Novus forlag · eISSN 1890-5455
DOI: https://doi.org/10.52145/mom.v114i2.2122
Maal og minne 2-2022 35–66

Kollokasjonar – språkets R2-D2 og C-3PO

Av Gunn Inger Lyse, Margunn Rauset og Helge Dyvik

I revisjonen av *Bokmålsordboka* og *Nynorskordboka* treng leksikografane reiskapar som effektivt og etterretteleg finn ordsekvensar som seier noko om typisk bruk og kva frekvente uttrykk ord inngår i. Artikkelen tek føre seg korleis ein definerer og studerer kollokasjonar i ulike tradisjonar, då fleire tilnærmingar til omgrep som *kollokasjon* og *fleirordsuttrykk* møtest i eit praktisk felt som leksikografi. Hovudvekta i artikkelen ligg på korleis ulike reiskapar kan nyttast for å finne kollokasjonar som er relevante å ha med i ordbøkene. Korpuskel-leks er eit korpusgrensesnitt utvikla for leksikografisk bruk. Der kan brukarane søkje etter kollokasjonar ved hjelp av statistiske assosiasjonsmål og ved regulære uttrykk på tvers av mange korpus. NorGramBank er ein trebank med søkjetemplat utvikla for leksikografisk bruk. Der er kollokasjonar ikkje berre eit reint statistisk fenomen av ord som førekjem saman og kan søkjast etter lineært, men søket kan òg referere til syntaktisk struktur og syntaktiske relasjonar mellom orda.

1. Innleiing[1]

Ordet *collocation* er belagt i Oxford English Dictionary frå 1750 og har vore brukt om ord som førekjem saman sidan 1917 (Rosqvist 2014: 7). Mange knyter likevel nemninga til den engelske lingvisten J. R. Firth, som på 1950-talet kom med dei bevinga orda «You shall know a word by the company it keeps» (Firth 1957: 11). Det handlar om at vi i vanleg kommunikasjon set saman fleire ord til uttrykk eller heile setningar, slik at

1. Korpuskel-leks og INESS-trebanken NorGramBank er utvikla ved UiB og tilgjengelege via språkinfrastrukturen CLARINO gjennom CLARINO-senteret i Bergen (De Smedt et al. 2015, Rauset et al. 2022).

dei i fellesskap dannar eit tydingsinnhald som igjen skal gje ei viss meining. Tydinga til eitt ord høyrer saman med tydinga til eit anna ord. Sitatet må ikkje tolkast som at tydinga til eit ord einast trer fram når ein kombinerer det med andre ord, eller at det leksikografiske prosjektet med å definere enkeltord som isolerte einingar er håplaust. Som kompetente språkbrukarar har vi jo som regel ei meining om kva eit ord betyr. Snarare peiker sitatet til Firth på at når vi lærer ord og tydingar i eit språk, gjer vi det ut frå erfaring med korleis dei einskilde orda blir brukte i bestemte språksituasjonar. Kontekst utgjer såleis ei mogleg kjelde til å finne ut noko om tydingane til ord. Når vi seinare tolkar ei ytring i ein bestemt språkbrukssituasjon, gjer vi ikkje det åleine ut frå ei generell oppfatning av kva kvart einskild ord tyder kvar for seg. Snarare bidreg konteksten til å utdjupe, presisere eller nyansere tydinga til einskilde ord.

Leksikografisk inneber sitatet til Firth to ting: I det leksikografiske arbeidet med å identifisere og skildre tydingane til eit ord er autentiske tekstdøme på bruken av eit ord ei viktig empirisk kjelde til kunnskap for leksikografen. Vidare er gode døme på den typiske bruken av eit ord eit nyttig supplement til ein ordbokdefinisjon, nettopp fordi eit viktig aspekt ved å «kjenne eit ord» er å kjenne til karakteristiske kontekstar og vanemessige ordkombinasjonar som ordet blir brukt i.

Kollokasjonsstudiar har blitt eit stort forskingsfelt internasjonalt dei siste tiåra, og ifølgje Stefanowitsch (2020: 233f) fordeler studiane seg i tre hovudgrupper. Den første typen er ei stor mengd studiar som med utforskande metode identifiserer kollokasjonane i store digitale tekstsamlingar, det vi omtalar som korpus. Stefanowitsch hevdar at forskarane bak oftast fokuserer på metode – korleis ein kan preprosessere korpusa, kva assosiasjonsmål ein skal bruke osv. – og i mindre grad på reint lingvistiske forskingsspørsmål eller einskilde kollokasjonar. Ei anna stor gruppe kollokasjonsstudiar er den bruksretta forskinga som nyttar frekvensstudiar til å finne kollokasjonar som er relevante å gjere greie for i ordbøker og anna undervisningsmateriale. Til slutt finst det ei lita gruppe med ofte deskriptive studiar som undersøkjer kva ord som støttar eller karakteriserer enkeltord eller eit avgrensa utval av ord. Stefanowitsch (2020: 234) peikar på at det generelt er eit relativt fråvære av teoretisk ambisiøse studiar som plasserer seg i denne gruppa, medan ein i norsk samanheng kan slå fast at vi manglar studiar i alle desse tre gruppene, då litteraturen om norske kollokasjonar framleis er nokså sparsam. For å nemne nokre har Fjeld og Vikør skrive om kollokasjonar og andre ord-

forbindelsar frå eit leksikografisk perspektiv (Fjeld og Vikør 2008, Fjeld 2009). Der er òg eit knippe datalingvistiske studiar av kollokasjonar i eit norsk materiale (Andersen 2011 og 2020, Dyvik, Losnegaard og Rosén 2019, Horvati 2005 og Lyse og Andersen 2012).

Med bakgrunn i leksikografi og utvikling av datalingvistiske verktøy for å studere grammatiske og leksikalske fenomen har forfattarane av denne artikkelen med seg perspektiv frå den andre og bruksretta delen av kollokasjonsforskingsfeltet. Artikkelen har like fullt mange trekk frå den første gruppa av kollokasjonsstudiar med vekt på metode og korpus, då formålet med artikkelen er å gjere greie for dei reiskapane i og utanfor Språksamlingane som er tilgjengelege når leksikografane ved Universitetet i Bergen skal velje ut kollokasjonar til standardordbøkene *Bokmålsordboka* og *Nynorskordboka*. Kunnskap om desse reiskapane håpar vi kan vere til nytte også for eit vidare språkvitskapleg miljø. Med Korpuskelleks og INESS (NorGramBank) har vi fått kraftfulle og supplerande verktøy for å finne norske kollokasjonar, og håpet vårt er at fleire skal få auga opp for forskingsfeltet og ta i bruk reiskapane.

I teoridelen (avsnitt 2) klarlegg vi kva ein forskar på når kollokasjonar er studieobjektet. Korleis kan kollokasjonar delast i ulike undergrupper, og korleis kan dei avgrensast i høve til nærskylde språkfaglege omgrep som fleirordsuttrykk, leksikaliserte uttrykk, idiom eller konstruksjonar med partikkelverb? Vi argumenterer for ein fraseologisk typologi med idiomklyngja, ordspråksklyngja og kollokasjonsklyngja som tre sekkestorleikar – som i staden for å prøve å skilje skarpt mellom kategoriane gjer eit poeng ut av å vise at dei ofte overlappar. Denne delen blir avslutta av ein kort gjennomgang av kollokasjonar i standardordbøkene.

I den metodiske hovuddelen ligg vekta på dei to ressursane Korpuskel-leks (avsnitt 3) og INESS-trebanken NorGramBank (avsnitt 4). Vi gjer greie for kva type ressursar dette er og korleis dei kan brukast i kollokasjonssamanheng. I avslutninga (avsnitt 5) samanfattar og samanliknar vi korleis ressursane kan brukast og utfyller kvarandre.

2. Kva er ein kollokasjon og kva er han ikkje?

2.1 Kollokasjonar i den korpusorienterte og systemorienterte tradisjonen

I leksikografien er det vanleg å skilje mellom to kollokasjonsteoriar eller -retningar: den korpusorienterte og den systemorienterte (Svensén 2004:

208). Meir enn at retningane utgjer motsetnader, ser dei ut til å fokusere på ulike spørsmål.

Den aller enklaste måten å forklare ein kollokasjon på, er at det er to eller fleire ord som ofte førekjem i lag. I den korpusorienterte tradisjonen, som har sitt opphav i J. R. Firth, handlar dette om ord som i autentiske tekstar opptrer oftare saman enn det som er statistisk sannsynleg, medan ein del forskarar innanfor den systemorienterte retninga trekkjer fram at det handlar om kva ord vi kan kombinere på ein måte som er naturleg og umarkert: «det er slik vi seier det på norsk» – eller det språket ein undersøkjer. Oftast er det slik at intuisjonen om kva som er «naturleg» eller «umarkert», passar med det som kan bli kvantifisert statistisk som «vanleg».

Når vi som barn lærer eit språk, skjer det i samspel med omgjevnadane. Ved å bli snakka til og høyre språket i bruk, møte det i skrift og sjå analogiar med reglar som gjeld for andre ord, lærer vi til dømes korleis vi byggjer opp ei setning med rett ordstilling, korleis ein bøyer ord og alt anna vi kan omtale som produktive grammatiske reglar. Og ved hjelp av desse produktive reglane kan ein setje saman setningar som «dette er vêr til å bli i godt humør av». Kollokasjonar er noko litt anna, det er ferdige pakkar eller einingar som vi lærer som heilskap, som for eksempel *fint vêr* og *dårleg vêr*. Viss ein prøver å byte ut eit ord i kollokasjonen med eit synonym, gjev det fort ordforbindelsen eit litt underleg preg. Det blir uidiomatisk å skifte ut *halde* i *halde eit foredrag* med *forrette*, då det verbet har ein bruksrestriksjon som knyter det til kyrkjelege handlingar, slik vi kjenner det frå kollokasjonane *forrette ved gudsteneste* eller *forrette ei gravferd*. Like uidiomatisk er det når somme direkteomset frå engelsk og seier *gje eit foredrag*.

Det finst forskingsdesign der kollokasjonar er avgrensa til ord som står rett ved sida av kvarandre, men innanfor den korpusorienterte tradisjonen er det vanleg å tillate ein avstand på opp til fem ord mellom komponentane (Stefanowitch 2020: 220). På norsk kan vi gjerne kalle denne avstanden *kollokasjonsrekkjevidde* (eng. 'collocation span'). Vi kan illustrere det med ordsekvensen *miste fatninga*, som ein i korpus kan finne i modifisere former som «mista ikkje lett fatninga», «mistar ein augneblink fatninga», «miste dei fullstendig grepet og fatninga», men der det uansett er *miste* + *fatninga* som utgjer einninga. Ein kan tenkje på orda som inngår i kollokasjonen, som nokre velkjende figurar frå Star Wars-universet, R2-D2 og C-3PO (fig. 1): «R2-D2 er én av to skikkelser som

Figur 1: R2-D2 og C-3PO. Foto: Mulyadi på Unsplash

er med i alle filmene i serien. [Den andre] er «protokolldroiden» C-3PO som R2-D2 *stadig opptrer sammen med* (Wikipedia, s.v. «R2-D2,» lesen 18.04.2022, vår kursivering)». Analogien til kollokasjonar er at det ofte er andre som er saman med og mellom dei, men det er likevel dei som har den tette koplinga mellom seg i form av syntaktiske relasjonar, ikkje berre lineære posisjonar.

I denne artikkelen, der hovudvekta ligg på verktøya nytta til kollokasjonsanalyse, baserer vi oss mest på den kvantitative korpusorienterte retninga. Kjensla av at to eller fleire ord «heng saman» som ei eining, heng ofte saman med ei erfaring med at dei ofte opptrer saman, og dette er noko vi kan kvantifisere, observere og telje statistisk. Som statistisk omgrep kan vi definere ein kollokasjon som ein statistisk signifikant samførekomst mellom ord (Sag et al. 2002), eller som det er formulert i *Nordisk leksikografisk ordbok*: «Termen kollokasjon kan også defineres ut fra ordenes typiske opptreden i tekstsammenheng, der frekvens og syntaktisk nærhet i konkrete tekster blir lagt til grunn når enkelte kombinasjoner blir betraktet som kollokasjoner» (NLO 1997: 155). Ein slik definisjon er utgangspunktet for statistiske mål på kollokasjon, som vi kjem tilbake til i del 3.2.2. Det å ta utgangspunkt i faktisk språkbruk og mønster ein kan oppdage i korpus, er heilt sentralt i korpuslingvistikken, der kollokasjonsstudiar er eit dominerande felt.

Vi skal likevel seie litt om den systemorienterte retninga, som har vore premissleverandør for mykje av den leksikografiske kollokasjonsforskinga i Norden (jf. t.d. Malmgren 2003, Svensén 2004, Rosqvist 2010 og 2014). Her legg forskarane gjerne vekt på kva ord vi kan kombinere på ein måte som er naturleg og umarkert, og dei har eit restriktivt syn på kva som er å rekne som kollokasjonar. Sentralt står å utvikle kriterium som gjev grupper av samanliknbare einingar med same semantiske relasjon mellom komponentane, som såleis kan gje innsikt i språksystemet. Den systemorienterte tradisjonen legg til grunn at kollokasjonar har ein hierarkisk struktur med eit hovudord (*basis*) som utgjer den tematiske kjernen, og eit biord (*kollokator*) som støttar eller karakteriserer basis. Basis utgjer den konstante delen av kollokasjonen, medan

kollokatoren til dels kan skiftast ut, utan at det endrar på tydinga til kollokasjonen. Basis blir vanlegvis brukt med ei av dei mest kjende tydingane sine, medan kollokatoren, som blir vald i lag med – ikkje uavhengig av – basisen, ofte er brukt i overført tyding, som verba i *vekkje mistanke* og *sløkkje tørsten* (Svensén 2004, Rosqvist 2014). Fleire stiller krav om at både basis og kollokator må høyre til dei opne ordklassene. Funksjonsord er ikkje sett på som moglege kollokatorar i den systemorienterte tradisjonen, og ordpar som *hus og heim*, *koppar og kar* er då ikkje å rekne som kollokasjonar, då det hadde vore ein ordforbindelse av to likestilte basisar.

2.2 Kriterium for å skilje kollokasjonar frå andre ordforbindelsar

Eit kriterium for å skilje kollokasjonane frå andre ordforbindelsar er fikseringsgrad. I *Nordisk leksikografisk ordbok* blir kollokasjonar rekna til dei mindre faste ordforbindelsane (NLO 1997: 119). Der blir kollokasjon definert som «ordforbindelse der betydningen har en klar sammenheng med betydningen til de enkelte delene» (154) – her blir altså komposisjonalitet trekt fram som det fremste kjenneteiknet ved frasetypen. På linje med den systemorienterte tradisjonen framhevar NLO at eit karakteristisk trekk ved kollokasjonane er at forbindelsen består av ein overordna og konstant basis og ein kollokator med karakteriserande funksjon.

Eit anna kriterium er variasjon – å sjå på kor fritt utskiftbare komponentane i frasen er. Når forskarane innan den systemorienterte tradisjonen gjer eit poeng av å drage linjer mellom kategoriane *frie kombinasjonar, kollokasjonar og idiom*, sjølv om det kan vere vanskeleg, er det til dels motivert av eit behov for å identifisere dei ordkombinasjonane som bør ordbokførast. Kollokasjonar ser dei som ein slags halvfabrikat ein kan hente fram frå minnet når ein treng dei, og som dei meiner ein bør gjere greie for i ordbøker, medan dei frie kombinasjonane, som blir kombinerte på føreseieleg vis, ikkje høyrer heime i ordbøker (Svensén 2004: 210). Leksikografane Fjeld og Vikør (2008) har til liks med *Nordisk leksikografisk ordbok* (1997) ei litt anna kategorisering, der dei reknar med to typar av kollokasjonar. Fig. 2 byggjer på skiljet i desse kjeldene mellom frie kombinasjonar, opne kollokasjonar og avgrensa kollokasjonar, som ei form for nyansering av den systemorienterte tradisjonen – og med ei presisering om at det ikkje finst faste grenselinjer mellom dei (Fjeld og Vikør 2008: 112). Frå venstre i figuren står dei *frie*

kombinasjonane, dei vi produserer ved hjelp av dei produktive syntaktiske reglane i eit språk. Dei fleste kombinasjonar av ord er frie, og det betyr at det enkelte ordet ikkje har avgjerande styring eller innverknad på valet av dei andre orda, men vi kan skifte dei ut alt etter kva ord vi treng for å uttrykkje det vi ønskjer. Kollokasjonar er underlagt større kombinatoriske restriksjonar, men likevel i ulik grad. Difor skil figuren mellom *opne kollokasjonar* (der kollokatoren til ei viss grad kan skiftast ut, utan at det endrar på tydinga til eininga (NLO 1997: 125)) og *avgrensa kollokasjonar* (som høyrer til dei faste ordforbindelsane som ikkje tillèt endring i kollokatoren utan at det endrar tydinga til eininga (NLO 1997: 117)).

Figur 2: Frie kombinasjonar og opne og avgrensa kollokasjonar.

Ser vi på den frie kombinasjonen «lese i ei bok», kan ein ifølgje dei semantiske og syntaktiske reglane i norsk like gjerne kjøpe ei bok eller selje henne, slå opp i eller leggje henne frå seg osv. *Lese til eksamen* har eit sterkare band mellom orda for å skildre prosessen med å lære seg pensum før ein eksamen. Når vi likevel har karakterisert kollokasjonen som open, er det fordi ein kan erstatte verbet med ein del andre verb frå same semantiske domene som *pugge til eksamen* og *studere til eksamen*. At vi har plassert *le hjarteleg* som ein open kollokasjon, handlar om at det finst eit knippe utskiftbare adverb som *inderleg, høgt, godt* osv., som i liten grad endrar på tydinga til eininga.

Dei avgrensa kollokasjonane heilt til høgre er dei som i minst grad har utskiftbare komponentar, og vi kan omtale dei som dei sterkaste kollokasjonane. Eit fellestrekk ved ein del av dei er at dei inneheld komponentar som sjeldan blir kombinert med andre enn dei vi ser her. Til dømes kan *leggje ned veto* ikkje uttrykkjast på så mange andre måtar.

Det er ikkje semje om at eit partikkelverb som *køyre opp* (for å *ta lappen* eller *få sertifikat*) høyrer heime i ein slik modell, ettersom den systemorienterte retninga legg til grunn at funksjonsord ikkje er moglege kollokatorar. Ein viktig grunn til å inkludere partikkelverb i slike modellar er at ein bør gjere greie for ordforbindelsen i ordbøker. Tek vi t.d. *vaske opp* 'reingjere koppar og kar' og *vaske ned* 'ta storreingjering av hus', illustrerer dei godt at mange av partikkelverba er ikkje-komposisjonelle og leksikaliserte uttrykk. Når dei t.d. kan stå utan utfylling, som i «eg køyrde opp i stad» og «i morgon skal eg vaske opp», har dei klare trekk av å vere kollokasjonar, men grensedraginga mot idiom er krevjande. Eitt viktig skilje mellom dei to typane ordforbindelsar er at vi vanlegvis brukar kollokasjonen som heilskap ikkje-metaforisk, sjølv om kollokatoren gjerne blir brukt i overført tyding, t.d. i *gripe sjansen* (Svensén 2004: 211).

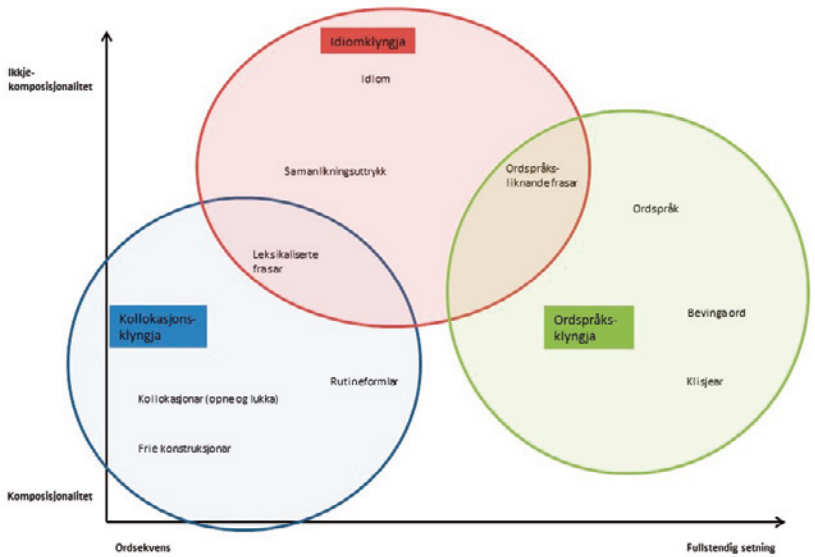
Fjeld og Vikør (2008: 113f) er dessutan blant dei som skil mellom grammatiske og leksikalske kollokasjonar. Dei grammatiske kollokasjonane har funksjonsord, ikkje minst preposisjonar, som biord, og ser vi på døme som *kunnskap om noko* og *erfaring med noko*, blir det tydeleg at ein må treffe med riktig preposisjon for å uttrykkje seg idiomatisk. Det er eit argument for å ha slike med i ordbøker som ein konvensjonalisert måte å uttrykkje eit innhald på. Den systemorienterte Svensén (2004: 213) hevdar i *Handbok i lexikografi* at denne typen ordforbindelsar ikkje er å rekne som kollokasjonar, men konstruksjonar.

Dei leksikalske kollokasjonane er kjenneteikna av at både hovudordet og biordet har sjølvstendige tydingar, som *hoppe strikk, hoppe bukk, gullande rein* og *sove søtt*. Elles ville *ta lappen* og *få sertifikat*, som vi nemnde ovanfor, begge med verb og substantiv som innhaldsord, vere rekna som leksikalske kollokasjonar etter systemorienterte kriterium.

I denne artikkelen baserer vi oss på den fraseologiske typologien som er utarbeidd i Rauset (2022: 37 ff), og som er oppsummert i fig. 3. Figuren gjer greie for eit utval sentrale frasetypar ut frå to dimensjonar: om det er ein ordsekvens eller ei fullstendig setning (x-aksen), og komposisjonalitet (y-aksen). Idiomklyngja og ordspråksklyngja representerer dei faste frasane (av typen ordspråk som *morgonstund har gull i munn,* bevinga ord som *Nu hvil dig borger, det er fortjent,* klisjear som *fordi du fortjener det,* ordspråksliknande frasar som *først til mølla,* idiom som *ha mange jern i elden*, samanlikningsuttrykk som *som olja lyn* og leksikaliserte frasar som *kald krig*), medan kollokasjonane er plassert innanfor sirkelen nedst til venstre, som del av dei mindre faste frasane i kollokasjonsklyngja. Figu-

ren skil mellom opne og avgrensa kollokasjonar, men ikkje eksplisitt mellom grammatiske og leksikalske (i tråd med argumentasjonen i denne artikkelen om at det er behov for å gjere greie for grammatiske kollokasjonar på linje med leksikalske i ordbøker). Typologien er basert på det kjende fraseologiske kontinuumet fri kombinasjon – kollokasjon – idiom/faste frasar (sjå t.d. Barkema 1996: 125, og som vi kjenner frå den systemorienterte tradisjonen), men tek sikte på å gjere greie for eit større utval av fleirordsuttrykk.

I fig. 3 blir det framheva at kollokasjonar er meir komposisjonelle enn idiom, samanlikningsuttrykk og andre faste frasar i dei romslege idiom- og ordspråksklyngjene, utan at ein hevdar at ein fullt ut kan slutte seg til tydinga til eininga ut frå tydinga til enkeltkomponentane i kollokasjonen (jf. partikkelverba *vaske opp* og *vaske ned*). Som før nemnt er eit problem ved dette kriteriet at det er gråsoner mellom kva som er fullt komposisjonelt og ikkje. I mange forskingsprosjekt vil det vere nødvendig med mykje klarare kriterium og grensedragingar, men bak figuren ligg ei praktisk-leksikografisk erkjenning av at det er svært krevjande, men ofte mindre nødvendig i leksikografisk samanheng, å skulle skilje dei ulike frasetypane frå kvarandre, og at dei ofte kan overlappe.



Figur 3: Typologi som skil kollokasjonsklyngja frå dei faste frasane i idiom- og ordspråksklyngjene.

2.3 Kollokasjonar i Bokmålsordboka *og* Nynorskordboka

I revisjonen av *Bokmålsordboka* og *Nynorskordboka* skal vi oppdatere innhaldet i to ordbøker med lang trykt og digital historie (Rauset 2019, Revisjonsprosjektet 2022). Leksikografane har tre hovudoppgåver: sikre at definisjonane er i tråd med dagens språkbruk, peike ut eit moderne og relevant ordtilfang i begge skriftspråka, og gjere utvalet av lemma i ordbøkene likare (Selback 2020). Leksikografane treng empirisk materiale for å vurdere lemmatilfang, fleirordsuttrykk, ordtydingar og syntaktisk åtferd, og finne bruksdøme. Det er då avgjerande med gode søkjeverktøy som lèt leksikografane jobbe etterretteleg og samstundes effektivt (Lyse 2020).

Standardordbøkene gjer ikkje greie for kollokasjonar (i den strenge tydinga av ordet, til skilnad frå t.d. idiom) som ein eigen kategori, men dei blir vanlegvis nytta som bruksdøme. Om tydinga til eininga avvik så mykje frå tydinga til komponentane at ho treng forklaring, legg vi kollokasjonane blant dei faste uttrykka. Dersom ein hadde laga heilt nye ordbøker, hadde det i dag vore naturleg å skilje ut kollokasjonar som ein eigen kategori, slik det t.d. er gjort i *Islex*, der kollokasjonane er gjorde greie for under dei ulike tydingane til ordet. I den ordboka kjem dei faste uttrykka (idiomklyngja og ordspråksklyngja i fig. 3) til slutt i ordartiklane, som signal om at avstanden mellom dei inngåande orda og den regulære tydinga i dei faste uttrykka er større enn i kollokasjonane.

3. Kollokasjonsstudiar i Korpuskel-leks

I arbeidet med kollokasjonar i standardordbøkene har vi fleire ressursar, og i denne artikkelen gjer vi greie for to av dei. Det eine verktøyet er Korpuskel-leks, som lèt oss gjere avanserte søk på enkeltord og frasar (Lyse 2020). Det andre verktøyet er trebanken NorGramBank for bokmål og nynorsk, der kvar setning har ein detaljert syntaktisk analyse (Dyvik et al. 2016).

3.1 Korpuskel-leks

Korpuskel er eit grensesnitt for å tilgjengeleggjere, søkje i og analysere tekstkorpus, og er eit generisk verktøy med over 60 korpus på ulike språk, format og mediatypar (Meurer 2012).[2] Korpuskel-leks er ei vidare-

2. https://clarino.uib.no/lex-prod/corpus/

utvikling av Korpuskel. Dette spesialtilpassa grensesnittet for leksikografisk bruk vart implementert for Revisjonsprosjektet i 2018, og lèt brukaren søkje på tvers av mange korpus samstundes. Som del av CLARINO er Korpuskel-leks ope tilgjengeleg, innanfor rammene av brukslisensane som gjeld for dei enkelte tekstkorpusa.

I tab. 1 listar vi dei 13 korpusa som per i dag ligg i Korpuskel-leks. Kvart av dei er lista med kolonnar som viser skriftspråk, storleik i millionar løpeord og skiljeteikn, tidsrommet korpuset dekkjer, og ei grovkategorisering av sjanger og tilgang. Metadata med meir informasjon om kvart korpus finn ein via nettsidene til Korpuskel-leks. Dette er eit heterogent materiale som er samla inn til ulike formål, og det kan difor ikkje kallast eit balansert korpus.

Styrken i materialet i Korpuskel-leks er mengda. Samla gjev Korpuskel-leks i dag eit materiale på 3,1 milliardar teikn, altså ord og teiknsetjing. Det totale materialet for bokmål er på 2,9 milliardar teikn, medan det totale materialet på nynorsk er på dryge 217 millionar, altså med ei fordeling på 93 % for bokmål og 7 % for nynorsk. Materialet har tekstar frå perioden 1765–2021, men med ei hovudvekt på nyare materiale. Om lag 75 % av materialet er tekst frå 1998 og fram til notid, altså om lag 2,2 milliardar teikn. Dette er naturleg nok viktig for revideringa av standardordbøkene, som skal beskrive eit moderne ordtilfang. Ser vi på skriftspråka kvar for seg, er 77 % av bokmålsmaterialet (2,9 milliardar) frå 1998 og framover i

Korpusnamn	Skrift-språk	Storleik (mill.)	Tidsrom	Sjanger	Tilgang
Forskning.no (2017)	nob	26,2	1998–2017	tekst: sakprosa	avgrensa
NAOB-tekster	nob	238,7	1846–2018	tekst: skjønnlitteratur	avgrensa
NBs frie tekster (bokmål)	nob	516,4	1765–2013	tekst: blanda	open
NBs frie tekster (nynorsk)	nno	46,0	1850–2010	tekst: blanda	open
Norsk aviskorpus (bokmål)	nob	2046,3	1998–2021	tekst: sakprosa	open
Norsk aviskorpus (nynorsk)	nno	21,1	1998–2020	tekst: sakprosa	open
Norsk aviskorpus annotert	nob	35,7	2001–2009	tekst: sakprosa	open
Norsk ordboks nynorskkorpus	nno	129,1	1866–2012	tekst: blanda	avgrensa
Norsk ordboks nynorskkorpus (2017)	nno	4,0	1926–2017	tekst: blanda	avgrensa
Talk Of Norway	nob, nno	64,3	1998-2016	tekst: sakprosa	open
Dialektendring (Talebank)	nno	5,6	1960–2020	tale: dialektkorpus	avgrensa
Industristad (Talebank)	nno	2,3	1948–2013	tale: dialektkorpus	avgrensa
Talesøk (Talebank)	nno	1,9	1998–2016	tale: dialektkorpus	avgrensa

Tabell 1: Korpus som er søkbare gjennom Korpuskel-leks.

tid; tilsvarande tal for nynorskmaterialet (217 millionar) er 56 %. Det må understrekast at tilgangen på eldre tekst òg er nyttig i ordbokarbeidet, mellom anna for å vurdere tradisjonsord og i kva grad dei er i bruk i dag samanlikna med tidlegare.

Ein veikskap i materialet er den skeive fordelinga mellom nynorsk og bokmål, og dessutan sjangerfordelinga. Tab. 2 viser at sakprosa utgjer nesten 70 % av det samla materialet, gjennom avistekstar og populærvitskapleg tekst (*Norsk aviskorpus bokmål* og *nynorsk, Forskning.no 2017* og *Talk of Norway*). I tillegg er der noko sakprosa i materialet som ligg i kategorien «blanda», og som både inneheld skjønnlitterær tekst og sakprosa (*Nasjonalbibliotekets (NBs) frie tekstar bokmål* og *nynorsk* og dei to nynorskkorpusa).

Sjanger	Storleik (mill.)	Prosent
skjønnlitteratur	238,74	7.61 %
sakprosa	2193,61	69.91 %
blanda (skjønnlitteratur, sakprosa)	695,52	22.17 %
dialektkorpus	9,78	0.31 %

Tabell 2: Sjangerfordeling mellom korpusa i tabell 1.

3.2 Korpusstudiar i Korpuskel-leks

Vi vil her gjennomgå dei to sentrale metodane i korpusstudiar; enkle søk og frekvenslister, og statistisk baserte rangeringar av ordsekvensar med høg kollokasjonsstyrke. Dataa i alle døma i del 3.2 kjem frå Korpuskelleks, med søk i alle korpusa i tab. 1, bortsett frå den annoterte delen av Aviskorpuset. For kvart søk er søkje-URL gjeve opp i ein fotnote. URLen gjev direkte det same søket, og føreset at brukaren har tilgang til dei aktuelle korpusa.

To terminologiske presiseringar må til. For det første omtalar vi for det meste sekvensar av ord. Men i eit elektronisk korpus finn vi teknisk sett strengar, òg kalla *teikn* (engelsk 'tokens') skilde med mellomrom, som omfattar både ord og teiknsetjing. Det vil seie at vi òg kan møte sekvensar av ord og teiknsetjing.

For det andre omtaler studiar med assosiasjonsmål vanlegvis ordsekvensar som *n*-gram, altså ein sekvens av ord med ei lengd på *n*. Eit *n*-gram med ei lengd på 2 kallar vi til dømes eit bigram, medan eit trigram har ei lengd på 3. Assosiasjonsmåla i Korpuskel-leks er implementerte som ei utrekning på forbindelsen mellom to teikn for å identifisere dei

komponentane *som stadig opptrer saman*, òg sjølv om dei ikkje alltid følgjer rett etter kvarandre – jamfør analogien til R2-D2 og C-3PO i avsnitt 2.1. Vi må difor skilje mellom bigram og trigram osv., som altså er ordsekvensar med direkte kontakt mellom orda, og statistisk sterke ordforbindelsar, som ikkje føreset slik direkte kontakt. I Korpuskel kan ein be om ei rangering av kollokasjonar for eit ord, t.d. *fatning*, med eit kontekstvindauge på t.d. [± 2]. Då vil det bli rekna ut kollokasjonsstyrke for *fatning* og alle ord som kjem i høvesvis første og andre posisjon framfor og etter. Eit setningsdøme som: *han tek tapet med* ***fatning*** *og seier at (...)* ville då gje følgjande ordpar som vi kan kalle kollokasjonskandidatar, skrivne mellom hakeparentesar, og der ein understrek indikerer at orda ikkje står direkte etter kvarandre: [tapet_ fatning], [med fatning], [fatning og] og [fatning _seier].

3.2.1 Regulære uttrykk og enkle frekvenslister

Ein kan gjere enkle søk på ordsekvensar ved hjelp av regulære uttrykk. La oss bruke verbet *hoppe* som døme. Eit typisk behov for leksikografen er å undersøkje kva ordsekvensar dette ordet inngår i. I Korpuskel-leks søkjer ein etter ein streng ved å skrive han mellom hermeteikn, og vi kan søkje etter ein vilkårleg streng ved å skrive ein tom hakeparentes. Søk (1) under vil vere eit søk på ordforma «hoppe» og det første ordet som følgjer etterpå (eit vanleg bigram), medan søk (2) representerer det tilsvarande trigrammet. Søk (3) vil gje treff på trigram som har ordforma «hoppe» i midten. Ofte er det nyttig å snevre inn eit søk tidsmessig, t.d. berre til treff etter 1970, som i søk (4). Dette er til hjelp dersom ein treng å undersøkje moderne bruk. Ofte treng vi å sjå alle bøyingsformene (søk 5), og i nokre tilfelle søkjer vi på leksemet «hoppe» (det vil seie alle bøyingsformene), som i søk (6). Det siste gjer redaksjonen relativt sjeldan, fordi det avgrensar treffa til det subsettet av korpus som har morfologisk analyse tilgjengeleg – dvs. korpusa *Norsk aviskorpus (annotert), Dialektkorpus, Forskning.no (2017), Industristad, Nynorskkorpusa, Talesøk* og *Talk of Norway*. Sjå elles nettsidene til Korpuskel-leks for ein meir omfattande dokumentasjon av søkjeuttrykk.[3]

(1) “hoppe” []

3. https://clarino.uib.no/lex-prod/documentation/korpuskel-documentation

(2) “hoppe” []{}[][4]
(3) [] “hoppe” []{}[][5]
(4) “hoppe” []{}[] :: year > “1970”[6]
(5) “hoppe|hopper|hoppar|hoppet|hoppa” []{}[][7]
(6) [/hoppe/ & {verb}] []{}[][8]

Søket i døme (2) gjev 45478 treff, fordelte på 12 korpus og innanfor tidsrommet 1838–2021.[9] Korpuskel-leks viser treffa på ulike måtar. Fana *Konkordans* viser treffa til søket i det klassiske visingsformatet med ei linje per treff (KWIC, eller KeyWord In Context). Treffordet/-orda er då midtstilte med ei viss mengd kontekst før og etter. Fana *Ordliste* oppsummerer kor mange gonger kvart mønster førekom, og er enkel å laste ned, som i uttrekket i tab. 3. Denne tabellen viser toppen av ei frekvensliste over korleis treffa på søket "hoppe" []{}[] er fordelte på dei ulike mønstera, med sekvensen «hoppe ut av» som den mest frekvente sekvensen.[10]

I det daglege ordbokarbeidet er slike frekvenslister eit effektivt verktøy for å få eit raskt overblikk over ordsekvensar som kan seie noko om

4. Søkje-URL: https://clarino.uib.no/lex-prod/corpus/avis-plain,avis-nno,dialekt,fn-new,industristad,naob,nb-fri-nob,nb-fri-nno,nnk,nnk-new,talesoek,ton/%22hoppe%22%20%5B%5D%5B%5D
5. Søkje-URL: https://clarino.uib.no/lex-prod/corpus/avis-plain,avis-nno,dialekt,forskning-no,fn-new,industristad,naob,nb-fri-nob,nb-fri-nno,nnk-new,talesoek,ton/%22hoppe%22%20%5B%5D%5B%5D
6. Søkje-URL: https://clarino.uib.no/lex-prod/corpus/avis-plain,avis-nno,dialekt,fn-new,industristad,naob,nb-fri-nob,nb-fri-nno,nnk,nnk-new,talesoek,ton/%22hoppe%22%20%5B%5D%5B%5D%20%3A%3A%20year%20%3E%20%221970%22%20
7. Søkje-URL: https://clarino.uib.no/lex-prod/corpus/avis-plain,avis-nno,dialekt,fn-new,industristad,naob,nb-fri-nob,nb-fri-nno,nnk,nnk-new,talesoek,ton/%22hoppe%7Chopper%7Choppar%7Choppet%7Choppa%22%20%5B%5D%5B%5D
8. Søkje-URL: https://clarino.uib.no/lex-prod/corpus/avis-plain,avis-nno,dialekt,fn-new,industristad,naob,nb-fri-nob,nb-fri-nno,nnk,nnk-new,talesoek,ton/%5B%E2%88%A5hoppe%E2%88%A5%20%26%20%7Bverb%7D%5D%20%5B%5D%5B%5D%20
9. Legg merke til at vi ikkje søkjer i det annoterte Aviskorpuset i Tabell 1 samstundes som vi søkjer i dei to vanlege aviskorpusa på høvesvis bokmål og nynorsk, ettersom dette ville kunne gje dupliserte treff.
10. Ein feil i ein del av inndata til Korpuskel-Leks førte til feilaktige treff. Denne feilen kjem til å bli fiksa i korpuset; i denne omgangen nytta vi eit søkjeuttrykk som ser bort frå den delen av korpus der feilen ligg. I staden for søkjeuttrykket i døme (2) på s. 48, nytta vi difor dette uttrykket: "hoppe" []{}[] @@ 0-93552955, 96452303-100000000000.

typisk bruk, og peike oss til frekvente uttrykk der fokusordet vårt inngår. I lista i tab. 3 ser vi t.d. sekvensar som *hoppe etter Wirkola, hoppe i det, hoppe i havet* og *hoppe i taket*. I søk på trigram, som dette dømet, kan vi òg intuitivt kjenne att toordssekvensar som er moglege kollokasjonskandidatar, som *hoppe av* og *hoppe bukk*. Slike reine frekvenslister er dessutan nyttige for å danne seg eit bilete av vanlege uttrykksmåtar som kan vere relevante som illustrerande bruksdøme i ordbokartikkelen (*hoppe ut frå, hoppe i fallskjerm*).

frekvens	"hoppe" []{}	frekvens	"hoppe" []{}
982	hoppe ut av	304	hoppe bukk over
571	hoppe etter Wirkola	299	hoppe opp og
563	hoppe i det	291	hoppe i taket
515	hoppe i sjøen	256	hoppe i havet
502	hoppe på ski	238	hoppe av ,
456	hoppe inn i	233	hoppe av isen
447	hoppe ut i	232	hoppe over til
374	hoppe av .	231	hoppe finaleomgangen .
326	hoppe i fallskjerm	229	hoppe i vannet
314	hoppe ut fra	213	hoppe av og

Tabell 3: Toppen av frekvenslista over treordsekvensar med «hoppe» som første ordform ved søk i Korpuskel-leks, med søk i alle korpus i Tabell 1 unnateke den annoterte versjonen av Aviskorpuset.

Frekvenslister som denne fangar berre inn dei ordsekvensane som samla har *høgast frekvens*. Men dei tek ikkje omsyn til i kva grad orda som inngår i ein sekvens, statistisk sett har sterkare tendens til å opptre saman enn med andre ord – altså om dei er ein sterk kollokasjon. Til dette har ein det ein kallar assosiasjonsmål.

3.2.2 Assosiasjonsmål

Såkalla assosiasjonsmål (AM) er statistiske metodar som prøver å kvantifisere ein intuisjon om ord som «høyrer saman». Enkelt sagt går dei fleste AM ut på å analysere relasjonen mellom kor ofte orda i ein sekvens opptrer saman, samanlikna med kor ofte dei førekjem kvar for seg.

I Korpuskel-leks er fem ulike mål på kollokasjon implementerte per i dag: *Frekvens, Relativ frekvens, Log Likelihood (LL), Mutual Information (MI)* og *MI * log (frekvens)*. Vi fokuserer her på korleis slike statistiske AM-mål kjem til nytte i det leksikografiske arbeidet, og for den interesserte lesaren viser vi til dokumentasjonen på nettsidene til Korpuskel-

leks og til Evert (2004), Lyse og Andersen (2012) og Pedersen et al. (2011).

I tab. 3 ovanfor såg vi ei frekvenssortert liste over treordssekvensar som startar med hoppe, altså eit søk på "hoppe" [][]. Fig. 4 viser tilsvarande dei ordpara som blir rangert høgast på statistisk grunnlag når *hoppe* er første ord og ein ser på kva ord som førekjem i posisjon éin eller to etter *hoppe*. I Korpuskel-leks kan vi søkje etter statistisk sterke kollokasjonar under fana *Kollokasjonar* (tredje fane øvst på skjermbiletet i fig. 4). Her kan ein velje mellom ulike attributt som er knytte til søkjeordet, og som regel vel vi den bøygde ordforma (attributtet *word*). Nedtrekksmenyane *venstrekontekst* og *høgrekontekst* handlar om kollokasjonsrekkjevidde og spesifiserer kor mange ord på kvar side ein vil ha for å berekne kollokasjonskandidatane til søkjeordet, og ein kan velje ein verdi frå 0 (ingen kontekst på den eine sida av søkjeordet) til 6 (alle kontekstord opp til 6 ord unna søkjeordet). For å svare til søket som gav frekvenslista i tab. 3, vel vi høgrekontekst = 2, altså to «plassar» til høgre for søkjeordet. I feltet *Terskel* i fig. 4 kan ein setje ei nedre grense for kor frekvente kollokasjonskandidatane må vere for å bli vurderte. Det er lurt å ikkje

Vis kollokasjoner | Attributt: word | ignorer storskriving | Venstrekontekst: 0 | Høyrekontekst: 2

Terskel: | Sorter etter: MI * log(Freq)

5965 kollokasjoner. ‹ 1 2 3 4 5 ... 199 › | Gå til side: | Last ned

Value	Frequency	Relative	Log likelihood	Mutual information
hoppe _ Wirkola	575	0.15481961	218854.69	13.366307
hoppe bukk	365	0.10951095	226106.02	12.866794
hoppe _ fallskjerm	344	0.062352728	173673.23	12.054243
hoppe finaleomgangen	299	0.024725048	92299.52	10.719764
hoppe _ strikk	106	0.051986266	271264.03	11.791922
hoppe _ flyvingen	36	0.13235295	463425.62	13.140109
hoppe skiflyging	135	0.03736507	213916.58	11.315481
hoppe _ sprette	124	0.03856921	225654.72	11.36124
hoppe _ høyet	104	0.043315284	254868.72	11.528667
hoppe _ trampoline	78	0.052419353	301999.62	11.80389

Figur 4: Dei høgast rangerte kollokasjonane for treordsekvensar med «hoppe» som første ordform ved søk i Korpuskel-leks.

setje ei slik grense før ein har sett den første rangeringa utan ein terskelverdi. Ser ein at det kom mange lågfrekvente og mindre interessante ordpar høgt på lista, kan ein eventuelt setje ei grense slik at alle kollokasjonskandidatar med frekvens lågare enn (t.d.) 5 blir filtrerte vekk. I Korpuskel-leks blir alltid alle fem statistiske mål rekna ut for kvart ordpar, og i resultatlista ser ein dei statistiske verdiane i kvar si kolonne. Unntaket er det statistiske målet *MI * log(Freq)*, som berre er produktet av verdiane i dei to siste kolonnane. I nedtrekksmenyen *Sorter etter* vel ein kva for statistisk mål ein vil sortere lista etter. Her viser vi resultata med målet *MI * Log(Freq)*, som ofte viser seg å gje den mest nyttige rangeringa. Det er slik sett kanskje noko uheldig at rekkefølgja av kollokasjonskandidatane i figuren ikkje samsvarer med nokon av kolonnane. Grunnen er ganske enkelt at dette femte statistikkmålet vart innført i etterkant av at dei fire første var implementerte, medan visinga med fire kolonnar vart ståande.

Samanliknar ein topptreffa i fig. 4 med den reine frekvenslista som vi såg i tab. 3, ser vi at orda i kvar kollokasjon intuitivt er langt meir bundne til kvarandre enn det som kom på topp i den reine frekvensordlista. Kvar linje (kvart rangerte ordpar) har ei kolonne *Frequency*. Vi ser at dei høgast rangerte ordpara i dette søket er relativt frekvente, som kollokasjonskandidatane [hoppe _ Wirkola] (som var funne 575 gonger) og [hoppe bukk] (som var funne 365 gonger). Understrekinga i den første av desse kollokasjonskandidatane betyr at det står eit ord mellom ord1 og ord2. Dette stemmer sjølvsagt òg med intuisjonen om dette dreier seg om uttrykket *hoppe etter Wirkola* 'å prøve seg mot nokon som er betre'. Den statistisk sterke kollokasjonen i dette uttrykket gjeld altså sambandet mellom dei to innhaldsorda, *hoppe* og *Wirkola*, og ikkje mellom *hoppe* og *etter* eller *etter* og *Wirkola*.

Dette er ei statistisk rangering av kollokasjonskandidatar som ikkje tek stilling til lingvistisk analyse. For leksikografen er altså dette eit supplement til vanlege lister av konkordansar eller ordlistefrekvens, men det er ikkje lister som gjev ferdige svar på kva ein eventuelt skal gjere med det som statistisk står som sterke kollokasjonar. Leksikografen vurderer deretter i kva grad dette kvalifiserer til å stå som eit fast uttrykk i ordboka med ei eiga forklaring, om det illustrerer ein bestemt bruk av fokusordet som bør inn som ei eiga tyding, eller om dette er ein illustrerande kollokasjon som kan brukast som bruksdøme under ei eksisterande tyding (uttrykket *hoppe etter Wirkola* er utan tvil eit fast og ikkje-komposisjonelt

idiom som treng eit eige oppslag og ei eiga forklaring i ordbøkene, og ligg allereie som eit eige oppslag der).

Kva så med den kollokasjonen som var rangert som nest sterkast i fig. 4, *hoppe bukk*? Intuitivt fortel språkkjensla oss at ordet *bukk* står ubøygd i denne kollokasjonen – dette er ikkje ein vanleg transitiv relasjon der ein t.d. kan «hoppe bukken». Søkjer ein etter ordsekvensen [] "bukk" i Korpuskel-leks (eit søk etter kva som helst direkte framfor ordforma *bukk*), vil ein raskt finne at verbet *hoppe* kan førekome i ulike bøyingsformer (med høgfrekvente treff som «hoppe bukk», «hopper bukk», «hoppet bukk»).

For å undersøkje kor variabelt uttrykket er, kan ein opne for strengar mellom «hoppe» og «bukk», og òg sjekke kva strengar som kjem imellom, t.d. slik:

"hopp.*" []{0,3} "bukk" []

Dette søket spesifiserer at ein vil ha alle strengar som begynner på «hopp» (*hoppe, hoppar, hoppa* osv.). Så kan det vere mellom 0 og 3 strengar før strengen «bukk», medan ein tom hakeparentes etter strengen «bukk» betyr at ein vil sjå kva strengar som følgjer rett etter «bukk». Dette gjev ei frekvensliste der ein raskt ser av dei 20 mest frekvente ordsekvensane (tab. 4) at det stabile er ulike bøygde former av verbet *hoppe*, ordforma *bukk*, og for det meste er *bukk* følgt av ordet *over*: *hoppe bukk over*. Så ser ein at det kan skytast inn modifiseringar, som t.d. «hopper statsråden ganske enkelt bukk over» og «hopper også bukk over». På dette grunn-

frekvens	kollokasjon	frekvens	kollokasjon
335	hopper bukk over	10	hoppes bukk over
304	hoppe bukk over	9	hoppe bukk "
202	hoppet bukk over	9	hopper elegant bukk over
30	hoppa bukk over	7	hopper han bukk over
20	hoppar bukk over	7	hopper man bukk over
19	hopper de bukk over	6	hoppe bukk og
16	hopper statsråden ganske enkelt bukk over	6	hoppet bukk i
14	hoppe bukk .	5	hopp bukk over
13	hopper også bukk over	5	hoppe bukk eller
11	hoppe bukk ,	5	hopper helt bukk over

Tabell 4: Toppen av frekvenslista ved søk på variantar i uttrykket : "hopp." []{0,3} "bukk" [] i Korpuskel-leks, med søk i alle korpus i Tabell 1 unnateke den annoterte versjonen av Aviskorpus.*

laget er det velmotivert å leggje inn *hoppe bukk over* som eit eige uttrykk med tydinga 'ikkje ta omsyn til'. I tillegg bør ordparet *hoppe bukk* (om barneleiken) gjerast greie for, sidan heller ikkje dette er ei gjennomsiktig og generell tyding av verbleksemet *hoppe*.

Den tredje høgast rangerte kollokasjonen i fig. 4, [hoppe _ fallskjerm], illustrerer ein statistisk sterk kollokasjon som intuitivt ser ut til å vere ganske komposisjonell. Den noverande første tydinga av verbet *hoppe* (etter revisjon) er slik: 'ta sats og sprette opp i eller gjennom lufta'.[11] Det å ta sats og sleppe seg gjennom lufta i fallskjerm verkar som ei heilt gjennomsiktig tyding, der «hoppe i fallskjerm» eventuelt kunne stå som eit døme på typisk bruk av verbet *hoppe* i tyding 1.

På denne måten kan leksikografen jobbe seg gjennom lister av statistiske kollokasjonar. Men ein må ha i mente at dei ikkje fortel meir enn at her er ein ordsekvens med ei sterk kollokasjonsstyrke. Om dette er ein bit av eit større uttrykk, og kva ein skal gjere med kollokasjonen, er likevel noko leksikografen må vurdere.

Som vi såg av dømet med *hoppe etter Wirkola* i fig. 4, kan det vere at assosiasjonsmåla ikkje fangar inn uttrykk med kollokasjonsrekkjevidde på meir enn to ord, sidan dei berre reknar statistikk på ordpar. Ved å leggje inn ein tom streng i søket som dannar utgangspunkt for å rekne ut kollokasjonar, blir det rekna assosiasjonsmål på treordsuttrykk framfor på toordsuttrykk. Eit døme kan vere adjektivet *fruktbar*, som står oppført som del av kollokasjonen *fruktbart samarbeid* i fig. 2 i avsnitt 2.2, og som òg blir brukt som døme på søk i ein syntaktisk trebank i avsnitt 4.2. Med ei søkjeeining som

"fruktba(r|rt)" []

får ein treff på ordformene *fruktbar* eller *fruktbart* følgt av ein vilkårleg streng (som t.d. «fruktbar dialog» eller «fruktbart virkemiddel»). Ein kan deretter få rangert kva som er statistisk sterke kollokasjonar til desse toordsekvensane. Av fig. 5 ser vi at det systemet gjer, er først å hente ut gode toordssekvensar (*fruktbar alder, fruktbart samarbeid, fruktbart å*) og å vise kva som er statistisk sterke kollokasjonar til desse igjen (*i fruktbar alder, et fruktbart samarbeid, fruktbart samarbeid med, være fruktbart å*). Så må leksikografen framleis vurdere om dette er sekvensar som kan fun-

11. https://ordbokene.no/nn/31276/hoppe

Vis kollokasjoner Attributt: word ignorer storskriving Venstrekontekst: 1 Høyrekontekst: 1 Terskel:

Sorter etter: MI * log(Freq)

6786 kollokasjoner. ‹ 1 2 3 4 5 ... 227 › Gå til side: Last ned

Value	Frequency	Relative	Log likelihood	Mutual information
i fruktbar alder	418	0.000005453039	[]	[]
et fruktbart samarbeid	309	0.00002391944	[]	[]
fruktbart samarbeid med	169	0.0000062779964	[]	[]
være fruktbart å	162	0.00003300429	[]	[]
mer fruktbart å	153	0.000043486973	[]	[]
og fruktbart samarbeid	115	0.0000017304623	[]	[]
fruktbart samarbeid mellom	108	0.00004907969	[]	[]
fruktbar alder .	99	6.042466e-7	[]	[]
lite fruktbar Midtøsten-reise	98	0.00013218306	[]	[]
fruktbar Midtøsten-reise Rapport	95	0.0033389567	[]	[]
lite fruktbart å	92	0.00012409023	[]	[]
en fruktbar måte	88	0.0000028333395	[]	[]
fruktbart samarbeid .	80	4.882801e-7	[]	[]

*Figur 5: Dei høgast rangerte kollokasjonane for fruktbar|fruktbart + kva som helst, med éin streng framfor eller éin streng etter. Vising av rangeringa med assosiasjonsmålet MI * log (frekvens).*

gere som gode bruksdøme på ei eksisterande tyding av *fruktbar*, eller om det bør leggjast inn som eit fast uttrykk.

Generelt er erfaringa til redaksjonen at det løner seg å starte med regulære søk på frekvente ordsekvensar som fokusordet vårt inngår i, før ein eventuelt òg sjekkar rangeringa med kvart av dei moglege assosiasjonsmåla. Ulike ord har ulik frekvensprofil, og det er lite føreseieleg kva kollokasjonar som kan vise seg nyttige for det aktuelle ordet ein jobbar med. Generelt har Log Likelihood ein tendens til å favorisere høgfrekvente, formulaiske sekvensar som typisk inneheld minst eitt grammatisk ord, som preposisjonar, subjunksjonar eller infinitivsmerket *å*. (t.d. *være fruktbart å* eller *et fruktbart og*). Mutual Information har ein tendens til å favorisere lågfrekvente kollokasjonskandidatar der eitt eller begge orda opptrer særleg eller berre i denne konteksten. Målet *MI * log (frekvens)* gjev større vekt til frekvensen til kollokasjonskandidaten som heilskap, og dette målet er sett opp som standardmålet når ein ikkje har valt noko anna. Dette er ikkje basert på ei vitskapleg undersøking, men erfaringsmessig gjev dette målet ofte dei mest relevante kollokasjonane med innhaldsord.

Det kan òg nemnast at redaksjonen har tilgang til å bruke ordskissene i Sketch Engine (Kilgarriff et al. 2014). Sketch Engine ligg på ein måte i grenseland mellom det vi finn i Korpuskel-leks og i NorGramBank i INESS. Sketch Engine er eit komplett verktøy for å søkje på ord og få forslag til statistisk sterke kollokasjonar, og ein kan få desse sortert i ulike grammatiske kategoriar (med kategoriar som *modifiers of, subjects of, objects of, and/or*). Denne ressursen er likevel sjeldan den første vi slår opp i, primært fordi det norske tekstmaterialet i Sketch Engine består av nedlasta tekst frå norske nettsider, som dekkjer alt frå offisielle nettsider til bloggtekst. Som diskutert i Lyse (2020: 4) kviler Revisjonsprosjektet på tilgang til kjeldegrunnlag av ein viss skriftleg kvalitet, og vi prioriterer difor empirisk materiale der teksten har vore gjennom ein redaksjonell prosess før publisering (i motsetnad til t.d. ein del bloggtekstar).

4. Kollokasjonsstudiar i INESS

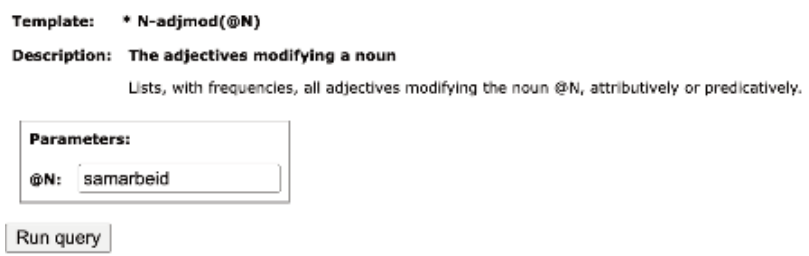
4.1 Trebanken NorGramBank

Ein trebank er eit syntaktisk analysert tekstkorpus – eit korpus der kvar setning er forsynt med ein syntaktisk (nokre gonger òg ein semantisk) analyse. Analyseformata kan variere, og trebankar kan vere bygde opp gjennom reint manuell analyse, gjennom manuell analyse i kombinasjon med større eller mindre innslag av automatisk analyse, eller gjennom reint automatisk analyse (parsing).

NorGramBank (Dyvik et al. 2016) er ein norsk trebank som er utvikla ved INESS- prosjektet ved Universitetet i Bergen (sjå Rosén et al. (2012) og nettsida for prosjektet http://clarino.uib.no/iness). Setningane i NorGramBank er analyserte automatisk med den komputasjonelle norske grammatikken NorGram, utvikla gjennom fleire år ved Universitetet i Bergen og basert på den syntaktiske teorien leksikalsk-funksjonell grammatikk (LFG). NorGramBank inneheld no ca. 160 millionar ord analysert tekst (av dette ca. 150 millionar på bokmål), som omfattar aviser, sakprosa, skjønnlitteratur, stortingsforhandlingar og somme andre teksttypar i mindre omfang. Det skjønnlitterære materialet og mykje av sakprosaen har prosjektet motteke i OCR-skanna form frå Nasjonalbiblioteket.

4.2 Nokre søkjetemplat som er eigna i kollokasjonsstudiar

Ein trebank tillèt søk etter og teljing av syntaktiske eigenskapar i tekstar. Dette gjer det mogleg å søkje etter fleirordsuttrykk på grunnlag av syntaktisk struktur og grammatiske relasjonar i uttrykka, og ikkje berre på grunnlag av den lineære plasseringa av elementa i dei. Eit døme kan vere kollokasjonen *fruktbart samarbeid*, som vi har vore inne på tidlegare. I vurderinga av om dette dømet er ein kollokasjon, er det mellom anna relevant å sjå på relative frekvensar. I jamføringa av frekvensane for ulike adjektiv som modifiserer substantivet *samarbeid*, er det då nyttig å få med døme der adjektivet ikkje berre står attributivt framfor substantivet, men òg predikativt i ulike konstruksjonar. Då kan lineært søk bli vanskeleg, medan ein trebank kan nyttast. Søkjeuttrykket for å finne adjektiv i ulike syntaktiske funksjonar kan bli komplekst, men til hjelp for leksikografane har NorGramBank ein reiskap kalla *søkjetemplat*, utvikla i samråd med leksikografane i Revisjonsprosjektet og i ordboka NAOB. Eit søkjetemplat er eit ferdig, parametrisert søkjeuttrykk der brukaren berre supplerer parameterverdiar, til dømes lemmaformer det skal søkjast etter. Namnet på templatet for søk etter adjektiv som kan modifisere eit bestemt substantiv, er **N-adjmod(@N)**. Når dette er valt frå ein meny, kjem brukaren til ei side der substantivet kan skrivast inn (fig. 6). Dette søket finn 729 ulike adjektiv, fordelte på 7 303 døme. *fruktbar* kjem på 28. plass med 43 treff; toppen av frekvenslista ser ut som i fig. 7.



Figur 6: Eit søkjetemplat med parameterverdi supplert av brukaren.

Nytta av å kunne søkje etter predikativ funksjon går til dømes fram ved adjektivet *viktig*, der nokre av dei 105 døma er viste i fig. 8. Desse døma kunne elles berre ha vorte funne gjennom lineært søk med stor kollokasjonsrekkjevidde, som ville ha gjeve mykje støy i form av irrelevante treff.

729 match types, 7303 matches. | Page 1 of 15 Next | Go to page: Go

Click on a row to see the matching sentences. | Copy format: plain NAOB

Count	#noun: atom	#p: value
1071	samarbeid	god
508	samarbeid	nær
478	samarbeid	internasjonal
454	samarbeid	nordisk
406	samarbeid	tett
172	samarbeid	økonomisk
130	samarbeid	interkommunal
127	samarbeid	europeisk
125	samarbeid	borgerlig
108	samarbeid	forpliktende
105	samarbeid	viktig
105	samarbeid	politisk
104	samarbeid	bred
90	samarbeid	konstruktiv
83	samarbeid	intim

Figur 7: Toppen av frekvenslista for adjektiv som modifiserer samarbeid.

125	samarbeid	borgerlig
108	samarbeid	forpliktende
105	samarbeid	viktig

Page 1 of 6 Previous Next | Go to page: Go | Download

Click on a row to go to the sentence. Mouse over a row to see the structures.

Treebank	Document	Trans.	Id	Sentence	
nor-stortinget_4	s140527	no	**1435**	Samarbeid mellom barnevernstjenesten og foreldrene blir ansett som viktig for at hjelpetiltakene skal ha en funksjon.	Copy
nor-stortinget_4	s140605	no	**1071**	Samarbeid med næringslivsaktører i utviklingspolitikken blir derfor viktigere.	Copy
nor-stortinget_4	s140619	no	**4190**	Finanskrisen viste oss at samarbeid på tvers av landegrenser er viktig for å løse internasjonale problemer.	Copy
nor-stortinget_4	s141105	no	**1308**	Jeg vil legge til at et styrket samarbeid mellom de enkelte nordiske lands utenrikstjenester faktisk er viktig av flere grunner.	Copy
nor-stortinget_4	s141111	no	**762**	Sjelden er et velfungerende samarbeid på tvers av landegrensene viktigere enn i slike situasjoner.	Copy

Figur 8: Døme med samarbeid modifisert av adjektivet viktig.

Trebanken kan òg nyttast til å skaffe oversyn over argumentrammer til verb, og over kva for verb eit substantiv oftast er argument til. Slik informasjon er nyttig i arbeidet med å finne typiske døme på verb-substantiv-kombinasjonar til ordbokoppslaga. Vi kan illustrere med eit døme frå fig. 2 i avsnitt 2.2, som viser nokre frie kombinasjonar og opne og avgrensa kollokasjonar. *kjøpe bil* og *selje bilen* er nemnde som døme på frie kombinasjonar, medan *starte bilen* er døme på ein open kollokasjon. Ei jamføring mellom desse uttrykka treng informasjon om kor ofte *bil* førekjem som argument ved ulike verb. Til dette kan ein nytte templatet

2234 match types, 26039 matches. | Page 1 of 45 Next | Go to page: Go

Click on a row to see the matching sentences. | Copy format: plain NAOB

Count	#A-arg2of: value	#B-noun: atom	#C-arg1of: value
1350		bil	være
1248		bil	stå
1081	ha	bil	
1040	kjøre	bil	
939		bil	kjøre
898		bil	komme
580	parkere	bil	
571	være	bil	
445		bil	stanse
363	ta	bil	
348	se	bil	
333	starte	bil	
309		bil	stoppe&stop
274	stanse	bil	
264		bil	svinge
258		bil	kunne
249		bil	passere
249		bil	exist
243		bil	skulle
240	kjøpe	bil	

Figur 9. Verb med bil *som argument 2 (patiens) til venstre og som argument 1 (agens) til høgre. (*exist *symboliserer presenteringskonstruksjon med* vere/være*, som i:* det er få bilar i gatene*).*

N-argofverbs(@N) med parameterverdien *bil*. Dette søket får 26 039 treff fordelte på 2 234 verb-argument-kombinasjonar; dei 20 mest frekvente kombinasjonane er viste i fig. 9.

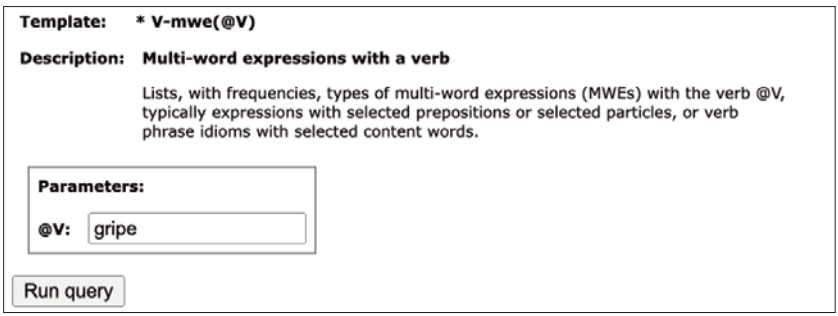
selge kjem på 51. plass med 85 treff. (Som tidlegare nemnt byggjer ikkje skiljet mellom frie kombinasjonar og kollokasjonar berre på skilnaden mellom frekvensane til uttrykka – ein må òg ta omsyn til frekvensane til kvart ord, som ein òg kan finne i trebanken.)

At trebanken ikkje berre tillèt søk etter syntaktiske funksjonar som subjekt og objekt, men òg etter argumentposisjonar, som svarar til semantiske roller, gjer at ei vidare mengd døme blir funne. Det gjeld til dømes setningar der *bil* som argument 2 (patiens) ikkje er objekt, men subjekt til passiv eller hovud for eit partisipp; sjå fire av dei 85 døma på *selge bil* i fig. 10, der berre det siste har *bil* som objekt til aktiv *selge*.

Grammatikken NorGram, som har gjeve opphav til dei syntaktiske analysane i NorGramBank, inneheld òg informasjon om mange fleirordsuttrykk, til dømes grammatiske kollokasjonar med selekterte preposisjonar eller partiklar, og visse idiom. Ein kan søkje etter døme på slike grammatikk-analyserte fleirordsuttrykk med templat som **V-mwe(@V)**

Nye biler som selges i dag, har en helt annen sikkerhet enn gamle biler.
Det betyr at det selges flere biler med lavere avgift, som igjen betyr at flere betaler avgift.
I klimaforliket fra juni i 2012 ble det enstemmig definert - med Fremskrittspartiets subsidiære tilslutning - et tak i 2017, eller 50 000 solgte biler.
Det er altså fire årstall som de som produserer drivstoff, importerer biler og selger biler, forholder seg til.

Figur 10. Ulike døme på bil *som argument 2 til* selge.



Figur 11. Templatet V-mwe(@V) med gripe *som parameterverdi.*

Count	#p: atom
3920	gripe*inn
1817	gripe*fatt*i
1806	gripe*etter
1618	gripe*om
1349	gripe*til
1033	gripe#tak*i
304	gripe*an
11	gripe*fatt*om

Figur 12. Søkjeresultat: grammatikk-analyserte fleirordsuttrykk med gripe *som hovudord.*

og **N-mwe(@N)**, som finn fleirordsuttrykk med eit bestemt verb eller substantiv som hovudord. Fig. 11 viser templatet **V-mwe(@V)** med verbet *gripe* som parameterverdi. Dette søket gjev treffa i fig. 12.

I slike høve blir ikkje frekvens eit kriterium for brukaren for å identifisere fleirordsuttrykk, sidan dei allereie er identifiserte i NorGram. Men det er ikkje uvanleg at grammatikken har funne meir enn ein analyse av slike konstruksjonar, sidan mange fleirordsuttrykk med ikkje-komposisjonelt innhald i tillegg vil ha ein gyldig komposisjonell analyse. *Han tenkjer på hytta* kan både tyde at han har hytta i tankane (grammatisk kollokasjon med selektert preposisjon), og at han tenkjer når han er på hytta (fri kombinasjon med adverbial preposisjonsfrase). Analysen i trebanken vil då

ofte vere vald på statistisk grunnlag. Dette inneber at søk etter slike grammatikk-analyserte fleirordsuttrykk ofte bør supplerast med søk etter tilsvarande frie kombinasjonar med komposisjonelt innhald, både fordi statistikken kan ha valt feil i nokre høve, og fordi det kan finnast fleirordsuttrykk som ikkje er dekte av grammatikken. Det finst òg templat for søk etter slike frie kombinasjonar.

I avsnitt 2 er den systemorienterte retninga nemnd, der kollokasjonar blir analyserte med eit hovudord og ein kollokator. I døme som *le hjarteleg, nekte blankt, tvile sterkt* og *protestere heftig* er verba hovudord. For at ein skal kunne identifisere slike, er det viktig å kunne jamføre med alternative moglege kollokatorar ved dei same verba og sjå på frekvensar og tydingar. Sidan trebanksøk kan referere til ordklassar og syntaktiske relasjonar, blir det mogleg å finne alle slike kandidatar gjennom eit enkelt søk etter ord i den relevante syntaktiske funksjonen – ein treng ikkje å søkje etter alle moglege kollokatorar enkeltvis. Templatet for å finne adjektiv som står adverbialt til eit bestemt verb, er **V-adverbialadj(@V)**. Eit søk med verbet *le* som parameterverdi gjev 6016 treff fordelte på 808 adjektiv; fig. 13 viser toppen av frekvenslista, der *hjertelig* kjem på tredje plass.

Count	#verb: atom	#w: value
1273	le	høyt
289	le	lavt
284	le	hjertelig
214	le	godt
143	le	hånlig
138	le	kort
123	le	rått
94	le	nervøst
88	le	hysterisk
84	le*av	høyt

Figur 13. Toppen av søkjeresultatet: dei mest frekvente adjektiva som står adverbialt til verbet le.

5. Oppsummering

Utgangspunktet for denne artikkelen har vore korleis ein definerer og studerer kollokasjonar i ulike tradisjonar, og korleis dei to verktøya Korpuskel-leks og NorGramBank kan nyttast i fraseologiske studiar av kollokasjonar, særleg med tanke på dei som kan vere relevante å ta med i ordbøker.

Både i den korpusbaserte tradisjonen og den systemorienterte tradisjonen ser forskarane på kollokasjonar som to eller fleire ord som «heng saman» som ei leksikalsk eining, og at kollokasjonar er ein slag friare forbindelse mellom ord enn t.d. idiom, der skilnaden frå meir frie til meir faste uttrykk heng saman med frekvens, med om tydinga til eininga er

komposisjonell, og med om ein kan variere delar av uttrykket utan å endre tydinga.

Noko av det som skil retningane, er at den systemorienterte tradisjonen har eit meir restriktivt syn på kva som er å rekne som kollokasjonar enn den korpusbaserte. I den systemorienterte er det sentralt å utvikle kriterium for å finne grupper av samanliknbare einingar med same semantiske relasjon mellom komponentane for å få innsikt i språksystemet. I denne artikkelen argumenterer vi snarare for ein fraseologisk typologi med ei idiomklyngje, ei ordspråksklyngje og ei kollokasjonsklyngje som tre sekkestorleikar, der vi i staden for å prøve å skilje skarpt mellom kategoriane gjer eit poeng av å vise korleis dei ofte overlappar.

Den korpusbaserte tradisjonen oppfattar kollokasjonar som noko som er uavhengig av syntaktisk analyse, og som først og fremst dreier seg om den lineære plasseringa av ord i høve til kvarandre og at to (eller fleire) ord førekjem saman oftare enn ein skulle vente ut frå frekvensane deira kvar for seg. Dette er metoden som ligg bak bruken av Korpuskelleks, der vi har vist døme på å søkje etter ordsekvensar med hjelp av regulære uttrykk eller gjennom statistiske assosiasjonsmål. Regulære uttrykk gjer det enkelt å finne dei mest frekvente mønstera av ord, medan assosiasjonsmål kan løfte fram samførekomstar som ikkje nødvendigvis er høgfrekvente, men som likevel er statistisk sterke kollokasjonar. Denne tilnærminga er enkel og effektiv, men seier berre at visse ord har ein tendens til å førekome saman. Det blir leksikografen sin jobb å vurdere om dette er ordsekvensar som står som typiske døme på ei bestemt tyding av eit ord, eller om kollokasjonen representerer ei eiga tyding som bør få sitt eige oppslag i ordboka. Dette er altså ein bruk av omgrepet *kollokasjon* som er langt meir open enn den vi finn i den systemorienterte tradisjonen.

Til samanlikning lèt NorGramBank oss søkje etter grammatiske mønster som er knytte til eit søkjeord, gjennom søk på syntaktisk struktur og grammatiske relasjonar. Dette opnar for å kunne finne syntaktiske variantar av same kollokasjon, som til dømes predikativ eller attributiv plassering av adjektiv, og det gjer det mogleg å søkje etter alle ord i ein viss syntaktisk relasjon til eit hovudord som eit grunnlag for å skilje ut dei som bør analyserast som kollokatorar (som i dømet *le hjarteleg*). NorGramBank inneheld dessutan informasjon om mange fleirordsuttrykk, til dømes visse idiom, og grammatiske kollokasjonar med selekterte preposisjonar eller partiklar, som ein då kan søkje direkte etter.

Slike trebank-søk kan såleis gje eit inventar av kollokasjonskandidatar med ord som ein så kan studere nærare i det mykje større Korpuskel-korpuset. Frå ståstaden til leksikografen utfyller dermed desse to ressursane kvarandre.

Litteratur

Ordbøker

Bokmålsordboka og *Nynorskordboka*. Språkrådet og Universitetet i Bergen. <ordbøkene.no> (april 2022).

Islex. Árni Magnússon-instituttet for islandske studiar, Det Danske Sprog- og Litteraturselskab, Universitetet i Bergen, Göteborgs universitet, Fróðskaparsetur Føroya og Helsingfors universitet. <islex.no> (april 2022)

NAOB = Det Norske Akademis ordbok. <naob.no> (april 2022).

NLO = *Nordisk leksikografisk ordbok*. Bergenholtz, Henning, Ilse Cantell, Ruth Vatvedt Fjeld, Dag Gundersen, Jon Hilmar Jonsson og Bo Svensén (red.). 1997. Skrifter utgitt av Nordisk forening for leksikografi. Universitetsforlaget.

Annan litteratur

Andersen, Gisle. 2011. Evaluation of alternative association measures for extraction of terminology based on a large Norwegian corpus. *SYNAPS – A Journal of Professional Communication* 26/2011, 62–68.

Andersen, Gisle. 2020. Phraseology in a cross-linguistic perspective: A diachronic and corpus-based account. *Corpus Linguistics and Linguistic Theory* 18(2), 365–389. https://doi.org/10.1515/cllt-2019-0057.

Barkema, Henk. 1996. Idiomaticity and terminology: A multi-dimensional descriptive model. *Studia Linguistica 50* (2), 125–160.

De Jong, Franciska, Bente Maegaard, Koenraad De Smedt, Darja Fišer & Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. I: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association (ELRA), 3259–3264. https://aclanthology.org/L18-1515.pdf

De Smedt, Koenraad, Gunn Inger Lyse Samdal, Rune Kyrkjebø, Hemed Ali Al Ruwehy, Øyvind Liland Gjesdal, Victoria Rosén & Paul Meurer. 2015. The CLARINO Bergen Centre: Development and

Deployment. I: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland, 1–12. Linköping Electronic Conference Proceedings 123:1. http://www.ep.liu.se/ecp/123/001/ecp15123001.pdf

Dyvik, Helge, Gyri Smørdal Losnegaard & Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. I: *Representation and parsing of multiword expressions: Current trends*, red. Yannick Parmentier & Jakub Waszczuk, 69–108. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.2579037

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse & Martha Thunes. 2016. NorGramBank: A 'Deep' Treebank for Norwegian. I: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, red. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis, 3555–3562. Portorož, Slovenia: ELRA.

Firth, John Rupert. 1957. A synopsis of linguistic theory 1930-1955 (1957). I: *Studies in Linguistic Analysis,* ei spesialutgåve av Philological Society, red. J.R. Firth, 1–32, Oxford: Blackwell.

Fjeld, Ruth Vatvedt. 2009. Leksikografisk dokumentasjon av flerordsenheter i norsk. *LexicoNordica 16*: 103–118.

Fjeld, Ruth Vatvedt og Lars S. Vikør. 2008. *Ord og ordbøker: Ei innføring i leksikologi og leksikografi*. Kristiansand: Høyskoleforlaget.

Evert, Stefan. 2004. The Statistics of Word Co-occurrences: Word Pairs and Collocations. Ph.D. thesis, IMS, University of Stuttgart.

Horvati, Eszter. 2005. Automatisk gjenkjenning av norske kollokasjoner. Masteroppgåve, Universitetet i Oslo

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1), 7–36. Springer.

Lyse, Gunn Inger & Gisle Andersen. 2012. Collocations and statistical analysis of n-grams. I: *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian*, red. Gisle Andersen, 79–109. Studies in Corpus Linguistics 49. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.49.05lys

Lyse, Gunn Inger. 2020. Ut med *adamsslekt* og inn med *arveprinsesse*? Leksikografiske metodar i revisjonen av *Bokmålsordboka* og *Nynorskordboka*. I: *Nordiska studier i lexikografi 15*, 215–224.

Malmgren, Sven-Göran. 2003. Begå eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet. *Språk och Stil 13* (ny följd), 123–168.

Meurer, Paul. 2012. Corpuscle – a new corpus management platform for annotated corpora. I: *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian*, red. Gisle Andersen, 29–50. Studies in Corpus Linguistics 49. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.49.02meu

Pedersen, Ted, Satanjeev Banerjee, Bridget McInnes, Saiyam Kohli, Mahesh Joshi & Ying Liu. 2011. The Ngram statistics package (text:: nsp): A flexible tool for identifying ngrams, collocations, and word associations. I: *Proceedings of the workshop on multiword expressions: From parsing and generation to the real world*, red. Paul Cook et al., 131–133.

Rauset, Margunn. 2019. Bokmålsordboka og Nynorskordboka – einegga, toegga eller siamesiske tvillingar? *LexicoNordica 26*: 155–175.

Rauset, Margunn. 2022. *Frasar til besvær? Studiar av norm og bruk i norsk fraseologi*. Doktorgradsavhandling, Universitetet i Bergen. https://hdl.handle.net/11250/2992943

Rauset, Margunn, Gyri Smørdal Losnegaard, Helge Dyvik, Paul Meurer, Rune Kyrkjebø & Koenraad De Smedt. 2022. Words, words! Resources and tools for lexicography at the CLARINO Bergen Centre. I: *CLARIN. The infrastructure for language resources*, red. Darja Fišer og Andreas Witt, 537–560. Berlin: deGruyter.

Revisjonsprosjektet. 2022. Revisjonen av Bokmålsordboka og Nynorskordboka. Oppdatert 11. mars 2022. https://revisjonsprosjektet.no.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer & Helge Dyvik. An open infrastructure for advanced treebanking. I: Jan Hajič, Koenraad De Smedt, Marko Tadić & António Branco (red.) *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, 22–29, Istanbul, Turkey, May 2012.

Rosqvist, Bodil. 2010. Ordförbindelser i SAOB – En undersøkning av beskrivningen av kollokationer. *Nordiska Studier i Lexikografi 10*, 455–469.

Rosqvist, Bodil. 2014. Hårt arbete *och* sträng vila. *I: Svenska kollokationer i lexikografisk och lexikologisk belysning*. Göteborgs universitet. http://hdl.handle.net/2077/35259

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Ann Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. I: *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science* 2276, red. Gelbukh, A. Berlin og Heidelberg: Springer, 1–15. https://doi.org/10.1007/3-540-45715-1_1

Selback, Bente. 2020. «Å nei, det ordet er ikkje lov på nynorsk!» Eller ...? Om parallell redigering av to norske ordbøker. I: *Nordiska studier i lexikografi 15*, 297–305.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. Textbooks in Language Sciences 7. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3735821

Svensén, Bo. 2004. *Handbok i lexikografi : Ordböcker och ordboksarbete i teori och praktik*. Stockholm: Norstedts Akademiska Förlag.

Unsplash. Bilete av R2-D2 og C-3PO. Foto: Mulyadi. Sett 18.04.22. unpslash.com/photos/JEfwd_okQGE

Wikipedia, s.v. «R2-D2,» lesen 18.04.2022, https://no.wikipedia.org/wiki/R2-D2.

Abstract

In the revision of the Norwegian dictionaries *Bokmålsordboka* and *Nynorskordboka*, the lexicographer needs efficient and user-friendly tools that can point to typical uses of individual words, and frequent phrases associated with them. This article presents and discusses the definition and study of collocations in different traditions, as multiple approaches to terms such as 'collocation' and 'multiword expression' meet in the practical field of lexicography. The main part of the article presents available digital tools for finding and studying collocations that may be candidates for inclusion in a dictionary. Corpuscle-lex is a corpus search tool adapted for lexicographical use. The user may search for collocations using statistical association measures and regular expressions across multiple corpora. NorGramBank is a treebank with syntactically analysed text. Hence a search may refer to syntactic structures and grammatical rela-

tions and not only to the linear positions of words in a text. NorGramBank comes with search templates enabling the lexicographer to make complex searches for the grammatical patterns around a word without engaging directly with query language expressions.

Gunn Inger Lyse
Universitetet i Bergen
Institutt for lingvistiske, litterære
og estetiske studium
Postboks 7805
NO-5020 Bergen
gunn.lyse@uib.no

Margunn Rauset
Universitetet i Bergen
Institutt for lingvistiske, litterære
og estetiske studium
Postboks 7805
NO-5020 Bergen
margunn.rauset@uib.no

Helge Dyvik
Universitetet i Bergen
Institutt for lingvistiske, litterære
og estetiske studium
Postboks 7805
NO-5020 Bergen
helge.dyvik@uib.no