+adjektiv

For denne kollokasjonstypen har jeg valgt adjektivet *beruset*, som også er med blant Svenséns eksempler for svensk. Analyse med DD gir følgende resultat:

> 10,34:2 overstadig · 6.43:2 synlig · 5,04:1 åpenbar · 3,42:2 lett · 3,59:1 tydelig · 2,42:2 sterk · 2,96:1 temmelig · 1,49:2 så · 1,9:1 litt · 0,18:1 for

BOB har *være synlig, overstadig b-,* som er i godt samsvar med DD-analysen, med *synlig* og *overstadig* som de mest frekvente. Men fordi dette er et lite vanlig adjektiv, får vi ikke mange nok treff i korpuset til at kombinasjonen blir merket som sterk kollokasjon. Og det kan godt hende at adverbene *åpenbar, temmelig* og *lett* kunne vært med, men vi trenger et større korpus for å vise dette.

Foreløpig kan vi sammenlikne resultatene fra DD med søk mot Norsk Aviskorpus og Google. Både Google (10.100 treff) og aviskorpuset (912 treff) har flest belegg på *overstadig beruset*, mens *åpenbart beruset* kommer på andreplass i resultater fra Google (2550 treff) og *sterkt beruset* i aviskorpuset (443 treff). Det er sannsynlig at en DD-analyse av et tilstrekkelig stort korpus ville vist tydeligere hva som er vanligst i forskjellige teksttyper og stillag. Her kommer også problemet med synonymi inn, det er bare der de forskjellige kollokatorene uttrykker samme meningsinnhold, at det er interessant å sammenlikne resultater fra forskjellige korpus. I dette tilfelle er *åpenbart* og *synlig* omtrent likeverdige, og til en viss grad *overstadig* og *sterkt*, og *lett* og *litt*.

Adjektivet *syk* er mer vanlig, og er også med i Svenséns eksempler. I DD-analysen får vi følgene kollokatorer for dette adjektivet:

> 7,81:5 **alvorlig** · 3,44:5 **hel** · 2,79:5 **så** · 3,75:4 **veldig** · 5,27:2 akutt · 4,72:2 kronisk · 5,59:1 uhelbredelig · 4,58:2 psykisk · 2,45:4 **for** · 4,42:2 mental · 2,09:4 **svær** · 1,71:4 mye · 3,55:2 skikkelig · 3,3:2 ordentlig · 2,08:3 litt ·

> 3,69:1 dødelig · 0,98:3 meget · 2,29:1 fysisk · 2,13:1 altfor · 1,54:1 temmelig · 0,13:1 virkelig · 0,09:1 enda

BOB har bare en kollokator her, nemlig *alvorlig s-,* som jo også er den mest frekvente. Det understøttes av Norsk Aviskorpus, som gir 1781 treff på kollokasjonen. NRO har bare med *dødelig syk*, og den er også med i DD-analysen, men med lav frekvens. Kollokasjonen gir 402 treff på Google og 27 treff i Norsk Aviskorpus, og det antyder at den etter hvert er sjelden. Den moderne uttrykksmåten her er *dødssyk*, med 14.600 treff på Google og 572 i Norsk Aviskorpus. En slik forskjell kommer automatisk fram med DD-analysen.

5.6. Partikkelverb og preposisjonsledd

Kollokasjoner med verb+adverb kan være frie temporale, lokative eller modale adverb (jf. 5.4 ovenfor), men de kan også være valensbundne adverbialkomplement til partikkelverb. Svensén har ikke med partikkelverb i sin oppstilling over det han kaller leksikalske kollokasjoner, men kategoriserer dem som grammatiske kollokasjoner og regner dem som en valensforbindelse. I tradisjonelle ordbøker er det imidlertid brukt mye plass på å registrere og beskrive slike konstruksjoner. Men det er gjort svært lite forskning på partikkelverb i norsk, og i tradisjonelle ordbøker er de sjelden skilt ut fra konstruksjoner med frie adverbialer. En gjennomgående analyse av resultatene fra DD vil kunne gi ny innsikt i og oversikt over denne konstruksjonstypen. DD gir f.eks. følgende verbpartikler ved *skrike*:

> 5,17:6 **ut** · 4,62:5 **opp** · 4,77:1 vill · 1,82:1 inne

DD viser at det ikke er verbalpartikler med *brøle* som grunnverb i korpuset, mens BOB har med *brøle ut* og *skrike op.* NRO har med både *skrike i, skrike op,* men ingen med *brøle*. Konstruksjonen *skrike i* med betydningen 'sette i et skrik' er foreldet i moderne norsk. Det er f.eks. ingen belegg på den konstruksjonen i Norsk Aviskorpus.

Preposisjonsledd til verb er også med i stort omfang i ordbøkene, uten at det er klart når de står i frie eller mer faste kombinasjoner. Om vi holder oss til preposisjonsledd ved de to samme verbene, finner vi med DD-analyse:

skrike

skrike til ...	6,6:3 PROP-hum · 0,28:3 <Hfam> · 0,72:2 slutt · 1,2:1 gutt · 0,47:1 hund
skrike av ...	5,78:3 smerte · 6,19:2 latter · 5,46:2 redsel · 6,03:1 fryd · 4,04:2 glede · 3,58:2 hals · 1,19:4 <f-psych> · 2,74:2 kraft · 3,32:1 sinne · 2,89:1 angst · 1,36:2 <percep-taste> · 2,19:1 munn · 0,57:2 <percep-l>
skrike gjennom ...	4,81:1 larm
skrike om ...	4,17:1 rasisme · 2,36:2 hjelp · 1,5:1 oppmerksomhet · 1,04:1 natt
skrike som ...	4,05:1 gris · 1,17:1 unge
skrike etter ...	2,89:2 oppmerksomhet · 2,69:2 svar · 1,87:1 mat · 1,54:1 jente · 0,91:1 kraft
skrike mot ...	3,88:1 asfalt · 2,68:1 himmel
skrike med ...	2,71:2 stemme · 2,19:1 munn · 0,57:2 <percep-l> · 0,09:1 øye
skrike til ...	3,38:1 PROP-hum
skrike uten ...	1,93:1 lyd
skrike bak ...	

brøle

brøle av ...	8,21:3 latter · 5,44:1 sinne · 2,59:3 <percep-l> · 3,96:1 glede · 2,22:1 kraft · 0,39:2 <f-psych> · 0,83:1 <percep-taste>
brøle mot ...	4,69:1 PROP-hum · 4,34:1 kjøkken
brøle til ...	0,75:2 <Azo> · 1,45:1 slutt · 0,6:1 <Adom>

Dette er alle sammen vanlige preposisjonsledd (bortsett fra noen automatiske analysefeil).

BOB har følgende preposisjonsledd under *skrike:* skrike *av* / *om* / *i munnen på hverandre* / *på* / *rundt* og under *brøle: b- av latter, smerte.* NRO har under *brøle:* brøle paa barnedejerne, og under *skrike* en lang rekke preposisjonsledd innledet med *over, for, fra, etter, om, på til, under, i, imot.* Det ligger et stort arbeid i å finne belegg på alle typene, noe en DD-analyse gir greit automatisk. Den gir samtidig et statistisk mål på kombinasjonen, som kan antyde om den er en kollokasjon eller ikke.

6. Oppsummering

Eksisterende ordbøker for norsk har dårlig eller ingen oversikt over flerordsenheter, særlig er typen kollokasjoner tilfeldig dokumentert. Det statistisk baserte analyseprogrammet DeepDict Lexifier brukt på et ordklassetagget og rimelig stort korpus vil gi et godt grunnlag for å utarbeide en oversikt over de mest vanlige kollokasjonene i norsk. En sammenlikning med resultater fra dette programmet og håndordboka Bokmålsordboka og den mer omfattende Norsk Riksmålsordbok viser at DeepDictLexifier ser ut til å være et godt verktøy for en gjennomgående revisjon av Bokmålsordbokas konstruksjonsopplysninger. I tillegg til bedre dekning kan dette verktøyet også gi innsikt i hva som er mest vanlig språkbruk i moderne norsk og dermed gjøre den til en bedre produksjonsordbok enn den er nå. Det samme gjelder større ordbøker som Norsk Riksmålsordbok, som har en artikkelstruktur og informasjonsmengde som gjør dem vanskelige å bruke som produksjonsordbøker.

7. Litteratur

Bick, Eckhard 2009: DeepDict – A Corpus-based Dictionary of Word Relations. I: *Proceedings of Nodalida 2009, Odense, Denmark.* NEAL proceedings series, Vol. 4, 268–271.

Cruse, D.A. 1991: *Lexical Semantics.* Cambridge, New York: Cambridge University Press.

Norsk-engelsk stor ordbok. 2001: Oslo. Kunnskapsforlaget.

Gundersen, Dag & Snorre Evensberget 2006: *Bevingede ord: ordtak, sitater og deres opprinnelse.* 4. utg. Oslo: Kunnskapsforlaget.

Kilgarriff, Adam & David Tugwell 2002: Sketching Words. I: Correard, Marie-Hélène (ed.), *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*, 125–135.

Malmgren, Sven-Göran 2008: Collocations in Swedish Dictionaries and Dictionary Research. I: *Lexicographica* 14, 149–158.
Norsk Ordbok 1966 –. Oslo: Det norske Samlaget.
NRO =*Norsk Riksmålsordbok* 1937–1957. Oslo: Aschehoug.
Nygaard, Lars, Joel Priestley, Anders Nøklestad and Janne Bondi Johannessen 2008: Glossa: Multilingual, Multimodal, Configurable User Interface. I: *Lrec 2008, European Language Resources Association (ELRA)*. Marrakech.
Olsen, Tone Rudi 2001: *Leksikografisk behandling av fraser med overført betydning*. Hovedoppgave INL 2001 http://www.duo.uio.no/sok/work.html?WORKID=2147
Rommetveit, Magne 2007: *Med andre ord*, *Den store synonymordboka med omsetjingar til nynorsk* 3. utgåva. Oslo: Det norske Samlaget.
Svensén, Bo 2004: *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. Stockholm: Norstedts Akademiska Förlag.
Svenskt språkbruk. Ordbok över konstruktioner och fraser. Stockholm: Norstedts 2003.

Ruth Vatvedt Fjeld
Professor
Institutt for lingvistiske og nordiske studier
Universitetet i Oslo
Box 1001 Blindern
N-0315 Oslo
r.e.v.fjeld@iln.uio.no