Lars G. Bagøien Johnsen

Å bruke språkteknologi til å undersøke medisinske ord

Michael 2023; 20: Supplement 31: 133–42.

Nasjonalbibliotekets digitaliserte samling av tekster utgjør en kilde til informasjon om både bøker og deres innhold. I dette kapitlet viser jeg hvordan slike kilder kan brukes til å studere medisinsk terminologi. Ved å følge tre medisinske termer skal jeg si noe om begrepet korpus *samt de to tilhørende begrepene* konkordans *og* kollokasjon. *Korpuset gir informasjon om bruk av ordene over tid, mens konkordanser og kollokasjoner gir oss informasjon om termenes bruk og til dels betydning.*

Data om medisinsk språkbruk gjør det mulig å si noe om hvordan ord brukes, hvordan alternativer forholder seg til hverandre og eventuelle endringer i bruk. Med algoritmer og digitale språkdata åpner det seg muligheter for å analysere ords forekomst i tekst. Ideelt sett kunne vi tenkt oss å få med alt som blir sagt og skrevet innen medisinsk behandling. Daglig produseres store mengder tekst blant annet i pasientjournaler. Med tilgang til all bruk av ord og uttrykk i alle relevante sammenhenger ville språkbildet være komplett. Juridiske og etiske betraktninger gjør at dette er umulig å få til. Det er likevel et stort tilfang av digitalt tilgjengelige tekster som kan benyttes til formålet, både på internett og i fra Nasjonalbibliotekets digitale bibliotek (1), som inkluderer bøker og tidsskrift[1]. Indirekte gir tekstene oss informasjon om bruk av termer gjennom den språklige konteksten de står i, og gjennom bruksfrekvens over tid får vi informasjon om mulige trender.

Korpus og informasjonsflyt

Vi skal se på hvordan man dynamisk kan bygge korpus for medisinsk tekst, som kan brukes i terminologisk arbeid. Med *korpus* forstår vi en samling

1 Nasjonalbiblioteket har digitalisert nesten samtlige bøker, mens tidsskrift er fremdeles under digitalisering.

tekster som har et sett egenskaper som gjør at vi kan trekke slutninger om bruksmønstre, betydning og grammatikk, basert på hvordan ordene opptrer i det. Det ligger informasjon både i selve byggeprosessen og hvordan det brukes i etterkant.

I leksikografisk og terminologisk arbeid er vi først og fremst ute etter ords egenskaper, ord som eksisterer i flere fagfelt og ord som har forskjellige egenskaper. I norsk vil ordet *morfologi* ha en betydning i både biologi og lingvistikk, men det er liten overførbarhet mellom de to domenene. For ordet *morfologi* betyr det at det må studeres i relevante kontekster. For å snakke om lingvistiske termer, bør korpuset derfor konstrueres for å fange inn den bruken vi er ute etter, at vi søker tekster innen lingvistikk og tilsvarende søker i medisinske tekster for medisinske termer. Men ofte er enkelte termer typiske for et bestemt fagfelt, som *canthus*[2] og *cochlea*, og da er ikke alltid korpusets definisjon som sådan like sentralt for å finne informasjon om termene selv. Medisinen benytter termer som også har en allmenn bruk, for eksempel *øye* og *hode*, og spesielt i slike tilfeller er det viktig å ha et korpus som definerer bruken av ordene i en medisinsk kontekst.

Ved Nasjonalbiblioteket er hele den tekstbaserte norske kulturarven digitalisert og gjort tilgjengelig både for lesing og for dataanalyse. I skrivende stund (november 2022) er over 600 000 bøker tilgjengelig digitalt, over 4 millioner aviser og ca. 100 000 tidsskrifter. For bøker er samlingen nesten komplett, aviser og tidsskrift er under kontinuerlig digitalisering.

Visning av bøker til allmennheten er begrenset av åndsverkloven, men det meste av dataanalysen vil ikke utfordre opphavsrettigheter. De tilgjengelige dataene om tekstene inkluderer for eksempel frekvenser og statistikk basert på kobling mellom metadata og ord. Slike data inneholder informasjon av en høykulturell verdi, som ikke bryter med lovverket.

Digital infrastruktur

Nasjonalbibliotekets infrastruktur for tekstanalyse består av to helt essensielle komponenter. Under digitaliseringen overføres en bok fra papir og trykksverte til digital tekst via bilder og *Optical Character Recognition* (OCR), som tar oss fra bilde til tekst. Samtidig knyttes metadata som foreligger om boken, til det digitale objektet. For å sikre en stabil gjenfinning av bøkene over tid innførte Nasjonalbiblioteket, i samarbeid med det finske nasjonalbiblioteket, en spesiell måte å referere til digitale objekter på med bruk av såkalt *Universal Resource Name* (URN), som sikrer stabil referanse til digitale objekter. Objektene har en helt bestemt intern struktur som identifiserer

2 Takk til Ole Kristian Våge for termen *canthus*.

objektets digitaliseringshistorie. En URN har også informasjon om objektets tilblivelse, det vil si når det ble digitalisert, hva som ble digitalisert og hvor det ble digitalisert. En URN kan for eksempel se slik ut: URN:NBN:no-nb_digibok_2012091906028. Her er det flere informasjonsbiter, av spesiell interesse er *digibok*, som betyr at objektet kommer fra en monografi med oppføring i nasjonalbibliografien[3], og der de åtte første sifrene forteller oss at boken er digitalisert 19.9.2012. De fem siste sifrene er et løpenummer. URN-ene gjør det mulig å arbeide med kopier av kopier av det digitaliserte materialet uten at man mister referansen til den originale teksten. På den måten utgjør URN-systemet et helt sentralt element i Nasjonalbibliotekets infrastruktur for digitale objekter. URN-ene sikrer også at alle undersøkelser som gjøres med dem som utgangspunkt, kan utføres på nytt på et senere tidspunkt. Dette er et viktig poeng i en vitenskapelig undersøkelse av språket.

Den digitale infrastrukturen består også av algoritmer som kan analysere tekster og gjøre søk i dem og på annet vis trekke ut informasjon. I tillegg til nettbiblioteket har Nasjonalbiblioteket et laboratorium for digital humaniora (DH-lab) som tilbyr programmatiske innganger til data via programmeringsspråk som Python og R, samt forskjellige apper som kan tilpasses forskningsprosjekter. Nedenfor har vi benyttet DH-labens muligheter for å hente ut informasjon.

Korpusbygging

Alle bøker og tidsskrifter er utstyrt med to typer metadata, det gjelder både de digitale med URN og de på papir[4]. Det ene er kataloginformasjon, som publikasjonsår, tittel etc, og det andre er klassifikasjon av innholdet, som gjøres med emneord og Deweys desimalsystem. I Deweys system er medisin innen desimaltall 610[5]. Vi kan derfor bygge et korpus med utgangspunkt i desimalsystemet, altså samle alle tekster som er blitt klassifisert med de to innledende sifrene 61. I det følgende viser vi hva vi kan gjøre med den informasjonen. Vi skal også bygge et korpus fra innholdsord, der tekstene er definert ut fra hva de inneholder fremfor hvordan de er klassifisert. I de følgende avsnittene skal vi bygge korpus etter begge metodene, og starter med å definere ut fra innhold.

3 Listen over alle bøker utgitt i Norge som vedlikeholdes av bibliotekene.

4 Mens URN-en gjelder for et bestemt eksemplar av en bok eller tidsskrift, vil metadata fra katalogene gjelde utgivelsen.

5 Søk for eksempel på Webdewey for en oversikt over Deweys klassifikasjonssystem.

Canthus, kantus og øyekrok

Vi samler tekster som inneholder ordet *canthus*. Det er en spesifikk term som beskriver øyekroken, og har termene *øyekrok* og *kantus* som alternativer. Termene *canthus* og *kantus* er homofone, altså at de har samme uttale, og også homofone med *cantus* som tilhører musikk-terminologien. De to første formene er unike for medisinsk bruk, altså at de kun benyttes i medisinske tekster.

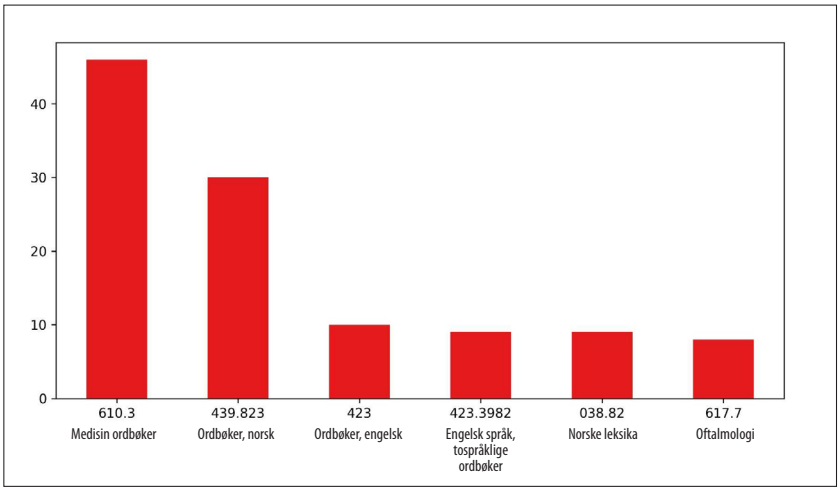
I det første korpuset starter vi med å definere samlingen av tekster som de tekster som inneholder selve ordet. Deretter ser vi på hva slags metadata bøkene i det korpuset har. Med DH-lab ber vi programmatisk om en liste med bøker (URN-er) som er publisert mellom 1920 og 2022, som har ordet *canthus* i teksten. Det gir oss en liste på 180 bøker.[6] I tabell 1 viser vi hvordan korpuset fortoner seg (tabell 1). Hver rad er en tekst, der kolonnene har data om teksten. Første kolonne, kalt urn inneholder URN-en. De to har med Deweys desimalnummer (kolonne ddc) og emneord (subjects). Resultatet viser en blanding av medisinsk litteratur, ordbøker og leksika.

Vi kan nå sjekke hva slags informasjon som ligger her om bøkene. Er det slik at *canthus* bare er knyttet til medisin eller kan ordet også forekomme i andre kontekster? Og hvordan fordeler kategoriene seg? I figur 1 er det

urn	title	authors	year	Ddc	Subjects
URN:NBN:no-nb_digibok_2012091906028	Store medisinske leksikon. 3 : I-M		2007	610.3 / 610.3 / 610.3	Medicine / Encyclopedias / Medisin / leksika
URN:NBN:no-nb_digibok_2008062304088	Medisinsk ordbok	Kåss , Erik / Hauge , Anton / Welle-Strand , G...	1992	610.3	Bokmål / Norsk språk / Medisin / Terminologi /...
URN:NBN:no-nb_digibok_2021062448623	Klassifikasjon av oftalmologiske diagnoser og ...		1975	617.7	Eye Diseases / classification
URN:NBN:no-nb_digibok_2017030248139	Oftalmologi : nordisk lærebok og atlas	Høvding , Gunnar	2004	617.7 / 617.7	Ophthalmology / Eye Diseases / augesjukdommar ...

Tabell 1. Deler av et korpus bygd over termen canthus

6 Søket kan også gjøres mot NB-digital (bokhylla.no eller nb.no) med omtrent samme resultat.



Figur 1. Deweys desimaltall for korpus basert på canthus

gjengitt en statistikk over de mest frekvente Dewey-desimalkodene (figur 1). De er talt opp ved å telle opp alle forekomstene i relevant kolonne i korpuset. Det er verdt å merke seg at ikke alle bøkene er klassifisert, av de 180 bøkene er det 45 som ikke har noen klassifikasjon.

Interessant nok er alle de fem numrene med høyest frekvens knyttet til medisinske ordbøker[7], via kodene 610.3 (medisinske ordbøker), 4XX (språkvitenskap/ordbøker) og 038.82 (norske leksika), mens på plass nummer seks kommer Dewey 617.7 (oftalmologi). Tilsvarende øvelse for ordet *cochlea* gir de faglige kodene 617.8 (otologi og audiologi), 612.8 (nervesystemet) før 610.3 (medisinske ordbøker).

Denne øvelsen gir oss en mulighet til å undersøke enkeltord gjennom metadata. Alle ord som vi kan anta har en rimelig snever bruk (eller spesifikke kontekster), vil kunne egne seg til en slik metode. I vårt tilfelle fikk vi vite at *canthus* og *cochlea* er entydig medisinske, og at termene ser ut til å trives godt i ordbøker. Ord med større bruksområder vil gi korresponderende flere koder.

Bygge et korpus og undersøke det

En mer tradisjonell korpuslingvistisk metode er å først definere korpuset med metadata og så studere fordelingen av ord inne i korpuset. I avsnittet over fløt informasjon fra ordet til metadata, som vi så benyttet til å danne

7 Som det fremgår av tabell 1, er noen koder ført opp flere ganger for en enkelt bok. Listene er redusert slik at det kun er én unik klassifikasjon per rad som telles og vises i figur 1.

oss et inntrykk av egenskapene til ordet. Nå snur vi om på informasjonsflyten og definerer et korpus med Deweys desimalkode og bruker den til å studere egenskaper ved ordene *canthus*, *kantus* og *øyekrok*. Ved å danne et korpus med utgangspunkt i medisinsk litteratur (Dewey 61X), vil vi kunne påberope oss en viss autoritet i påstander om bruken av de ordene.

I eksemplet under er det bygget et korpus basert på over 12 000 bøker. Med Deweys system kan det også bygges korpus over medisinske tidsskrifter, men nedenfor begrenser vi oss til et korpus over bøker basert på Deweys desimalkode 61X. DH-lab har apper som gjør det forholdsvis enkelt å bygge korpus og undersøke dem. Alle resultatene i dette kapitlet kan reproduseres med en av disse appene. Den grunnleggende infrastrukturen består av databaser og REST-API-er mot databasene[8]. I praksis betyr det at korpuset kan ligge på datamaskin (server) forskjellig fra den maskinen man arbeider på. Fra API-ene er det bygget programbibliotek som gjør det mulig å lage brukervennlige grensesnitt for leksikografer.

Undersøkelsen deles i to deler. Først ser vi på bruksfrekvens for de tre ordene, deretter søkes det i korpuset etter forekomster (konkordanser) og bygges kollokasjoner.

Trender

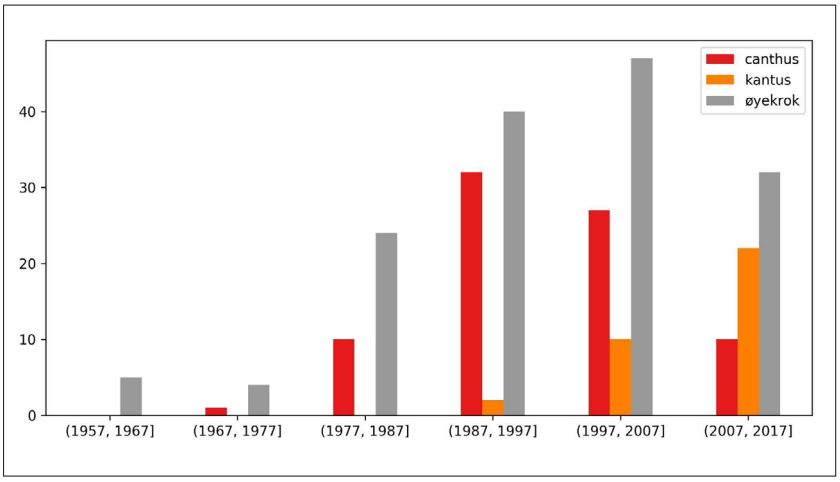
Det er to måter å nærme seg et korpus på. Den ene er å la korpuset, selve tekstene, være implisitt, og den andre er å hente ut URN-ene (dvs. tekstreferansene) eksplisitt. Med bruksfrekvenser over tid bruker vi DH-labens innganger til å telle ord i et implisitt korpus, altså antallet ganger et ord forekommer i korpuset.

Ved å telle de tre termene *canthus*, *kantus* og *øyekrok* i et korpus definert ved Dewey 61X, fra 1950 til 2020, målform bokmål, trer det frem et tydelig mønster som vist i figur 2. Her er det tydelig at ordet *øyekrok* var den hyppigste termen i starten på perioden, der både *canthus* og *øyekrok* øker i frekvens frem mot slutten av 1980-årene. Samtidig starter *kantus* sitt løp, og frem mot vår tid ser det ut til å overta stafettpinnen. Diagrammet i figur 2 er laget med en variant av NB n-gram[9] (1), der det er mulig å begrense trendlinjene til metadata.

Et viktig poeng kan være at det er kun de tre ordformene som er gjenstand for telling. Ordet *øyekrok* har flere former, som *øykroken* og *øyekroker*,

8 For å dele, trenger vi grensesnitt for utveksling av informasjon mellom datamaskiner, såkalte *programmeringsgrensesnitt* eller API-er (Application Programming Interface).

9 Nasjonalbibliotekets søketjeneste NB N-gram gir mulighet til å finne og sammenligne ordfrekvenser, for eksempel når og hvor ofte ord forekommer i bøker og aviser i et historisk perspektiv. Tjenesten er basert på digitaliserte bøker og aviser i Nasjonalbiblioteket, og inneholder materiale fra 1810 til i dag.



Figur 2. Oversikt over bruksfrekvens for canthus, kantus og øyekrok i et medisinsk korpus. X-aksen er gruppering over perioder, og y-aksen viser samlede antall forekomster i hver periode.

mens *canthus* ser ikke ut til å ha noen morfologi. For *kantus* forekommer *kantussen.*

Konkordanser

Konkordanser for ord (også kalt keyword in context, KWIC) består av mindre tekstutdrag som viser en liten kontekst for ordet (2). Ved å hente ut alle URN-ene kan vi ved hjelp av dem søke i enkelttekstene (her er korpuset tidsskrift publisert etter 1950 og klassifisert med Dewey 61X) og finne konkordanser, som gir en viktig informasjon for leksikografisk arbeid. Konkordansenes små tekstutdrag[10] er ofte nok for å se hvilken betydning ordet er brukt i og hva det står sammen med. Et søk etter ordene *canthus* og *øyekrok* gir 212 treff på første ordet i korpuset og 247 treff på siste[11].

Sammen med konkordansene leveres også litt metadata fra teksten de er hentet fra, og i praktisk bruk vil det være mulig å klikke seg inn på Nasjonalbibliotekets nettbibliotek for eventuelt å kunne lese mer fra den aktuelle teksten. Eksempler:

10 DH-lab gir maksimum 20 ord i konteksten, et konservativt antall ord som ikke skal utfordre opphavsrett.

11 En feilkilde med korpus definert gjennom Dewey er at klassifikasjonen er mangelfull før 1950. Etter den tid har de fleste bøker fått kode, men det åpner for flere måter å bygge korpus, avhengig av hva man er ute etter.

... Og jo nærmere **canthus** den angrepne kjertel sitter desto sterkere pleier kemosen og i det hele betendelsessymptomene å være ....[12]

... Neseroten var breddeforøket, og pasienten hadde en hevelse som strakte seg ned mot mediale **canthus** bilateralt, mest uttalt… [13]

I kloroformnarkose spaltedes huden fra **canthus** externus ved snit langs orbitalranden til dens midte, forsigtig lagdes derpaa snit gjennem...[14]

... Et arr ved laterale **øyekrok** kan være skjemmende, men ved suturering med tynn tråd og lukking i tre lag...[15]

... at stemmen sprekk og tåran enno blinke' i ein **øyekrok** Og dæm skjelv' litt når dæm sjer pendel'n i...[16]

... tårekanalen med lett trykk med en finger i mediale **øyekrok**. For enkelte gamle kan dette imidlertid være noe vanskelig...[17]

Som det fremgår av konkordansene, er det en konkordans for øyekrok som skiller seg litt ut fra de andre mer teknisk medisinske utdragene.

Kollokasjoner

Kollokasjoner er en «forbindelse av to eller flere ord som vanligvis opptrer sammen, for eksempel *gjøre fremskritt* og *felle en dom*», ifølge ordboka (naob.no). Vi kan også kalle det for aggregerte konkordanser. I DH-lab bruker vi begrepet i en litt videre betydning enn slik det er definert i ordboka, slik at kollokatene for et ord ikke trenger å stå ved siden av ordet, men at det holder at de befinner seg innenfor en kontekst, og er assosiert.

Med kollokasjoner kan man se på større eller mindre kontekstvindu, og også begrense dem til bare noen få ord til høyre eller venstre for å undersøke sammenhengen mellom ord og hva de typisk opptrer med. Kollokatene, altså de ordene som antas å være knyttet til et målord, og eventuell endring i dem kan fortelle oss noe om forskjellige betydninger. I DH-lab får kollokatene en relevansverdi basert på frekvensforskjellen mellom forekomster i konkordansene (kontekst) og korpuset. I termer av prosent, om ordet har

12 Tidsskrift for Den norske lægeforening 1937; 57: 200. Tidsskrift for Den norske legeforening (trykt utg.) : The Journal of the Norwegian Medical Association. 1937 Vol. 57 Nr. 4 (nb.no)

13 Tidsskrift for Den norske lægeforening 2000; 120: 2252. https://www.nb.no/items/d64561a30ce438034e2977f2b5bd6f54?page=23

14 Tidsskrift for Den norske lægeforening 1914; 34: 442. https://www.nb.no/items/bd4b85886592a98d361e0bdce72b7f15?page=465

15 Tidsskrift for Den norske lægeforening 1990; 110: 3098. https://www.nb.no/items/5ca029472701277f8c64ce150f9f9325?page=21

16 Tidsskrift for Den norske lægeforening 1992; 112: 3126. Tidsskrift for Den norske legeforening (trykt utg.) : The Journal of the Norwegian Medical Association. 1992 Vol. 112 Nr. 24 (nb.no)

17 Tidsskrift for Den norske lægeforening 1997; 117: 2021. https://www.nb.no/items/f955f099fc128440542d391ea5dcdf31?page=19

en frekvens på 2 % av alle ordene i kontekst for målordet, samtidig som det opptrer i 0,5 % i korpuset ellers, kan vi si at det opptrer fire ganger så ofte med målordet enn utenfor. Frekvensforskjellen gir et mål på assosiasjonen mellom kollokatet og målordet, jo høyere desto sterkere assosiert. Det bør skytes inn at målet er følsomt for lavfrekvente ord. I praktisk bruk er det nødvendig å sjekke den faktiske frekvensen.

For *øyekrok* og *canthus* er de mest relevante kollokatene (fem på topp) angitt i tabell 2 og tabell 3. Den siste kolonnen inneholder relevanstallet. Kollokatet øyelokk forekommer bare 22 ganger i omgivelsene for *øyekrok*, men det er ca. 8000 ganger hyppigere enn frekvensen det har i det samlede materialet. Interessant nok er *øyelokk* også et høyt rangert kollokat for *canthus*.

ord	frekvens	relevans
øyelokket	9	10142
øyelege	9	9130
øyelokk	22	8871
Pannen	6	5813
fremmedlegemer	6	5522

Tabell 2. Kollokater for øyekrok, dvs. forbindelse av to eller flere ord som vanligvis opptrer sammen med øyekrok. Relevanstallet står til høyre: for eksempel kollokatet øyelokk forekommer bare 22 ganger i omgivelsene for øyekrok, men det er ca. 8000 ganger hyppigere enn frekvensen det har i det samlede materialet.

ord	frekvens	relevans
laterale	16	21645
directly	32	14495
enh	15	10125
øyelokk	20	9740
edge	16	8392

Tabell 3. Kollokater for canthus, dvs. forbindelse av to eller flere ord som vanligvis opptrer sammen med canthus. Relevanstallet står til høyre: for eksempel kollokatet laterale forekommer 16 ganger i omgivelsene for canthus, men det er ca. 21.000 ganger hyppigere enn frekvensen det har i det samlede materialet.

Oppsummering

I artikkelen har vi sett på hvordan ressursene som tilbys ved Nasjonalbiblioteket, kan benyttes i medisinsk-terminologisk arbeid. Informasjonsflyten kan gå fra ord til tekst og så til klassifikasjonsdata, og omvendt fra klassifikasjon til tekster og så til ord og termer. Selv om ikke alt som er publisert innen medisin er tilgjengeliggjort, er det likevel store mengder informasjon som kan hentes ut.

Ressurstilgang

Ressursene som er beskrevet over, er under stadig utvikling. Relevante nettadresser kan være nettstedet for DH-lab (https://www.nb.no/dh-lab), som har pekere til flere apper som kan være av nytte i et leksikografisk arbeid.

Litteratur

1. Birkenes MB, Johnsen LG, Lindstad AM et al. From digital library to n-grams: NB N-gram. I: P*roceedings of the 20th Nordic Conference of Computational Linguistics*. Linköping: Linköping University Electronic Press, 2015: 293–5. https://ep.liu.se/ecp/109/039/ecp15109039.pdf (12.12.2022).
2. Sinclair J, red. *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press, 1991.

Lars G. Bagøien Johnsen
lars.johnsen@nb.no
Nasjonalbiblioteket
Postboks 2674 Solli
0203 Oslo

Lars G. Bagøien Johnsen er forskningsbibliotekar ved Nasjonalbiblioteket i Oslo. Han har doktorgrad i lingvistikk og har siden 2013 spilt en sentral rolle i utviklingen av Nasjonalbibliotekets digitale forskningsinfrastruktur.