OSLa Johannessen, Janne Bondi & Kristin Hagen (red.) Leksikografi og korpus. En hyllest til Ruth Vatvedt Fjeld, Oslo Studies in Language 11(1), 2020. 101–124. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

LEKSIKOGRAFISK BOKMÅLSKORPUS (LBK) – BAKGRUNN OG BRUK

RUTH VATVEDT FJELD, ANDERS NØKLESTAD OG KRISTIN HAGEN

SAMMENDRAG

Denne artikkelen er en introduksjon til Leksikografisk bokmålskorpus (LBK). Vi starter med en historisk oversikt over ordboksarbeid som er utført for norsk språk, og forklarer bakgrunnen for at LBK ble bygd opp på den måten det ble. Deretter gir vi en oversikt over innholdet i korpuset, før vi til slutt viser hvordan man kan søke i korpuset ved hjelp av korpussøkeverktøyet Glossa.

[1] INNLEDNING

Leksikografisk bokmålskorpus (LBK) er et representativt, vektet korpus laget for leksikalsk utforsking av moderne bokmål. LBK inneholder tekster fra perioden 1985 til 2013, i alt omlag 100 millioner ord. Korpuset er bygd opp ved Universitetet i Oslo og lagt inn i søkegrensesnittet Glossa, som er administrert av Tekstlaboratoriet ved Institutt for lingvistiske og nordiske studier på Humanistisk fakultet. Korpuset er fritt tilgjengelig for språkforskning, det kreves kun at man logger seg inn og godkjenner et sett av bruksregler. Prosjektet ble ledet av professor Ruth Vatvedt Fjeld, ved daværende avdeling for bokmålsleksikografi ved Institutt for lingvistiske og nordiske studier, Universitetet i Oslo.

Korpusets design og oppbygning er spesielt tilrettelagt for utforsking av leksikalsk og morfologisk variasjon i moderne norsk bokmål, og er også et godt materiale for å dokumentere endringer i bruken av bokmålsordforrådet over en periode på tretti år. Korpuset er bygd opp etter mønster fra blant annet *Den Danske ordbogs elektroniske korpus* og *British National Corpus*.

Denne artikkelen er tredelt: I kapittel 2 gis det en historisk oversikt over leksikografi i Norge og Norden. Her gjør vi også rede for behovet for et elektronisk, leksikografisk innrettet tekstkorpus av en viss størrelse.

I kapittel 3 beskriver vi en del hindere og problemer som oppsto i arbeidet med å bygge et balansert korpus, og hvilke løsninger som ble valgt.

Kapittel 4 gir en praktisk og instruktiv innføring i bruken av korpusets mange søkemuligheter og hvordan man kan velge ut subkorpus for spesialundersøkelser og få resultatene vist og eventuelt lastet ned i ulike formater.

[2] BAKGRUNN OG METODEVALG

[2.1] *Materiale for tradisjonelt ordboksarbeid*

De fleste ordbøker for norsk språk er skrevet uten en systematisk oppbygd ordboksbase som materiale (jf. Nordisk leksikografisk ordbok, 1997:199), dvs. man har ikke undersøkt tekster etter en gjennomtenkt plan for å finne hvilke ord som skal dokumenteres i ordboka, men har samlet ord og uttrykk etter relativt tilfeldige metoder. Det betyr i praksis at ordbøkene ikke er redigert ut fra et tekstmateriale som er metodisk innsamlet for å dekke et visst ordforråd. Man trenger for eksempel et annerledes tekstmateriale for å lage spesialordbøker enn om man vil lage allmennordbøker som skal dekke hele det ordforrådet en språkkultur normalt bruker.

Tradisjonelt har norske leksikografer tatt som utgangspunkt de oppslagsordene de fant i eldre ordbøker, og lagt til mer eller mindre tilfeldig registrerte nye ord eller uttrykk som de mente manglet. Metoden var å notere seg ord man kom på, leste eller hørte, som man så skrev ned på sedler sammen med diverse andre opplysninger. Sedlene ble så systematisert alfabetisk i arkiver som utgjorde grunnlaget for redigering av den planlagte ordboka. Et slikt materiale ble selvsagt svært preget av den eller de som hadde samlet sedlene.

Et annet problem var at de eksisterende ordene ikke ble undersøkt for bruks- eller betydningsendring, noe som skjer jamt og trutt med ordforrådet i alle levende språk.

Et tidlig unntak fra denne tradisjonen i Norge er *Norsk riksmålsordbok* (Knudsen og Sommerfelt (red.) 1937-1957), som hadde som mål å dokumentere det norske litterære skriftspråket. Utgangspunktet var den norske delen av Brynhildsens *Norsk-engelsk ordbok*, men redaksjonen ble fort klar over at den ble for knapp, og de brukte også andre tilgjengelige ordbøker i tillegg til systematisk å ekserpere mange utvalgte norske forfatterskap som tilleggsmateriale. Etter hvert hadde redaksjonen en omfattende beleggsamling fra mønstergyldige tekster skrevet på norsk riksmål, blant annet ble Ibsens ordforråd godt dokumentert. Resultatet ble en normativ og vel dokumenterende ordbok over det norske litteraturspråket.

De fleste andre norske ordbøker er kommet i stand på den tradisjonelle måten, bygd på mer eller mindre tilfeldig innsamlet grunnlagsmateriale i tillegg til tidligere ordbøkers ordforråd. Det kan bli gode ordbøker av det, men svakheten ved en slik metode er at ordbøkene lett kan dokumentere språklige avvik eller nyheter i stedet for det vanlige og mest brukte ordforrådet, siden det er det spesielle eller avvikende som det er lettest å merke seg. (Ifølge den danske professor i leksikografi Henning Bergenholtz (Bergenholtz, 1996:6) kan resultatet i verste

fall bli en samling av perversiteter.) Grunnen til det er at de dagligdagse ordene og uttrykkene i det språket en er flytende i, lett overses, og dermed kan de komme til å mangle i ordbøkene. Særlig kan metoden være farlig om ordboksforfatteren har et personlig mål eller en sterk ideologi med arbeidet sitt. En av de mest kritiserte ordbøkene i så måte er Arakins russisk-norske ordbok (1963), som omtales som notorisk upålitelig (Nesset & Trosterud 205:273). Arakins mål var å lansere ei ordbok med størst mulig lemmautvalg, og redaktøren komponerte derfor selv mange udokumenterte, men tenkbare sammensetninger. Av ideologiske grunner kom nok også Knud Knudsen til å konstruere mange udokumenterte ord i norsk i den ordboka han selv regnet som sitt storverk, *Unorsk og norsk, eller fremmedordenes avløsning* (Knudsen 1881), der målet var å gi alternativer til fremmedord i den norske språket. Som kjent fikk svært få av hans forslag innpass i norsk språk.

Beleggsamlinger der en og samme ekserptor kan ha bidratt med flere hundre registreringer, vil uvegerlig også medføre en skjevhet i ordutvalg og bruken av ordene. Svært ofte er det ordboksredaktørenes eget ordutvalg som fører til skjevhet eller feil beskrivelser og merking. Det er også lett å finne eksempler på bruksmarkeringer vi er uenige om i mange ordbøker.

Alle ordbøker må gjøre et utvalg av ord som skal beskrives, et såkalt lemmautvalg. Det fins internasjonalt noen få uttømmende ordbøker som er laget manuelt over berømte tekstsamlinger eller viktige forfatterskap i såkalte tesauruser eller konkordanser. I tesauruser lemmatiserer man med stor nøyaktighet hvert eneste ord som forekommer i de utvalgte tekstene og utarbeider definisjon eller forklaring av bruken av dem, og ordene blir vanligvis ordnet etter begrep, altså etter semantisk betydning, ikke etter alfabetet. Konkordanser, derimot, er ordnet alfabetisk, og hvert ord har henvisning til hvor det er brukt i den teksten man har gransket. Både tesauruser og konkordanser var før dataalderen en meget tidkrevende og kostbar dokumentasjon av ordforråd, og ble bare gjort med svært viktige tekster, som regel historiske eller religiøse skrifter. For norsk fins bare *Norsk bibelkonkordans* fra 1907 som er laget på den måten, med en oversikt over 70 000 av de ordene som var brukt i den første norske Bibelen fra 1904.

[2.2] *Nye muligheter for leksikografien i dataalderen*

I siste halvdel av 1990-tallet forsøkte mange leksikografer å finne nye og bedre metoder for ordbokskriving, og med framveksten av elektronisk databehandling av tekst ble det plutselig mulig å behandle store mengder data med letthet. Da datamaskiner ble tatt i bruk i ordforskningen, ble det en faglig revolusjon for leksikografene, og det ga faget en helt ny status. Tidligst ute med bruk av edb (elektronisk databehandling) i norsk leksikografi var sannsynligvis professor

Harald Noreng, som i samarbeid med daværende NAVFs edb-senter for humanistisk forskning i Bergen i perioden 1975-1986 utarbeidet en fullstendig oversikt over ordforrådet i Ibsens verker (Noreng 1993). Det var et banebrytende arbeid i norsk leksikografi.

I dataalderen ble det mulig å samle tekster i søkbare databaser, kalt elektroniske tekstkorpus. Utbredelsen av moderne korpusleksikografi begynte med John Sinclairs forelesninger og studier utover 1980-tallet, og hans lærebok *Corpus Concordance Collocation* fra 1991 fikk stor betydning verden over. Internasjonalt begynte mange ordboksredaksjoner å lage egne, søkbare tekstsamlinger som grunnlag for sine ordbøker. Med moderne elektroniske tekstkorpus kunne man lage både konkordanser og tesauruser ut fra et nesten uendelig stort tekstmateriale, noe som åpnet opp for en helt ny æra i ordboksarbeidet. Sinclair hadde lenge hevdet at man måtte ha store tekstsamlinger for å kunne si noe viktig eller sikkert om språk i det hele tatt, og når det gjaldt leksikografi, formulerte han det slik: «It is, therefore, necessary to have access to a large corpus because the normal use of language is highly specific, and good representative examples are hard to find.» (1991:101). Han kritiserte ellers de tradisjonelle ordbøkene for bare å være egnet for resepsjon, ikke for produksjon. Produksjonsordbøker krever mer dyptpløyende semantisk beskrivelse av de leksikalske enhetene, og mer detaljert grammatisk informasjon om bruken av dem. Det var særlig konstruerte eksempler som Sinclair var motstander av, han mente at bare ekte brukseksempler kunne anvendes som dokumentasjon på ord i bruk (Krishnamuthy 2008:237); med hans egne ord: «usage cannot be thought up - it can only occur» (Sinclair 1987:3). Språklig evidens kan man bare finne enten ved introspeksjon eller observasjon, og på leksikalsk nivå er bare observasjon brukbart, hevdet han. Sinclair mente at leksikografi først og fremst var en deldisiplin av språkteknologien. Han demonstrerte hvordan man kunne lage ordbøker ved hjelp av språkteknologisk metode i Cobuild-prosjektet *Looking up!* (Sinclair 1987). Rapporten fra dette prosjektet havnet i posthylla som en gave til en stipendiat ved det daværende Leksikografisk institutt, som seinere ble prosjektleder for LBK. Den boka virket som en indirekte oppfordring: norsk leksikografi har stort behov for et skikkelig leksikografisk korpus, og rapporten hadde stor betydning for at arbeidet med LBK ble igangsatt.

Men det er ikke bare å sette i gang og samle inn tekster og tro man lager et nyttig leksikografisk korpus. Hovedproblemet med å bruke korpus som materiale for ordboksarbeid er at man må være svært bevisst på hvilke tekster man anvender, hvor store tekstmengder man trenger og hvordan tekstmassen bør være sammensatt for å få fram et så dekkende bilde som mulig av et språks ordforråd og hvordan det brukes i autentisk språkbruk. Utvelgelsesprosessene her

er avgjørende for kvaliteten på ordbøker som lages med korpus som materiale. Samtidig med at det ble utviklet metoder for å bygge opp store tekstkorpus, vokste innsikten i statistiske metoder brukt i språkforskningen. Særlig innen leksikografien har det blitt utviklet gode og brukervennlige programmer, som Sketch Engine (Kilgarriff et al. 2004) og den nordiske DeepDict analysis (Bick 2010). Ved hjelp av et sett algoritmer kan man nå analysere store tekstmengder automatisk og regne ut hvilke fraser og sammensetninger som er vanlige, og hvilke som er uvanlige eller må regnes som feil. Særlig nyttig er slike hjelpemidler ved kartlegging av et språks fraseologi.

[2.3] *Korpusbasert leksikografi i Norden*

Den danske ordbog (DDO) ble laget med et systematisk sammensatt tekstkorpus på 40 millioner ord som materiale. Ordboka kom ut i 1995 som den første korpusbaserte ordboka i Norden. Alle ord som forekom i korpuset, ble vurdert som kandidater for ordboka. Etter en grundig utvelgelsesprosess i redaksjonen ble de aktuelle ordene lemmatisert, definert og belagt med levende og autentiske eksempler fra korpuset. Å redigere en ordbok ut fra slikt materiale var da noe helt nytt i nordisk leksikografi, og det lå en annen tenkning bak selve redigeringsarbeidet enn slik man hadde gjort ved de tradisjonelle ordbøkene. Leksikografi ble med disse metodene ikke sett på som en gudbenådet samleroppgave, som man tidligere tenkte om fagfeltet. Det var særlig den første internasjonale læreboka i leksikografi som dokumenterte dette fagsynet: *The art and craft of lexicography* (Landau 1984).

De nye datamaskinelle metodene bød på en mer vitenskapelig og nøyaktig utforsket presentasjon av et språks ordforråd og hvordan det fungerte i reell bruk. For ordboksredigering er korpus selvsagt bare ment som et tryggere grunnlag og hjelpemiddel, både for lemmaseleksjon, definisjon og bruksmarkeringer. En redaksjonell vurdering av enhver opplysning i en ordbok vil naturligvis aldri bli overflødig. Men grunnlaget vurderingen tas på, er mye bedre enn før.

Metodene er beskrevet i seinere nordiske lærebøker i leksikografi, som Bo Svenséns *Handbok i lexikografi* (2005), Sven Göran Malmgrens *Svensk lexikologi – Ord, ordbildning, ordböcker och orddatabaser* (1994) og Fjeld og Vikørs *Ord og ordbøker* (2008). Språkdata ved Göteborgs universitet var tidlig ute med systematiske ordforrådsstudier, med Sture Allén og den store forskergruppen rundt ham, som utførte korpuslingvistiske undersøkelser helt tilbake på 1960-tallet. Det fins mange store og gode korpus over svensk språk, men det har ennå ikke blitt utarbeidet noe balansert og vektet korpus over svensk skriftspråk som grunnlag for ordbøker.

I Norge var det på slutten av 1990-tallet fortsatt stor mangel på systematisk innsamlet materiale for å lage ordbok, og da det ble lyst ut midler til forskningsinfrastruktur på HF-fakultetet ved Universitetet i Oslo, søkte vi og fikk i 1999 aksept for å bygge et bokmålskorpus på 40 millioner løpeord. Ruth Vatvedt Fjeld, Avdeling for leksikografi ved Institutt for lingvistiske og nordiske studier, sto som prosjektleder med Anders Nøklestad ved Tekstlaboratoriet på samme institutt som teknisk fagansvarlig.

Det ble samtidig søkt om midler til å bygge opp et korpus for nynorsk under ledelse av Lars S. Vikør. Nynorskprosjektet valgte å ta inn alle tilgjengelige tekster uten å relatere til vekting eller en viss spredning i materialet, sannsynligvis fordi det var mye mindre tekst å velge blant for nynorsk. Dermed utviklet disse to korpusene seg svært forskjellig, og det var lite reelt samarbeid underveis om design og oppbygning av dem. Det nynorske korpuset skulle først og fremst være grunnlag for redigering av det nasjonale prosjektet *Norsk Ordbok*, der nynorsk skriftspråk og norske dialekter skulle dokumenteres.

[3] OPPBYGGING AV DET LEKSIKOGRAFISKE BOKMÅLSKORPUSET

De eldste tekstene som ble samlet inn for LBK, var fra 1985, siden det var det første året det ble vanlig å publisere tekster elektronisk. Målet var å lage et åpent («open-ended») korpus, der man fylte på nye tekster i vektet mengde for hvert år, inntil man hadde et korpus som var stort nok for det man ville dokumentere. Målet på 40 millioner ord ble nådd i 2008. I løpet av perioden 1999-2008 var elektronisk publisering blitt mye mer vanlig og lettere tilgjengelig. Derfor søkte vi om flere midler for å bygge et korpus på 100 millioner ord, slik at størrelsen skulle bli sammenliknbar med andre internasjonale korpus i tiden, bl.a. British National Corpus.

Med et korpus på 100 millioner løpeord kunne vi dokumentere bedre både enkeltord og fraseologien i moderne norsk bokmål. Forskningen på fraseologi hadde skutt fart på totusentallet. Et språks inventar består ikke bare av enkeltord, mange leksikalske enheter består av flere ord som gjerne opptrer sammen i flerordsenheter eller fraser, kalt MWEs («multiword expressions»). Statistiske analyser av store tekstmengder kan gi bedre oversikt over hvilke flerordsenheter som eksisterer i et språk, eller hvilke som er gangbare i moderne språkføring, og hvilke som er foreldet. Flerordsenheter, både med og uten metaforisk betydning, endrer seg mye raskere i et språk enn enkeltord gjør, og en god ordbok bør kunne vise det. De statistiske analysene som trengs for kartlegging, krever imidlertid et svært omfattende materiale om resultatene skal bli gode nok. Etter en statistisk analyse av MWEs i LBK2008 (Bick 2010) fant vi ut at et materiale på 40 millioner løpeord var for lite for å sortere tilfeldige sammenstillinger av ord fra

flerordsenheter med egen leksikalsk betydning. Ved 100 millioner løpeord ble resultatene mer treffsikre og nyttige i kartlegging av faste fraser og flerordsenheter i forskjellige teksttyper. Fagspråk og allmennspråk har som regel svært ulik fraseologi, dermed er det også viktig å kunne sortere tekstene i subkorpus etter teksttype og sjanger for å kunne beskrive fagrelevant fraseologi bedre.

Vi fikk forlenget prosjektet med dette målet for øye, med støtte fra Universitetet i Oslos infrastrukturprogram. Tekstinnsamlingen ble avsluttet i 2013 med vel 100 millioner løpeord, og LBK inneholder dermed tekster fra en periode på 28 år.

At et korpus er balansert, vil si at korpuset består av en beregnet mengde forskjellige typer tekster etter et definert mål. Vi kom fram til at det skulle være 20 % periodika, 45 % sakprosa, 25 % skjønnlitteratur, 5 % tv-tekst, og 5 % upublisert, og dermed unormert, tekst. Balansen ble bestemt etter en avveining av statistiske data fra Norsk Mediebarometer i 2003 om folks lesevaner[1], samt hva det faktisk var mulig å få tak i av elektroniske tekster uten større kostnader enn prosjektet kunne tåle. Mediebarometeret viste at folks lesevaner det året var 35 % internettlesing, 15 % bøker, 40 % aviser, 5 % tidsskrifter og 5 % tegneserier. Det var lite spesifiserte tall, for eksempel visste man ikke hva slags tekster folk leste på internett, og vi jenket fordelingen noe etter hvordan man hadde designet andre korpus, særlig var British National Corpus et viktig forbilde, samt det korpuset som lå til grunn for utvelgelse av grunnbegrepene i EuroWordNet, i og med at LBK også var ment å være grunnlag for utvikling av et norsk ordnett (Vossen, P. et al 1998, Fjeld, R.V et al 2004). I EuroWordNet la man stor vekt på balanse med hensyn til teksttype og tekstforfatternes kjønn, alder og sosiale tilhørighet. Dette var også data som ble registrert for de fleste tekstprodusentene i LBK. Under innsamlingen prøvde vi bevisst å holde disse variablene så balansert som mulig, med de begrensninger materialtilgangen ga. Det ble en avveiing av hva slags materiale som var lett tilgjengelig, og hva vi spesielt trengte å finne for å holde balansen. I den prosessen gjorde de vitenskapelige assistentene en stor innsats for å kartlegge de sosiale variablene. Et unntak gjorde vi med avistekstene, som ofte var svært korte og var skrevet av personer som det var vanskelig å skaffe bakgrunnsdata på. Da ble bare avisnavn og publiseringsdato registrert.

[3.1] *Hvorfor lage et balansert leksikografisk korpus for norsk bokmål?*

Et godt balansert korpus over moderne norsk er et nyttig hjelpemiddel for å kartlegge hvilke ord som er brukt av så mange og i så forskjellige sammenhenger at de bør være oppslagsord i en allmennordbok. Et balansert korpus gir kunnskap om det vanligste ordinventaret i forskjellige typer tekst og er et viktig redskap

[1] https://www.ssb.no/a/publikasjoner/pdf

for å kunne foreta en godt fundert lemmaseleksjon. Et godt lemmautvalg inkluderer selvsagt alle rimelig frekvente ord, men også nye ord, såkalte neologismer. At ordbøkene skal ha med nye ord, er lett å forstå, bare man ikke tar med alt nytt som kanskje brukes et fåtall ganger av noen få; det må foreligge en viss konvensjonalisering før et ord lemmatiseres.

Videre er det viktig at de ordene som lemmatiseres, er brukt med en viss frekvens og spredning innen det aktuelle språksamfunnet. Dialektuttrykk tas vanligvis ikke med, eller de merkes som typiske for et område eller et spesielt kommunikasjonsfelt. En god ordbok har derfor et sett av såkalte diasystematiske markeringer som gir opplysninger om bruksforhold eller andre pragmatiske restriksjoner. Men å være treffsikker med disse markeringene forutsetter god innsikt i språkbruken generelt, det er ikke nok med egen intuisjon. Et balansert korpus skal nettopp kunne gi det nødvendige bakgrunnsmaterialet for å gjøre slike nødvendige avveiinger ved oppføring av bruksmarkeringer.

Ved hjelp av et korpus kjørt mot lemmalisten i en eksisterende ordbok kan man også få en viss oversikt over hvilke ord som ikke eller sjelden brukes lenger, slik at man ekskluderer eller merker dem som anakronismer, avhengig av ordbokstype. I skoleordbøker og innlæringsordbøker er det ofte hensiktsmessig å utelate foreldede ord, da denne typen brukere sjelden leser eller anvender all informasjon i en komplisert ordboksartikkel, og dermed kan komme i skade for å skrive anakronismer eller rett og slett utdødde ord i sine moderne tekster, noe som selvsagt ikke er til hjelp for dem. I større dokumentasjonsordbøker og historiske ordbøker bør ord som ikke lenger er vanlige i samtiden, være med, slik at brukerne kan lese og forstå dem. Tidligere ble slike ord gjerne slettet av plasshensyn. I elektroniske dokumentasjonsordbøker bør de bli stående med bruksmarkering som «sjelden» eller «eldre».

Levende språk er i stadig endring, og særlig endrer ordinventaret seg i takt med den generelle samfunnsendringen. Kartlegging av neologismer og anakronismer er derfor en forutsetning for godt ordboksarbeid (Fjeld og Nygaard 2012), og i det arbeidet er gode og moderne korpus av uvurderlig verdi.

[3.2] *Innsamling av tekstene i Leksikografisk bokmålskorpus*

Tekstene som ble samlet inn til korpuset, ble sortert og klassifisert i følgende teksttyper:

Aviser og ukeblader	AV00
Riksaviser	AV01
Regionaviser	AV02
Lokalaviser	AV03
Ukeblad	AV04
TV	TV00
Teksting av nyhets- og dokumentarprogram	TV02
Teksting av TV-serier og filmer	TV03
Skjønnlitteratur	SK00
Epikk	SK01
Drama	SK02
Lyrikk	SK03
Korttekster	SK04
Religiøse tekster	SK05
Sakprosa	SA00
Lærebøker	SA01
Fagbøker	SA02
Biografier	SA03
Artikkelsamlinger (antologi, festskrift o.l)	SA04
Avhandling	SA05
Juridisk dokument	SA06
Rapport	SA07
Oppslagsverk, leksikonartikler	SA09
Reiseskildring	SA10
Uspesifisert	SA99
Fagtidsskrifter	SA11
Unormert materiale, småtrykk, annet	UN00
Reklame, brosjyre, bruksanvisninger m.m.	UN01
Korrespondanse & forretningsdokumenter	UN02
Diskusjonsgrupper	UN04
Blogger	UN05

Korpuset har registrert tre sosiale variabler hos tekstforfatterne: kjønn, alder og oppvekststed. Idealet var å ha jamn fordeling på alle sosiale variabler, slik at

det blant annet ville være mulig å undersøke hvordan forskjellige brukere anvender den åpne bokmålsnormen, utføre statistiske analyser og få sammenliknbare resultater.

Resultater fra korpus bør alltid vurderes etter to vesentlige dimensjoner, frekvens og spredning. Svært mange undersøkelser i store tekstmengder, som hos Google eller liknende, gir gode anslag for frekvenser, men det er ofte vanskelig å finne ut noe etterprøvbart om spredning. I beste fall kan man finne spredning over tid, da publiseringsår ofte registreres, men spredning på fagfelt, sosiale variabler hos forfattere som kjønn og alder eller andre avgjørende variabler er det ikke mulig å sortere ut. Vil man ha sammenliknbare kvantitative resultater etter sosiale variabler fra LBK, kan man i søkegrensesnittet lett skille ut subkorpus med lik størrelse. Et balansert og godt merket korpus kan dessverre heller ikke bli helt perfekt eller treffsikkert, men det kan i hvert fall gi bedre grunnlag for videre undersøkelser enn et søk på for eksempel Google kan.

Det var mange hindringer som skulle overvinnes for å få tilgang til nok elektroniske tekster, særlig ga lov om åndsverk store problemer i de første årene prosjektet pågikk. De ble løst ved at vi avtalte med forlagene å slette ca. 20 % av sidene i hvert verk, tilfeldig spredt utover i teksten, slik at den ble umulig å rekonstruere for piratutgaver, som var det man fryktet mest ved oppstarten av prosjektet. Det lettet søknadsprosessen overfor tekstprodusentene i stor grad. Redselen for misbruk og rekonstruksjon avtok imidlertid etter hvert som digitalisering og nettpublisering ble mer vanlig utover 2000-tallet, og fra ca. 2010 var det få som ba om slik tekstreduksjon. I tillegg ble det undertegnet forpliktende avtaler med alle tekstgivere om at tekstene ikke skal brukes til kommersielle formål, kun til språkforskning, og at tekstene bare offentliggjøres gjennom et avtalt grensesnitt som er tilgjengelig for registrerte brukere via Tekstlaboratoriet ved Universitetet i Oslo.

I begynnelsen var det også vanskelig å få fatt i nok elektronisk tilgjengelige tekster uten vederlag. Men gjennom gode kontakter, først og fremst forlagsredaktør Øystein Eek i Kunnskapsforlaget, fikk vi avtaler med forlagene Gyldendal, Kunnskapsforlaget og Aschehoug. Forlagsredaktør Inger-Ma Gabrielsen har også bidratt jevnlig ved å gi UiO kopi av det meste som ble utgitt på Cappelen forlag. Seinere fikk vi også avtale med Verbum forlag, samt en lang rekke mindre tekstprodusenter.

Tekstene har ellers blitt samlet inn på svært forskjellige måter, bl.a. sendte prosjektlederen i 2001 ut et bønneskrift til alle universitetsansatte i Norge, med overskriften «Bli ordgiver du også». Det var svært mange som ga positiv respons på det og bidro med sin egen faglige produksjon, og slik ble mye av fagprosadelen i korpuset mettet, og det ga god spredning på fagfelt. Vi fikk i den forbindelse

også både publiserte og upubliserte vitenskapelige tekster. Sakprosatekstene er fordelt på humaniora, samfunnsfag, realfag, juss, helsefag samt sport og fritid. Dermed er det nå mulig å kartlegge faguttrykk som er gått inn i allmennspråket, gjerne med utvidet eller endret betydning, og som da bør defineres ut fra både faglig og allmennspråklig bruk.

Populærvitenskapelige tekster fant vi for det meste på Internett og i aviser. Også venner og bekjente av mange som arbeidet på korpuset, har bidratt med tekster. Vi har ellers fått tekster fra en lang rekke forfattere som har gitt oss sine råmanuskripter slik de var før de ble vasket av forlagene, slik at de kunne sammenliknes med de endelige, publiserte tekstene. Slike dubletttekster er interessante for leksikografisk forskning og for normeringsarbeid generelt, og materialet har blitt analysert i interessante studentoppgaver som blant annet ble publisert på prosjektets lanseringsseminar i 2008 (Ims 2008). Dette materialet kan lett sorteres ut som et subkorpus som det er viktig å hegne spesielt om, da råtekstene i prinsippet er upubliserte og uferdige.

I tillegg til forlagstekster og privat materiale har vi høstet inn offentlige tekster på Internett etter hvert som det ble en vanlig publiseringskilde, så mot slutten av arbeidsperioden ble det viktigere å velge de rette tekstene, mens det i starten var et hovedproblem å få tak i tekster i det hele tatt. Hele tiden måtte vi ha den rette balansen for øye, og sørge for at balansen skulle være der for hvert enkelt årstall.

Det mest krevende var å få inn tekstet talemål fra tv og radio, noe som ble løst ved å legge inn NRKs teksting for hørselshemmede over en viss periode. På den måten fikk vi i det minste registrert det nesten-muntlige ordforrådet, som er viktig i et leksikografisk underlagsmateriale. Tekstekontoret ved NRK fortjener en ekstra takk for fleksibilitet ved å gi oss tilgang til dette.

[3.3] *Resultatet*

Korpuset består av 27 082 dokumenter som til sammen utgjør 99 959 468 token, dvs. ordformer og tegnsettingstegn. Alle metadata er registrert manuelt. Målet med korpuset var som nevnt en balansert sammensetning av teksttyper. Videre ønsket vi å ha antall løpeord balansert for hvert enkelt år, samt god spredning på forfatternes sosiale variabler. Det har ikke vært lett å holde den balansen, og dersom man ønsker å gjøre direkte statistiske analyser ut fra visse variabler, kan det bli nødvendig å velge ut likevektige subkorpus. For eksempel var det vanskelig å finne like mange tekster skrevet av kvinner som av menn. Siden det totale korpuset er veldig nær 100 millioner token, er det enkelt å finne prosentandel for de enkelte variablene. For eksempel er det knapt 53 prosent tekst produsert av menn og tett oppunder 30 prosent av kvinner. De resterende tekstene har vi

ikke opplysninger om forfatterens kjønn på. Det gjelder først og fremst avistekster.

Om vi ellers ser på balansen i det ferdige korpuset, får vi også et noe annet bilde enn det som var målet. I forhold til de oppsatte målene har korpuset knapt 5 % mer sakprosa enn planlagt og 10 % mer skjønnlitteratur, og 15 % mindre fra periodika. Samtidig med arbeidet med LBK ble det utviklet et omfattende aviskorpus ved Universitetet i Bergen, Norsk Aviskorpus. Siden LBKs annoteringssystem var krevende for nettopp korte avistekster, valgte vi derfor å øke mengden skjønnlitteratur, og redusere avistekster. Man kan lett kontrollere ordforrådet i avistekster fra LBK mot Norsk Aviskorpus, og vi anser ikke denne balanseforskyvningen som betydningfull for verdien av LBK. Norsk aviskorpus er dessuten et særdeles godt hjelpemiddel i den redaksjonelle vurderingen som alltid foretas før et ord ordboksføres.

LBK er et ikke-kommersielt produkt som kan brukes av alle som logger seg inn og godkjenner lisensavtalen. Republisering er ikke tillatt, ellers kan korpuset gjerne brukes i språkforskning som gir salgbare produkter. Avtalen vi har med forlagene og andre tekstgivere, er at Universitetet i Oslo har rett til å bruke tekstene til forskningsformål mot at tekstgiverne selv også får tilgang til å bruke korpuset. Videre kan Universitetet i Oslo gi andre forskere utvidet tilgang etter spesiell avtale.

En viktig tanke bak LBK var nemlig at det ikke bare skulle fungere som materiale for ordboksredigering, men også kunne brukes i språkforskning generelt. For eksempel skulle korpuset kunne dokumentere bruken av moderne norsk morfologi, og spesielt vise hvilke brukere som foretrakk de forskjellige formene. Det er særlig interessant i språk som norsk, der det har foregått en politisk styrt normering over mange år, samtidig som normen er svært vid og gir mange valgmuligheter. Korpuset kan også gi svar på spørsmål om i hvilken grad eksisterende normering har virket, eller hvilke ikke-tillatte former som er så mye i bruk at de burde tas inn i normen. Dette er bl.a. demonstrert tydelig i masteroppgaven til Kjersti Wictorsen Kola (Kola 2014).

Korpuset har foreløpig ikke blitt utnyttet til å lage en korpusbasert ordbok, slik DDO ble laget over et eget korpus, men det har vært benyttet ved revisjon av både Bokmålsordboka og flere andre bokmålsordbøker. I Det Norske Akademis ordbok (NAOB, 2017) ble korpuset brukt flittig, både som leksikografisk verktøy for å undersøke morfologi, kollokasjoner og annet, og for å hente ut sitater. Pr. 2019 er det 30277 ordboksartikler i NAOB som har sitat fra LBK. Ellers er det skrevet flere fagartikler, masteroppgaver og PhD-avhandlinger med grunnlagsmateriale fra LBK.

[4] SØKEPROGRAMMET GLOSSA

Leksikografisk bokmålskorpus er gjort tilgjengelig i søkesystemet Glossa (Nøklestad et al. 2017), som er utviklet av Tekstlaboratoriet ved Universitetet i Oslo. Glossa er web-basert, noe som innebærer at man ikke trenger å installere programvare for å bruke det; alt man trenger er en nettleser og en datamaskin som er koblet til internett. Systemet støtter innlogging med Feide[2], noe som betyr at ansatte og studenter på norske universiteter og høyskoler (og på mange andre offentlige institusjoner) vil kunne logge inn med det brukernavnet og passordet de bruker på institusjonen sin. Det samme gjelder personer ved utenlandske akademiske institusjoner som støtter innlogging med eduGAIN[3], og personer med CLARIN-konto[4]. Om man ikke har tilgang til noen av disse innloggingsmetodene, kan man også få tilgang ved å kontakte Tekstlaboratoriet[5].

[4.1] *Søkegrensesnitt*

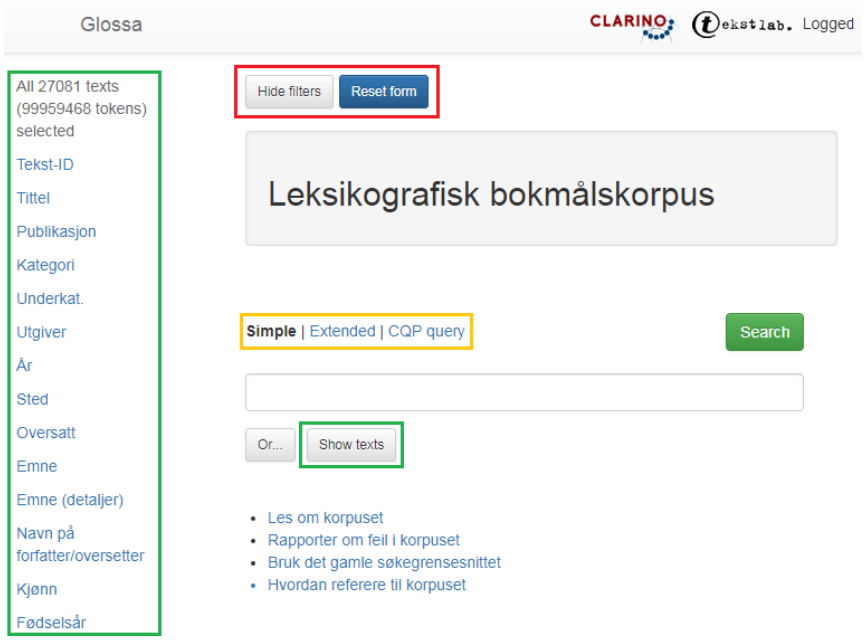
Glossa gjør det mulig å søke etter bestemte ord, eller deler av ord, fraser, lemma (grunnformer eller oppslagsformer), ordklasser og grammatiske trekk. Man kan søke i et utvalg av tekstene, for eksempel bare i aviser eller i tekster av en bestemt forfatter. Resultatene blir presentert som en konkordans, dvs. en liste med tekstutdrag som viser søkeordet med litt kontekst foran og etter. Konkordansen kan lastes ned i Excel-format eller som tab- eller kommaseparert tekst, noe som gjør det enkelt å jobbe videre med resultatene i annen programvare, som Excel eller statistikkprogrammer. Man kan også få resultatene i form av frekvenslister.

En ulempe med mange tidligere korpussøkegrensesnitt er at de krever teknisk kunnskap, både om såkalte regulære uttrykk og om kodene som representerer metadatakategorier og grammatisk informasjon i korpuset. Glossa er derimot utformet med tanke på at det skal være enkelt å bruke for personer uten slike spesielle tekniske kunnskaper.

Hovedsøkesiden til LBK er vist i Figur 1, der fargede rammer er satt inn for å utheve viktige deler av grensesnittet. Til venstre (grønn ramme) er alle metadatakategoriene det går an å søke i. Disse kategoriene representerer ulike typer informasjon om tekstene i korpuset: Tekst-ID, Tittel, Publikasjon, Kategori, Underkategori, Utgiver, År, Sted, Oversatt, Emne, Emne (detaljer), Navn på forfatter/oversetter, Kjønn og Fødselsår. Glossa sørger for at man ikke kan velge inkompatible verdier fra de ulike kategoriene; har man først valgt Dag Solstad under Navn på forfatter/oversetter, så vil Tittel-lista blir redusert til bare å inneholde boktitler av Solstad. Over metadatakategoriene kan man se hvor mange

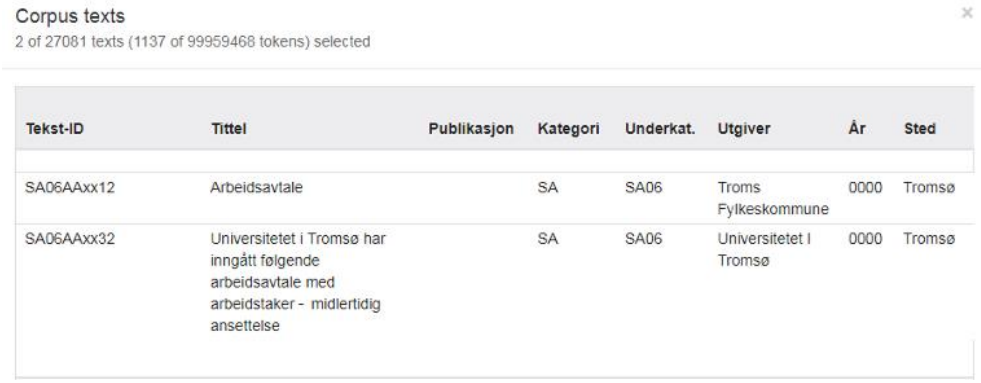
[2] https://www.feide.no/
[3] https://edugain.org/
[4] https://www.clarin.eu/
[5] tekstlab-post@iln.uio.no

tekster man har valgt, og hvor mange token (ordformer og tegnsettingstegn) de inneholder til sammen. Knappen *Show texts* (også grønn ramme) gir en oversikt over alle tekstene eller det utvalget tekster som har blitt valgt (se Figur 2). Øverst i Figur 1 finner vi to knapper (rød ramme). Med *Hide filters* kan man skjule metadatakategoriene til venstre, mens *Reset form* gir en blank søkeside.



FIGUR 1: Hovedsøkesiden til LBK.

Glossa tilbyr tre forskjellige søkegrensesnitt (gul ramme i Figur 1), der valgene fra venstre mot høyre gir økende muligheter for å formulere avanserte søk, men også en viss økning i vanskelighetsgrad. Grunnleggende søk etter ordformer eller fraser kan foretas ved hjelp av en enkel søkeboks som minner om det man finner i Google eller andre web-søkemotorer, og som de fleste brukere derfor vil være kjent med. Det er dette grensesnittet som er vist i Figur 1.


Corpus texts

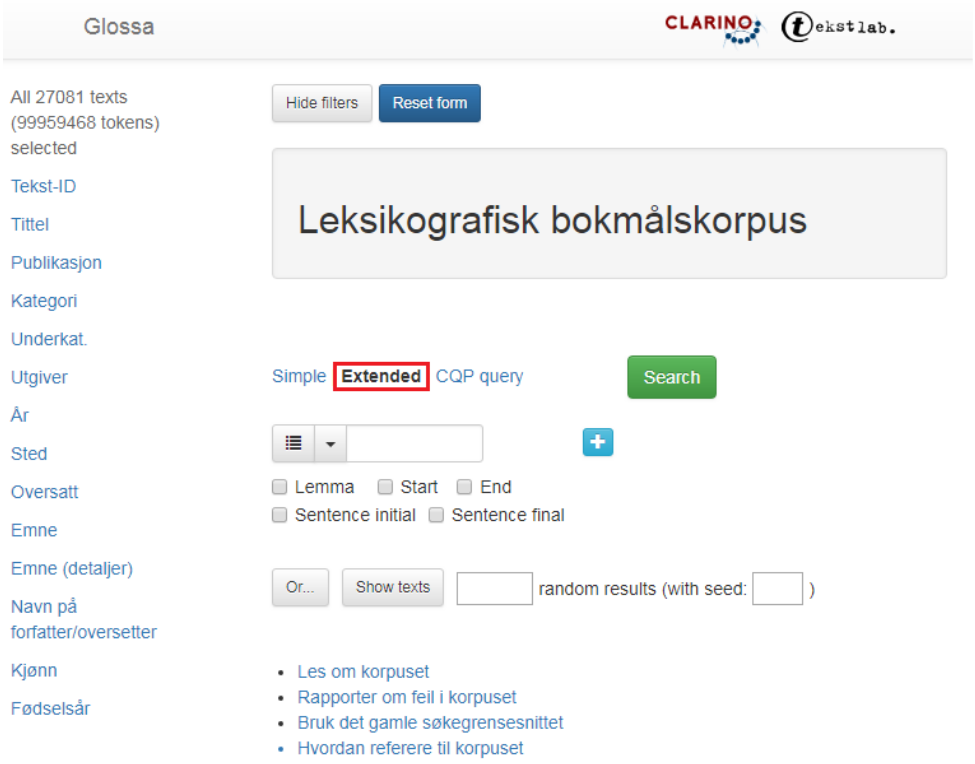
2 of 27081 texts (1137 of 99959468 tokens) selected

Tekst-ID	Tittel	Publikasjon	Kategori	Underkat.	Utgiver	År	Sted
SA06AAxx12	Arbeidsavtale		SA	SA06	Troms Fylkeskommune	0000	Tromsø
SA06AAxx32	Universitetet i Tromsø har inngått følgende arbeidsavtale med arbeidstaker - midlertidig ansettelse		SA	SA06	Universitetet I Tromsø	0000	Tromsø

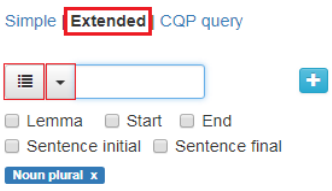
FIGUR 2: Oversikt over tekstene som er valgt.

Vil man søke etter lemma, deler av ord eller grammatiske trekk, kan man bytte til et grensesnitt som presenterer slike valgmuligheter i form av avkrysningsbokser, knapper og nedtrekksmenyer. Det gjør man ved å klikke på *Extended* over søkeboksen, som vist Figur 3. Man får da opp en søkeboks med en rekke avkrysningsknapper under, og disse gjør det mulig å spesifisere at søkeordet skal være lemma, begynnelsen eller slutten på et ord, eller i begynnelsen eller slutten av en setning. I dette grensesnittet er det også mulig å spesifisere at man vil ha et antall tilfeldig utvalgte resultater i stedet for alle resultater. Dette kan være nyttig hvis man vil ha ut et begrenset antall eksempler, men likevel vise variasjonen som fins i kildetekstene (siden man ellers bare ville få eksempler fra de første tekstene i korpuset).

Vil man oppgi grammatiske kriterier, kan man gjøre det ved å klikke på en av knappene til venstre for søkeordet, som vist i Figur 4. Knappen rett til venstre for ordet åpner en meny der man raskt kan velge ordklasse. Vil man i tillegg oppgi bøyingstrekk, kan man i stedet klikke på knappen lengst til venstre og få opp et vindu med ordklasser og tilhørende morfosyntaktiske trekk, som avbildet i Figur 5. Merk at det er fullt mulig å spesifisere både ordform eller lemma og grammatiske kriterier for ett og samme søkeord.

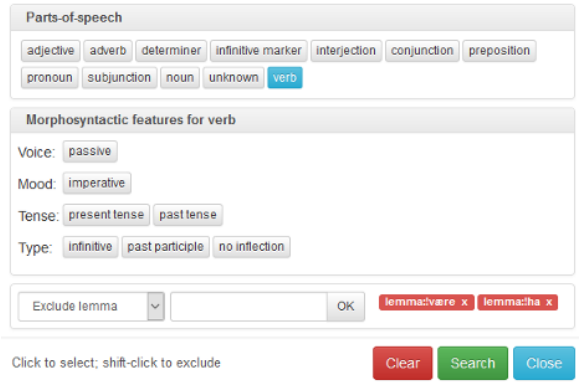


FIGUR 3: Utvidet søk.



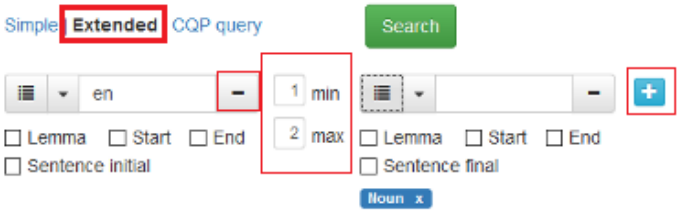
FIGUR 4: Knappene for ordklassesøk og søk etter andre morfologiske trekk.

Nederst i vinduet i Figur 5 kan man også utelukke bestemte ordformer eller lemma fra søket. Denne muligheten er nyttig hvis man for eksempel vil søke etter verb, men utelukke de verbformene som har «være» eller «ha» som lemma.



FIGUR 5: Vindu med ordklasser og tilhørende bøyningstrekk, samt mulighet for å spesifisere eller ekskludere lemma eller ordform.

Hvis man vil søke etter en frase, kan man legge til flere søkeord ved å klikke på den blå pluss-knappen lengst til høyre for søkeuttrykket. Vil man fjerne et av ordene igjen, kan man klikke på minus-knappen til høyre for det aktuelle søkeordet. Man kan også oppgi minimum og/eller maksimum antall uspesifiserte ord som skal kunne forekomme mellom to oppgitte søkeord; dette gjør det mulig for eksempel å søke etter ordet «en» etterfulgt av et substantiv, men med minst ett og maks to uspesifiserte ord (for eksempel adjektiver med eventuelle adverb foran) mellom. Disse funksjonene er uthevet i Figur 6.

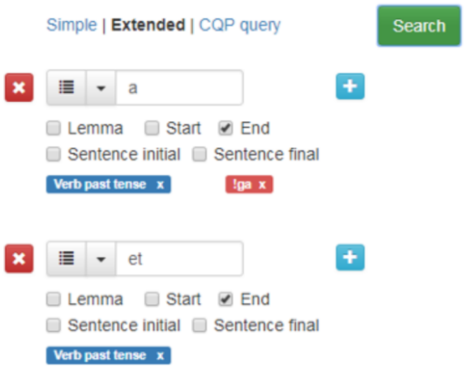


FIGUR 6: Søk på flere ord.

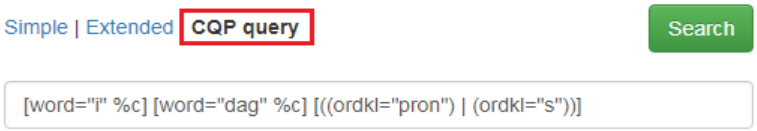
Det er også mulig å søke etter flere alternative varianter av et ord eller en frase. Hvis man klikker på knappen merket *Or...* under søkeboksen, får man opp en ny rad der man kan oppgi et alternativt søkeuttrykk. Figur 7 viser et eksempel på dette, der man har valgt å søke etter verb i preteritum som slutter på *-a* eller *-et*. Ordformen *ga* er ekskludert fra søket ved hjelp av mekanismen som ble vist i Figur 5.

Til slutt kan man velge å bruke et grensesnitt som gir tilgang til alle former for avanserte søk som er støttet av den underliggende søkemotoren (*The IMS Open Corpus Workbench*, Evert og Hardie 2011), men som til gjengjeld krever kunn-

skap om regulære uttrykk og grammatiske koder. Figur 8 viser hvordan søkeboksen ser ut med et regulært søkeuttrykk. I stedet for å utforme et slikt søkeuttrykk fra bunnen av kan man spesifisere søket så langt det er mulig under *Extended* og deretter bytte til *CQP query*-grensesnittet (som da vil vise uttrykket for søket man har spesifisert så langt) og justere det videre der.



FIGUR 7: Samtidig søk etter alternativer.

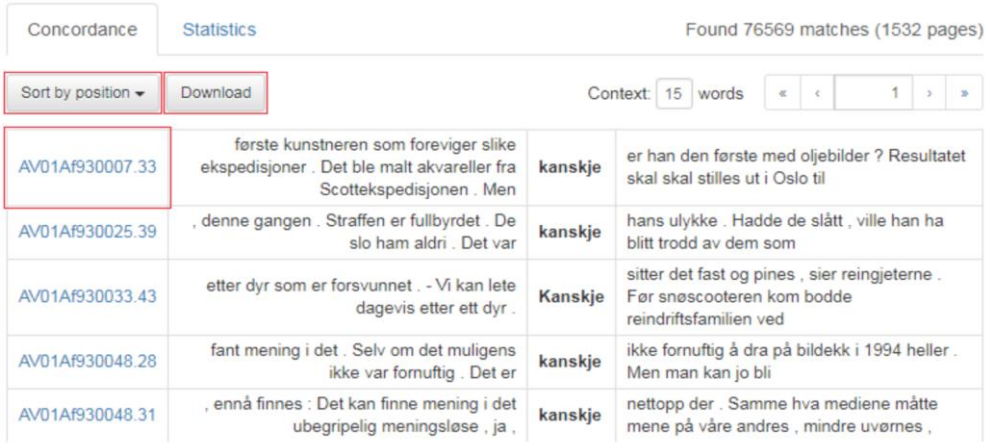


FIGUR 8: Avansert søk med regulære uttrykk. Søkeuttrykket inneholder ordformen *i* direkte etterfulgt av ordformen *dag*, igjen direkte etterfulgt av et ord med ordklasse «pron» (pronomen) eller «s» (substantiv). «%c» spesifiserer at forskjellen mellom store og små bokstaver skal ignoreres.

[4.2] *Søkeresultater*

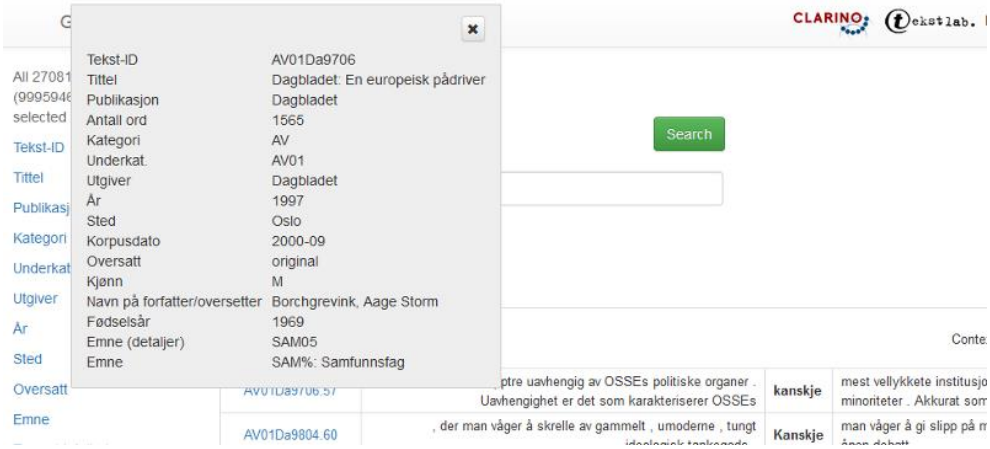
Søkeresultatene for LBK blir presentert i form av konkordanser og frekvenslister, med konkordanser som standardvisning (se Figur 9). Hvert søkeresultat blir vist som en rad i tabellen, med kolonner for tekst-ID, kontekst før søkeordet, selve søkeordet, og kontekst etter søkeordet. Kontekstlengden kan justeres ved å skrive inn et nytt tall i boksen over tabellen og trykke *Enter*. Ved hjelp av de uthevede knappene over tabellen kan resultatene sorteres eller lastes ned i Excel-format eller tab- eller kommaseparert format. Hvis man holder musa over søkeordet eller et ord i konteksten, får man opp informasjon om

lemma og grammatiske tagger, som vist i Figur 10. Ved å klikke på tekst-IDen i første kolonne får man listet opp metadataene som er registrert for teksten som søkeresultatet er hentet fra (Figur 11).



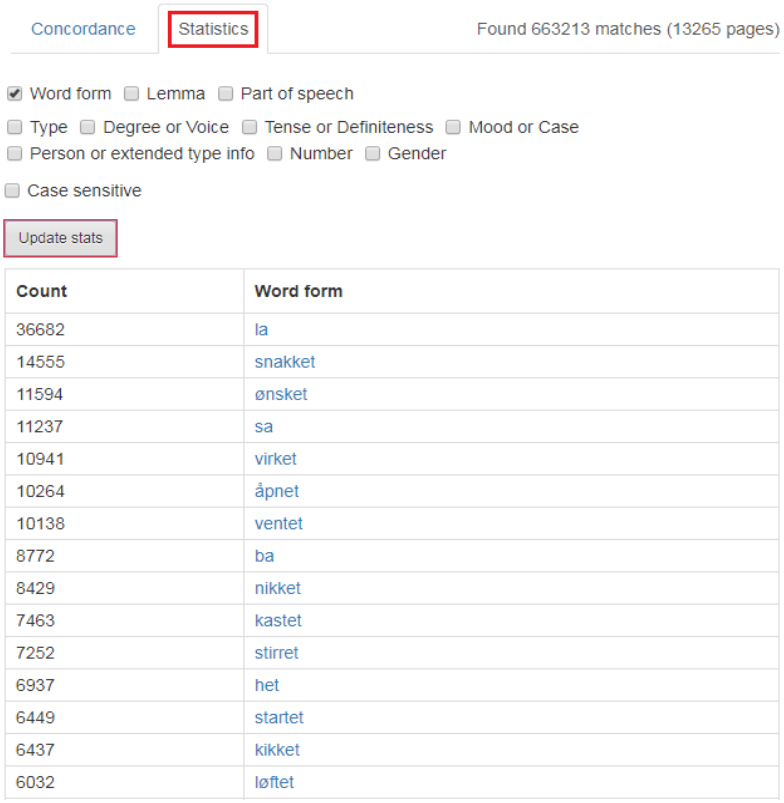
FIGUR 9: Konkordansvisning av søkeresultater.

FIGUR 10: Visning av lemma og grammatisk informasjon for et ord i søkeresultatet.



FIGUR 11: Visning av metadata for et søkeresultat.

Under den andre fanen med søkeresultater, *Statistics,* kan man få vist frekvenser over søkeresultatet; se Figur 12. Man kan velge å se frekvenser for ordformer, lemma og/eller grammatiske trekk. I figuren har vi valgt å vise ordformer. Hvis man klikker på en ordform, får man se en konkordans med bare de søkeresultatene som inneholder den bestemte ordformen.



Count	Word form
36682	la
14555	snakket
11594	ønsket
11237	sa
10941	virket
10264	åpnet
10138	ventet
8772	ba
8429	nikket
7463	kastet
7252	stirret
6937	het
6449	startet
6437	kikket
6032	løftet

FIGUR 12: Frekvenser for ordformene i et søkeresultat.

[5] OPPSUMMERING

LBK er et av flere norske tekstkorpus som er utviklet på 2000-tallet. Det som skiller LBK fra andre korpus, er at det er relativt godt balansert og har svært mange metadata. Dermed kan LBK utnyttes til flere formål enn å kartlegge leksikalske enheter. De enkelte ordboksoppslag (lemmaer) kan utstyres med rikere informasjon enn det som er vanlig i norske ordbøker, f.eks. morfologisk variasjon, leksikalske forskjeller hos forskjellige grupper språkbrukere, kjønnsspesifikt ordvalg, leksiko-syntaktiske endringer hos yngre brukere, for å nevne noe. Korpuset

har et vell av metadata, både relatert til språkbrukerne og til de forskjellige teksttypene og –sjangrene. Ved grundig analyse av LBK kan man dessuten framskaffe sikre data i den spesielle norske språkpolitiske og språkideologiske debatten. Søkesystemet Glossa gjør at det er enkelt å anvende for statistisk analyse og er gjort greit tilgjengelig for språkforskere.

Korpuset har alt i alt kostet rundt 10 årsverk, alt finansiert av Universitetet i Oslo, og vil være en viktig ressurs for moderne språkforskning og leksikografisk dokumentasjon av moderne norsk bokmål for perioden 1985-2013.

TAKK

Det har vært mange medarbeidere på prosjektet for tekstsamling, tekstrensing og merking. Lars Nygaard, Anne Engø, Preben Wik, Rune Lain Knudsen og Arash Saidi har etter tur vært tilsatt på timebasis som dataingeniører, og har bidratt mye til den teknologiske utviklingen av korpuset, hele tiden under ledelse av senioringeniør Anders Nøklestad. Det har vært mange vitenskapelige assistenter som har arbeidet med innsamling av tekster, tekstrensing og innlegging i korpuset. Ålov Runde var tilsatt som vit.ass. i oppbyggingsfasen og gjorde et stort arbeid med praktisk tilrettelegging av selve databasen. Videre i perioden har mange assistenter bidratt, særlig bør nevnes vit.ass.-ene Carina Nilstun, Julie Torjusen og Kjersti Wictorsen Kola, som alle har vært studenter på bachelorkurset i leksikografi og har bidratt vesentlig i arbeidet med LBK, både med gode ideer, systematisering og organisering av korpusinnholdet, og som har fungert som hjørnesteiner i et særdeles godt og kreativt arbeidsmiljø over mange år.

REFERANSER

Arakin, Vladimir D. (red.). 1963. *Norsk-russisk ordbok*. Moskva: Statsforlaget for ordbøker på fremmede språk og nasjonalitetsspråk.

Bergenholtz, Henning. 1996. Korpusbaseret leksikografi. I *LexicoNordica* 3, 5-18.

Bick, Eckhard. 2010. DeepDict - et korpusbaseret relationelt leksikon. I Ruth Vatvedt Fjeld & Henrik Lorentzen (red.), *Lexico Nordica 17 - 2010, Leksikografi og språkteknologi i Norden*, 17-34 . LexicoNordica. ISSN 0805-2735.

British National Corpus. http://www.natcorp.ox.ac.uk/corpus/creating.xml

Den danske ordbog. https://ordnet.dk/ddo

Den Danske ordbogs elektroniske korpus. https://ordnet.dk/ddo/fakta-om-ddo/metode-og-kilder/en-korpusbaseret-ordbog

Det Norske Akademis ordbok (NAOB). https://www.naob.no/

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. I *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Fjeld, Ruth E Vatvedt, Wik, Preben & Lars Nygaard. 2004. Managing Complex and Multilingual Lexical Data with the Simple Editor. I *Euralex international congress;* 2004-07-06.

Fjeld, Ruth E. Vatvedt & Lars S. Vikør. 2008. *Ord og ordbøker.* Fagbokforlaget.

Fjeld, Ruth E. Vatvedt & Lars Nygaard. 2012. Lexical neography in modern Norwegian. I *Exploring newspaper language : using the web to create and investigate a large corpus of modern Norwegian*. John Benjamins Publishing Company, 221-240. ISBN 978-90-272-0354-0.

Ims, Ingunn. 2008. *Kan dublett-tekstene fortelle noe om et forlags normeringsfilter?* Presentasjon på lanseringsseminar, Institutt for lingvistiske og nordiske studier ved Universitetet i Oslo.

Kilgarriff, Adam. et al. 2004. The Sketch Engine. I G. Williams og S. Veissier (red.), *Proceedings of the Eleventh EURALEX International congress, Euralex 2004*, Universite de Bretagne-Sud, 105-116.

Knudsen, Knud. 1881. *Unorsk og norsk eller fremmedordenes avløsning*. Cammermeyer forlag.

Knudsen, Rune Lain & Ruth Vatvedt Fjeld. 2013. LBK2013: A balanced; annotated national corpus for Norwegian Bokmål. I *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013; May 22-24; Oslo; Norway.* NEALT Proceedings Series 19.

Knudsen, Trygve & Alf Sommerfelt (red.). 1937-1957. *Norsk Riksmålsordbok.* Utgitt av Riksmålsvernet.

Kola, Kjersti Wictorsen. 2014. *Bokmålsbruk – hvorledes/hvordan/åssen og hvorfor?: Om bruken av morfologiske og ortografiske varianter i bokmålsnormalen.* Masteroppgave, Universitetet i Oslo.

Krishnamurthy, Ramesh. 1997. Corpus-driven lexicography. I *International Journal of Lexicography* vol. 21, nr. 3, 231-242.

Landau, Sidney I. 1984. *The Art and Craft of Lexicography.* Schribner.

Malmgren, Sven Göran. 1994. *Svensk lexikologi – Ord, ordbildning, ordböcker och orddatabaser.* Studentlitteratur AB.

Nesset, Tore & Trond Trosterud. 2005. Ny norsk-russisk ordbok: Ei leksikografisk storhending. I *LexicoNordica 12*, 273-284.

Nordisk leksikografisk ordbok. 1997. Universitetsforlaget.

Noreng, Harald. 1993. *Konkordans over Henrik Ibsens dramaer og dikt*. Ibsensamlingen. Oslo.

Norsk aviskorpus. http://uni.no/nb/uni-computing/clu/norsk-aviskorpus/

Nøklestad, Anders, Kristin Hagen, Janne Bondi Johannessen, Michal Kosek, & Joel Priestley. 2017. A modernised version of the Glossa corpus search system. I Jörg Tiedemann (red.), *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, 251-254.

Sinclair, John. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. Collins ELT.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Svensén, Bo. 2005. *Handbok i lexikografi. Principer och metoder i ordboksarbetet.* Esselte 1987 og 2005.

Vossen, Piek, Laura Bloksma, Paul Boersma, Felisa Verdejo, Julio Gonzalo, Horacio Rodriquez, German Rigau, Nicoletta Calzolari, Carol Peters, Eugenio Picchi, Simonetta Montemagni & Wim Peters. 1998. *EuroWordNet Tools and Resources Report.* EuroWordNet (LE-4003) Deliverable D021D025, University of Amsterdam.

KONTAKT

Ruth Vatvedt Fjeld
Institutt for lingvistiske og nordiske studier, Universitetet i Oslo
r.e.v.fjeld@iln.uio.no

Anders Nøklestad
Tekstlaboratoriet, Institutt for lingvistiske og nordiske studier, Universitetet i Oslo
anders.noklestad@iln.uio.no

Kristin Hagen
Tekstlaboratoriet, Institutt for lingvistiske og nordiske studier, Universitetet i Oslo
kristin.hagen@iln.uio.no