OSLa Hagen, Hjelde, Stjernholm & Vangsnes (red.) *Bauta: Janne Bondi Johannessen in memoriam*, Oslo Studies in Language 11(2), 2020. 449–464. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

GLOSSA SOM FORSKNINGSVERKTØY HVA FOLK SØKER ETTER OG HVA RESULTATENE BRUKES TIL

ÅSHILD SØFTELAND,[1] ANDERS NØKLESTAD,[2]
JOEL PRIESTLEY[2] & KRISTIN HAGEN[2]

[1]*Høgskolen i Østfold,* [2]*Universitetet i Oslo*

SAMMENDRAG

Tekstlaboratoriets korpus og søkegrensesnittet Glossa er utvikla til bruk for språkforskere og andre interesserte fra hele Norden og resten av verden. I denne artikkelen viser vi hvordan søkesystemet Glossa er utvikla fram mot i dag, hva korpusbrukerne søker på, og hvordan korpusdataene kan brukes i vitenskapelige publikasjoner.

[1] INNLEDNING

Glossa er et søkeverktøy for tekst- og talespråkskorpus som er utvikla ved Tekstlaboratoriet gjennom mange år. Hensikten med Glossa er at både enkle og kompliserte søk skal kunne gjøres intuitivt og lett, uten at brukeren trenger intensiv kursing eller kunnskap om formelle søkespråk. Seksjon 2 i denne artikkelen vil gi en kort presentasjon av hvordan søkeverktøyet er utvikla. Seksjon 3 viser kort hvordan Glossa ser ut i dag. I seksjon 4 tar vi for oss søkestatistikken for Glossa. Hvilke søk er det forskerne gjør? Er det enkle ordsøk eller mer kompliserte søk der metadata og for eksempel ordklasse er involvert? I femte seksjon ser vi på hva resultatene fra søka brukes til i vitenskapelige artikler. I 4 og 5 har vi begrensa oss til talespråkskorpusene *Nordisk dialektkorpus* (NDC) og *Amerikanordisk talespråkskorpus* (CANS). I seksjon 6 oppsummerer vi.

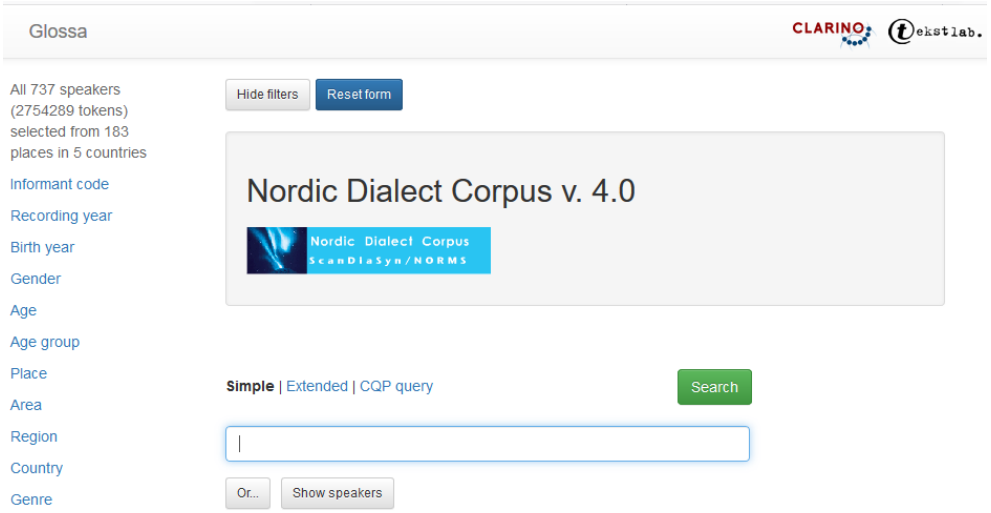
[2] UTVIKLING AV GLOSSA

Språkforskere trenger data som viser hvordan tekst og tale er i faktisk bruk. Da Tekstlaboratoriet ble oppretta i 1992, ble en viktig oppgave å samle tekster på norsk og andre språk og finne verktøy som kunne behandle slike data. Janne ble leder av Tekstlaboratoriet, og så tidlig behovet for et enkelt søkeverktøy som språkforskerne kunne ha nytte av uten å måtte bruke mye tid på opplæring eller

nedlasting av søkegrensesnitt og fonter. Janne ble dermed pådriver og hjernen bak alle søkegrensesnittene som ble utvikla ved Tekstlaboratoriet.

Det første websøkegrensesnittet ble laga for *The Oslo Corpus of Bosnian Texts* i 1997-1998[1]. Grensesnittet er enkelt og bygger på tekstdatabasen IMS Corpus Workbench, den samme databasen som Glossa bruker i dag. I dette korpuset kan man søke på enkeltord i en søkeboks, men for mer avanserte søk må man bruke et formelt søkespråk. *Oslo-korpuset av taggede norske tekster*, lansert i 1999, var det første søkegrensesnittet der alle søk kan gjøres ved å bruke menyer eller klikke i bokser[2]. Både Oslo-korpuset og Bosnisk-korpuset kan fortsatt brukes i dag.

Neste steg ble å utvikle søkegrensesnitt for talespråkskorpus, også via web, med transkripsjoner knytta til lyd- og videoopptak slik at søkeresultatene kunne høres og ses ved siden av visninga av transkripsjonene. Det første talespråkskorpuset, *NoTa-Oslo*, ble lansert i 2005, fulgt av *BigBrother* og *TAUS.*[3]



FIGUR 1: Hovedsøkesiden for nye Glossa med Nordisk dialektkorpus. Metadatamenyen er til venstre og en enkel søkeboks i midten. Her kan det søkes på ett eller flere ord. Et klikk på Show speakers viser hvilke informanter som er valgt i metadatamenyen.

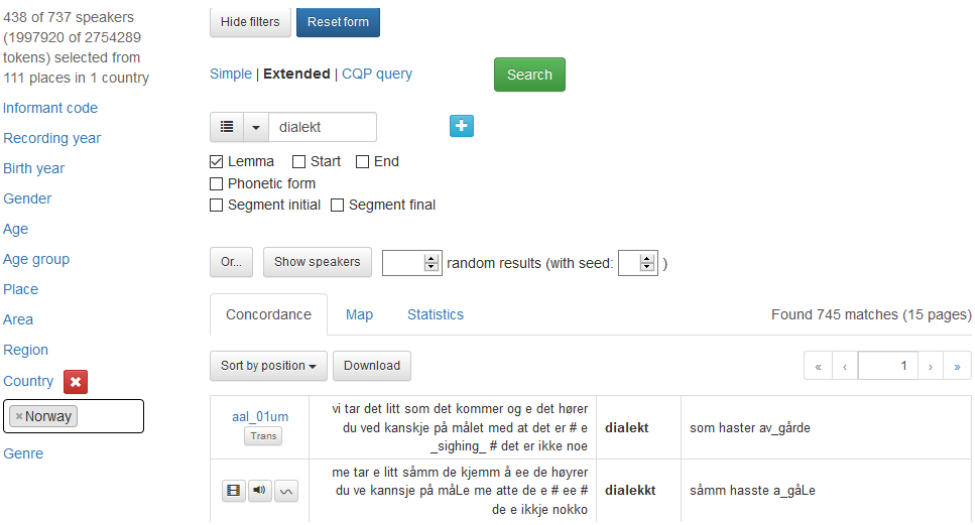
Parallelt med utviklinga av talespråkskorpusene utvikla Tekstlaboratoriet et nytt grensesnitt som skulle kunne brukes til både tekstkorpus, talespråkskorpus og flerspråklige (parallell-)korpus. Resultatet ble første versjon av Glossa,[4] der

[1] Hovedprogrammerer for *The Oslo Corpus of Bosnian Texts* var Diana Santos.
]2] *Oslo-korpuset av taggede norske tekster* ble videreutvikla av Sigurd Schiøth og Anders Nøklestad.
]3] Joel Priestley var hovedprogrammer for de første talespråkskorpusene.
[4] Den første versjonen av *Glossa* ble utvikla av Lars Nygaard og Joel Priestley.

alle talespråkskorpusene etter hvert ble lagt inn i tillegg til korpus som *Oslo Multilingual Corpus* og *Leksikografisk bokmålskorpus*. Med Glossa fikk Tekstlaboratoriet et fleksibelt verktøy der det var enklere å legge inn nye korpus. Det var brukervennlig, fungerte stort sett på samme måte for alle korpusene, og ga dem også likt utseende.

[3] NYE GLOSSA

I 2012 fikk Tekstlaboratoriet mulighet til å utvikle en ny versjon av Glossa gjennom *CLARINO*-prosjektet. Denne versjonen skulle bygge på erfaringene Tekstlaboratoriet hadde gjort seg med den første versjonen av Glossa, inkludert respons fra brukerne og programmererne. I nye Glossa[5] er fokuset på brukervennlighet videreført. I denne versjonen er inngangen for brukere uten tekniske forkunnskaper enda enklere (se Figur 1 og 2), samtidig som den tilbyr muligheter for svært avanserte søk for brukere som vil tilegne seg den nødvendige kunnskapen om det underliggende søkespråket (se Figur 3). Resultatene blir vist som ryddige konkordanser (jf. Figur 2).

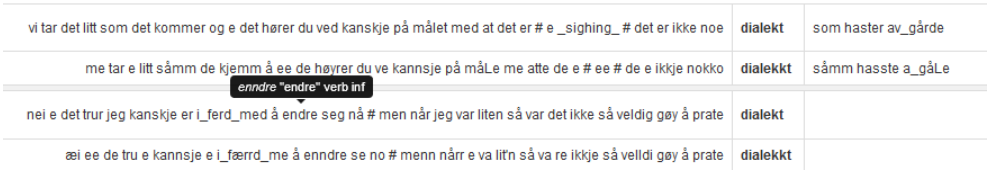


FIGUR 2: Under den utvidede søkeboksen kan man spesifisere søket. I dette eksempelet er det søkt på alle former av ordet *dialekt*. Et klikk på plusstegnet til høyre gir en ny søkeboks (og et nytt plusstegn). Øverst til venstre over metadatamenyen kan man se hvor mange informanter og tokens som er valgt. Søkeresultatet vises som en ryddig konkordans. Et klikk på film- eller lydsymbolet til venstre for søkeresultatet gir video-/lydvisning. Lydbølgen gir et spektrogram. Resultatene kan også vises som kart (under *Map*) eller lister over treff-ordene med frekvenstall (under *Statistics*).

[5] Hovedutvikler av nye Glossa har vært Anders Nøklestad med Joel Priestley på talesspråkskorpus-delen.

FIGUR 3: I CQP query-boksen kan man skrive avanserte regulære søkeuttrykk. Her er søket fra Figur 2.

De norske talespråkskorpusene har blitt automatisk annotert med grammatisk informasjon ved hjelp av en statistisk ordklassetagger som også er utvikla ved Tekstlaboratoriet. Annotasjonen følger *Norsk referansegrammatikk* (Faarlund, Lie & Vannebo 1997). I nye Glossa kan man se denne informasjonen ved å holde musa over et ord i søkeresultatet, slik Figur 4 viser. Ved å klikke på menysymbolet ved siden av søkeordet får man opp en boks der det er mulig å søke på grammatisk informasjon som ordklasse, kjønn, tall, tempus osv., som vist i Figur 5.



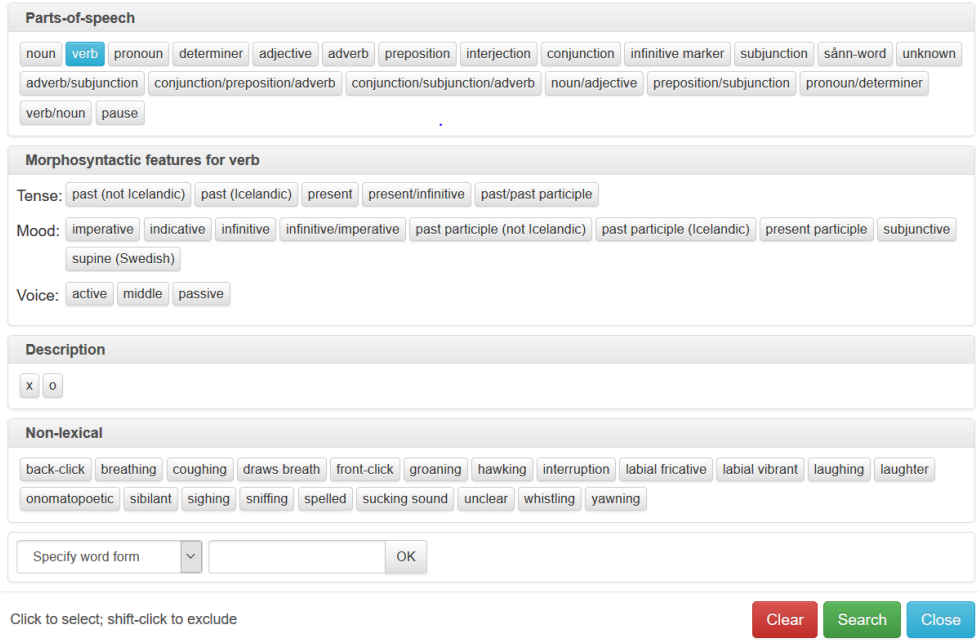
FIGUR 4: Grammatisk informasjon for ordet *endre* (*enndre* i den talemålsnære transkripsjonen, lemmaet er også '*endre*') fra søkeresultatet i Figur 2.

Noen av de norske talespråkskorpusene har to ulike transkripsjoner, en talemålsnær og en ortografisk. (Se Figur 2 og Figur 4 for eksempler på dette.) I nye Glossa blir det i utgangspunktet søkt i den ortografiske informasjonen, men dersom man krysser av for *fonetisk* ('Phonetic form') under søkeboksen, søkes det direkte i den talemålsnære transkripsjonen.

Nye Glossa kan brukes med Feide-, eduGAIN- eller CLARIN-innlogging, og har et system for håndtering av ulike brukerlisenser for korpus. Via nye Glossa vil det også være mulig å søke i korpus som ligger på andre servere enn der Glossa selv er installert. Dette kan gjøres gjennom infrastrukturen *federated content search* i CLARIN.

Som den gamle versjonen er den nye tilgjengelig som åpen kildekode og kan lastes ned på en åpen lisens om man vil bygge egne korpus. Den nye versjonen av Glossa brukes i dag til så å si alle Tekstlaboratoriets tekst- og talespråkskorpus. Den nye versjonen videreutvikles kontinuerlig, og gjennom infrastrukturprosjektet CLARINO+ er det planlagt mange forbedringer, spesielt

med hensyn til resultathåndtering. Les mer om nye Glossa i Nøklestad et al. (2017) og på nettsidene til CLARINO (se referanselista).



FIGUR 5: Søkeboksen for grammatisk informasjon samt ikke-språklig annotasjon som latter. Ved å klikke på en ordklasse (her verb) får man mulighet til å søke på mer informasjon om ordet.

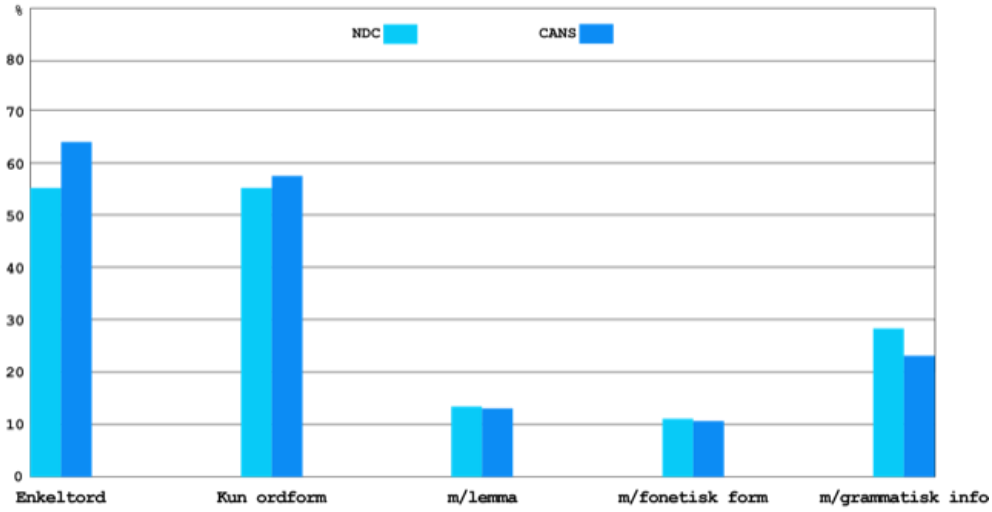
[4] HVA SØKER FOLK ETTER?

De siste 15 åra har Tekstlaboratoriet først og fremst fokusert på utvikling av talespråk, noe som har resultert i en lang rekke talespråkskorpus, bl.a. NoTa-Oslo, BigBrother og TAUS som nevnt ovenfor, men også Nordisk dialektkorpus (NDC), Amerikanordisk talespråkskorpus (Corpus of American Nordic Speech, CANS), LIA-korpusene (norsk og samisk) og korpus for diverse etiopiske språk (se Tekstlaboratoriets nettsider for mer informasjon om alle korpusene). Vi har sett nærmere på hva slags søk som blir gjort i to av disse, NDC og CANS, som begge er transkribert både talemålsnært og ortografisk.

Analysene av søk er basert på informasjon som blir lagra i en søkedatabase. Informasjon om enkeltsøk blir ikke knytta til brukeren som utfører søket, og er derfor fullstendig anonymisert. For frekvensanalysen av selve søkeuttrykkene bruker vi data som er innhenta fra 2017 til februar 2020. Informasjon om metadatautvalg har bare blitt registrert over en kortere periode, og derfor er

statistikken over metadata-søk begrensa til en periode på to måneder.[6] Glossa lagrer ikke informasjon om hvilken side i søkegrensesnittet som er brukt for hvert enkelt søk, *simple, extended* eller *CQP query* (jf. Fig.2).

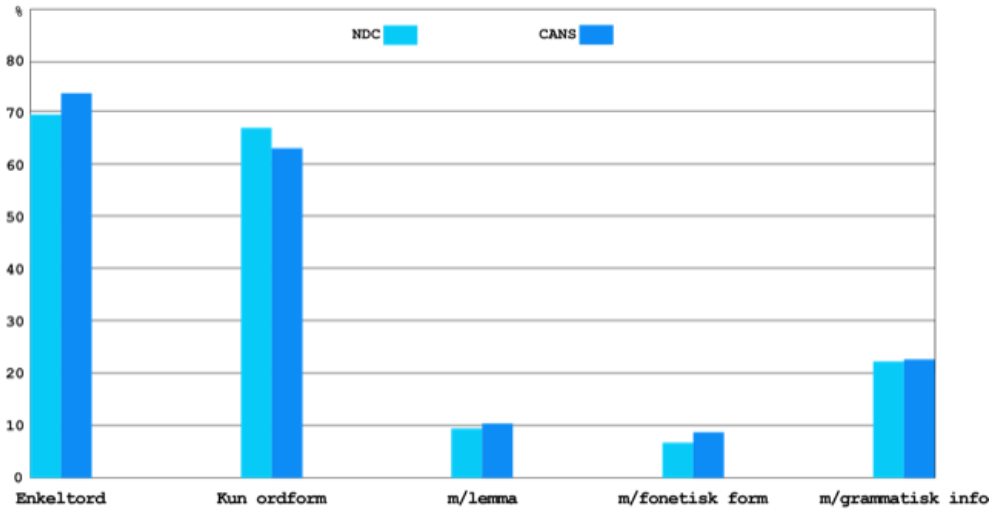
Resultatene fra frekvensanalysene er presentert i Figur 6 og 7.



FIGUR 6: Fordelinga av unike søkeuttrykk. I NDC var det registrert 16 548 unike søk, i CANS 4 448.

Figur 6 viser fordelinga av *unike søkeuttrykk* (forskjellige søk), mens Figur 7 viser fordelinga av søk totalt. Med andre ord: Hvis for eksempel ordformen *jeg* har blitt søkt etter 100 ganger, blir det bare telt én gang i Figur 6, men 100 ganger i Figur 7. Søylen «Enkeltord» viser andelen søk som bare inneholder ett ord, altså ikke flerordsuttrykk eller fraser. «Kun ordform» er søk (på enten enkeltord eller fraser) der bare den ortografiske formen til ordet er oppgitt, og ikke annen språklig informasjon. De tre siste søylene (som også inkluderer både enkeltord og fraser) viser derimot andelen søk der brukeren har spesifisert henholdsvis *lemma* (grunnform av ord), *fonetisk form* (med søk i lydnær transkripsjon) eller *grammatiske attributter* (f.eks. ordklasse, bøyingsform) som en del av søket. (Merk at et søk ikke nødvendigvis må inkludere et bestemt ord). Siden ett og samme søk kan spesifisere flere attributter, for eksempel både lemma og grammatisk informasjon, blir ikke summen av søylene 100 %.

[6] Nærmere bestemt fra og med 05.12.2019 til og med 04.02.2020.



FIGUR 7: Fordelinga av antall søkeuttrykk totalt. I NDC var det registrert 40 048 søk, i CANS 10 769.

Som disse to figurene viser, ser vi de samme tendensene i begge korpusene, uavhengig av om man ser på bare unike eller totalt antall søk:

- Rundt to tredjedeler av søka er søk etter bare ett søkeord, altså ikke fraser eller flerordsuttrykk.
- Rundt to tredjedeler av søka (både på enkeltord og fraser) angir bare ordformer (altså ikke lemma, fonetisk form eller grammatikk el.a.).
- Omtrent ti prosent inneholder søk etter et lemma.
- Omtrent ti prosent inneholder én eller flere spesifiseringer av fonetisk form.
- Rundt en fjerdedel inkluderer søk på grammatisk informasjon.

Det er verdt å merke seg det andre punktet ovenfor, som sier at to tredjedeler av søka ikke involverer verken lemma, fonetisk form eller grammatikk. Denne typen søk kan man utføre bare ved hjelp av *simple*-versjonen av grensesnittet, noe som illustrerer hvor stor verdi det har at Glossa tilbyr denne enkle søkemuligheten.

Analysene av bruk av metadatautvalg viser at 74 % av søka i NDC og 79 % av søka i CANS inkluderer metadata. Disse talla må imidlertid tas med en klype salt,

siden Glossa automatisk gjør et nytt søk hver gang man endrer metadatautvalget. Så hvis man for eksempel velger tre metadataverdier (innenfor enten samme eller ulike metadatakategorier), vil det bli registrert som tre søk med metadata. I slike tilfeller er det vanskelig å si om brukeren virkelig har vært interessert i hver enkelt metadataverdi eller bare har ment å bygge opp et sammensatt metadatautvalg.

Av samme grunn oppgir vi ikke konkrete tall når det gjelder frekvensen av ulike metadatakategorier, men vi kan slå fast at de klart mest brukte kategoriene i NDC er informasjon om informantens hjemsted (*Country*, *Region*, *Area* og *Place*), aldersgruppe, kjønn og informantkode, mens opptaksår, sjanger og konkret alder eller fødselsår forekommer mye sjeldnere. For CANS er de mest brukte kategoriene informantens arvespråk (norsk eller svensk), informantkode og opptaksår. Det virker rimelig at informasjon om hjemsted er mindre brukt i et arvespråkskorpus enn i et dialektkorpus, og siden de fleste talere av amerikanorsk eller -svensk er eldre mennesker, er det heller ikke overraskende at alder er sjeldnere valgt i CANS enn i NDC.

[5] BRUK AV KORPUSSØK I FORSKNINGSARTIKLER

I denne seksjonen skal vi se på hvordan ulike typer korpussøk blir brukt i vitenskapelige forskningsartikler, både søk etter enkelt- og flerordsuttrykk og med og uten presisering av fonetiske eller grammatiske detaljer i søket. Informasjon om metadatautvalg blir også nevnt. Forskningsarbeidene som blir omtalt, er publiserte tidsskriftartikler som viser til bruk av NDC eller CANS der også den konkrete søkeprosessen kommer fram. Utvalget av artikler er ment å vise bredde både i typen søk og i språklige varieteter og variabler som er undersøkt. I tillegg til mange tidsskriftartikler som ikke omtaler hvilke korpussøk som er gjort, er språkdata fra NDC og CANS også brukt i mange bokkapitler i antologier, i monografier og i mange master- og doktoravhandlinger. Noen forskningsarbeid beskriver også alternative måter å bruke korpusdata på som ikke kommer med under avgrensningene her. Ett eksempel er Nygård (2016) som analyserer subjektsellipser, altså ord som ikke er uttalt, og dermed må finne alternative måter å komme fram til eksemplene på. Et annet er Lohndal & Westergaard (2019) som analyserer V2-brudd i subjektsinitiale og ikke-subjektsinitiale setninger, syntaktiske funksjoner som ikke er annotert i korpuset, og dermed har de gått manuelt gjennom hele transkripsjoner for å sortere alle forekomstene.

Søfteland & Borthen (2018) er et eksempel på hvor langt man kan komme med søk på enkeltordsuttrykk. Arbeidet bak denne artikkelen om *den pragmatiske partikkelen 'sjø' i midt-norsk*, bygger i prinsippet på ett enkelt søk på *sjø* i NDC

(norsk del), uten fonetiske eller grammatiske avgrensninger. Dette søket ga 425 treff, inkludert 27 eksempler på *sjø* i betydninga 'hav, vann' som ble sortert bort (2018, s. 252). Det ble også gjort kontrollsøk på ulike fonetiske former (*sje*, *sju, sji* etc.), men dette genererte bare noen få eksempler til (2018, s. 252). Videre ble *geografiske* metadata brukt for å beskrive utbredelsen av *sjø*, med vekt på det trønderske målområdet, og i tillegg er metadata rundt *alder* og *kjønn* en del av analysen. Basert på de rundt 400 eksemplene fra NDC, gjør Søfteland & Borthen en detaljert pragmatisk-semantisk analyse av bruksmønstrene til partikkelen. Analysen viser at ytringer med *sjø* noen ganger er en forklaring og andre ganger en bekreftelse, noen ganger et hint om at innholdet i ytringa er nytt for mottakeren og andre ganger et tegn på at samtalepartnerne er enige, men at dette kan samles i én pragmatisk-semantisk analyse innenfor det relevansteoretiske rammeverket (2018, s. 278).

Riksem (2018) er et annet eksempel der utgangspunktet er ett enkelt søk, men med mer krevende sorteringsarbeid i etterkant for å sile bort ikke-relevante treff. Artikkelen handler om *språkblanding i nominalfraser i amerikanorsk*, og er et godt eksempel på hvordan korpusenes annotering med taggen 'Language X' kan brukes til datainnsamling. 'Language X' er en merkelapp som blir satt på ord som ikke står i Bokmålsordboka, både dialektord, engelske eller andre lånord og slanguttrykk. Arbeidet bak denne artikkelen starter med ett enkelt søk i CANS bare avgrensa med taggen 'X', altså etter alle ord i korpuset som har fått denne merkelappen (2018, s. 484). Deretter har Riksem gått gjennom hele trefflista og spesifikt valgt ut alle *engelskspråklige substantiv*. Dette resulterte i 1265 treff, som deretter ble kategorisert med tanke på språkblandingstematikken i artikkelen, med vekt på engelske ord i en ellers norskspråklig kontekst. Majoriteten av slike tilfeller har også norsk *morfologisk* kontekst, totalt 730 av eksemplene (2018, s. 485). Riksem gjør videre en detaljert syntaktisk analyse av engelsk–norsk språkblanding hos arvespråkstalere, innenfor et *eksoskeletalt* rammeverk, med følgende hovedfunn: Majoritetsspråket engelsk har satt preg på arvespråket amerikanorsk bl.a. gjennom at engelsk vokabular har blitt blanda inn i det. Det typiske mønsteret er at engelske innholdsord blir kombinert med funksjonelle elementer fra norsk. I tillegg er det en god del eksempler på engelske substantiv med engelsk *s*-ending, i ellers norskspråklige kontekster. Riksem konkluderer også med at den eksoskeletale tilnærminga fungerer godt som analyseverktøy for ulike typer språkblanding i et korpus med flerspråklig informantdata.

Spilling & Haugen (2013) omhandler *gradbøying*, med data fra NDC (norsk del) i tillegg til NoTa-Oslo og BigBrother-korpuset, totalt ca. 3,5 mill. ord. Det er to typer korpussøk som er gjort på dette materialet, begge uten fonetisk søkeavgrensning. For å finne gradbøyde *suffiksformer*: Søk etter alle ord som

slutter på *-re*, *-st* eller *-ste*. For å finne *perifrastiske* former: Søk etter *mer*, *mere* og *mest* etterfulgt av et åpent, valgfritt ord. I begge tilfeller er ikke-relevante treff manuelt sortert bort i etterkant. For de perifrastiske nevner Spilling & Haugen at dobbeltformer som *mer rarere* er luka ut. For 'slutter på'-søka er ord som ikke innebærer gradbøying tatt bort, men noen svært høyfrekvente suppletivformer er ikke gjennomgått, noe som bl.a. medfører at en infinitiv som *å bedre* kan ha blitt værende sammen med de mange eksemplene på en komparativform som *bedre* (2013, s. 6).[7] Spilling & Haugen gjør en bruksbasert analyse av tre hovedtyper gradbøying: *Metakomparasjon* (197 treff), *absolutt* gradbøying, inkludert komparativ med klassifiserende funksjon (450 treff) og *vanlig* gradbøying (ca. 20.000 treff totalt, men som nevnt er noen leksemer svært frekvente (jf. 2013, s. 19, tabell 2 og 3). Hovedfunna i artikkelen er at metakomparasjon (*serviset er mer grønt enn grått*) nesten bare brukes med perifrastisk bøying, mens absolutt gradbøying (*en bedre middag*) nesten bare brukes med suffiksbøying. For vanlig gradbøying (*Ola er større enn Per*) brukes begge typer, men ulike leksemer er gjerne sterkt knytta til én av bøyingsmåtene. Ellers er morfologisk enkle adjektiv typisk bøyd med suffiks, mens mer komplekse er bøyd perifrastisk (2013, s. 37). Spilling & Haugen konkluderer med at at de to bøyingsmåtene har delvis overlappende betydning, men også sine spesialfunksjoner, og at perifrastisk gradbøying trolig er den mest produktive av de to.

Med data fra CANS er det gjort mye forskning på nominalfrasen i amerikanordisk de siste årene, både i språkblanding/kodeveksling (jf. Riksem 2018 mfl.) og for ordstilling og kongruens. Lohndal & Westergaard (2016) analyserer genus i amerikanorsk gjennom søk etter både enkelt- og flerordsuttrykk: 1) ubestemt artikkel etterfulgt av et substantiv, eventuelt med adjektiv imellom 2) possessiv 3) substantiv i bestemt form. I artikkelen blir funna også sammenlikna med data fra NDC. Lohndal & Westergaard finner ca. 1000 fraser med ubestemt artikkel og substantiv i CANS, og blant disse er 76 % hankjønn, 17 % hunkjønn og 7 % intetkjønn (2016, s. 7). Sammenlikning i rene tall viser at dette er mindre bruk av intetkjønn enn i NDC, ca. like mye hunkjønn som de *eldre* i NDC (de yngre der har mindre) og litt mer hankjønn enn de *yngre* i NDC (de eldre der har mindre). I den videre analysen av ulike attesterte *enkeltord*, kommer det derimot fram en klar forskyvning mot hankjønn (F>M, N>M) hos de amerikanorsk-talende, og delvis også mot hunkjønn (N>F) (jf. 2016, s. 9, tabell 3). Mønsteret med forskyvning mot hankjønnsmorfologi gjelder ikke for bestemt

[7] Spesifisering av grammatiske trekk ser ikke ut til å være brukt i søkeprosessen; søka er ikke avgrensa etter f.eks. 'adjektiv med komparativ bøying'. En mulig forklaring på dette kan være at det har generert for få og/eller for mange treff, altså at den grammatiske taggeren ikke er helt treffsikker når det kommer til komparativ- og superlativ-bøyingsendinger.

form-suffikser, disse er gjennomgående målspråkslike (2016, s. 12). En av konklusjonene er at dataene illustrerer *attrisjon* og ikke ufullstendig språktilegnelse (2016, s. 12).

Johannessen & Larsson (2015) har delvis samme forskningstema, men inkluderer også amerikas*vensk*. Analysen er basert på korpusdata etter søkekombinasjonen *(determinativ) + adjektiv + substantiv*. De finner totalt 171 fraser der genuskongruens er et relevant studieobjekt (131 det+adj+subst og 58 adj+subst). Bare 21 av disse har ikke-målspråkslik kongruens, nesten alle av typen *med* determinativ (jf. 2015, s. 7, tabell 3). De viser også at det er store individuelle forskjeller, og en stor del av artikkelen inneholder detaljerte analyser av enkeltinformanter fra begge språk, inkludert noen som har hatt tett kontakt med Europa-norsk/-svensk i voksen alder. En av konklusjonene er at trekjønnssystemer i seg selv ikke er mer sårbare enn tokjønnssystemer (2015, s. 13). Genus er generelt godt bevart, men det er flere målspråksavvik i mer komplekse nominalfraser, noe Johannessen & Larsson tolker som prosesseringsvansker (2015, s. 18). Metadata-informasjon bidrar til den overordna analysen at startalder for innlæring av majoritetsspråket engelsk er av stor betydning for attrisjonsgraden.

Anderssen, Lundquist & Westergaard (2018) ser på ordstilling i possessivkonstruksjoner i CANS med utgangspunkt i korpussøk etter possessiv i prenominal og postnominal posisjon [poss+subst/subst+poss]. Etter at faste uttrykk er sortert bort, står det igjen 756 treff. Blant disse er et klart mindretall prenominale (17 %), noe som er (enda) mindre enn i tidligere studier av voksne førstespråkstalere av norsk (25–27 % (2018, s. 755)). Dette er et litt overraskende funn siden engelsk bare har prenominale possessiver, og man dermed kunne forvente en overgang til dette som det mest høyfrekvente mønsteret, i hvert fall hos noen av informantene (2018, s. 754). Anderssen et al. viser at det er store forskjeller mellom informantene, både for possessiv-plassering og dobbel definitthet, som også er analysert i samme materiale. Informantene deler seg i to grupper, de som foretrekker de typisk norske mønstrene (subst+poss, dobbel definitthet), og de som overbruker engelske mønstre (poss+subst, enkel definitthet) og samtidig blir målt med lavere norskkompetanse. En av konklusjonene er at den sistnevnte gruppa er berørt av CLI, *tverrspråklig påvirkning*, mens mønstrene i den første gruppa heller kan analyseres som *tverrspråklig overkorreksjon,* CLO (2018, s. 760).

Det finnes mange eksempler på morfologi-studier av NDC; i tillegg til Spilling & Haugen kan vi nevne Garbacz (2014) om dativ i norske dialekter og Knooihuizen (2014) om et utvalg bøyingssuffiks i færøysk. Garbacz gjør flere ulike søk med fonetisk avgrensning, etter varianter av hankjønnspronomenet

(*honom, hånom, honnom, hånnom, håno, hono, hånno, honno*), og varianter av flertallsbøyde substantiv, da med søkefunksjonen 'ord som slutter på' (*-om, -åm, -å* og *-o*). Resultatet etter sortering er 17 treff på hankjønnspronomen med dativform, fordelt på 7 målepunkter, og 156 treff på flertallsord med dativform, fordelt på 35 målepunkter. Gjennom bruk av kartfunksjonen i korpusgrensesnittet viser Garbacz at den geografiske utbredelsen av dativ er i Midt-Norge, aller mest i Oppland og Hedmark, men også i Møre og Romsdal og Trøndelag. Et hovedfunn er at preposisjonsstyrt dativ er klart mer frekvent enn verbstyrt dativ.

Knooihuizen analyserer morfologisk variasjonsdistribusjon med mål om å svare på hvilken retning utvikling av standardtalespråk for færøysk tar. *-st* er en tradisjonell bøyingsending for 2. person entall av sterke verb, og *-ir* og *-ur* er frekvente bøyingsendinger i både substantiv, adjektiv og verb, men begge tilfeller er kjent som språktrekk med mye variasjon. Når det gjelder *-st* finner Knooihuizen 266 aktuelle verbformer i NDC (færøysk del). 89 av disse er eksempler på *vita* brukt som diskursmarkør, og der er formen *tú veit* gjennomgående leksikalisert uten *-s-* (2014, s. 98). Analysen av de 177 andre treffa viser at *-st*-ending er klart mer i bruk i Torshavn enn f.eks. Sandur, og at *-st*-verbform er mer frekvent når det står foran subjektet (V–S) enn når det står etter (S–V) (jf. 2014, s. 97, tabell 3). Dialektgeografiske forskjeller ser ikke ut til å bli utflata hos de unge, heller forsterka (2014, s. 98). Mulige *-ir*/*-ur*-kontekster er høyfrekvent, og derfor har Knooihuizen her brukt bare rundt halvparten av korpusmaterialet. Utvalget gir ca. 1500 eksempler, ¾ *-ur* og ¼ *-ir* i ortografisk transkripsjon. Hovedspørsmålet i analysen av disse er om de to endingene har gått sammen til én. Dataene viser en klar tendens til 'nøytralisert' uttale av begge, men samtidig at det tradisjonelle systemet er bevart ved ikke-nøytral uttale (2014, s. 99-100). En av konklusjonene er at korpusmaterialet er for snevert til å kunne drive mer omfattende sosiolingvistisk forskning, men at dataene som er analysert peker mot at det ikke er noen større dialektnivelleringsprosess på gang i færøysk (2014, s. 101).

En syntaksstudie av NDC som inkluderer alle de fem språka i korpuset, er Bentzen (2014). Her er temaet ordstilling i setningsinnledning med fokus på adverbet *kanskje* (svensk *kanske*, islandsk *kannski,* færøysk *kanska*, dansk *måske*). Analysen bygger på følgende korpussøk: *kanskje*+V, *kanskje*+XP og XP+*kanskje*, altså setningsinitial *kanskje* etterfulgt av finitt verb (V2-struktur), setningsinitial *kanskje* etterfulgt av en annen frase og finitt verb senere (ikke-V2) og en annen frase setningsinitialt etterfulgt av *kanskje* og finitt verb senere (ikke-V2). Fra dansk, islandsk og færøysk er det få relevante treff i korpuset, men artikkelen inneholder også grammatikalitetsvurderinger av tilsvarende setninger fra

Nordisk syntaksdatabase (NSD), og der kommer det bl.a. fram klare forskjeller i dansk og norsk for setninger som *Kanskje kommer Peter ikke* (norsk: lav, dansk: høy) og *Kanskje han ikke kommer* (norsk: høy, dansk: lav) (2014, s. 228/231). For svensk og norsk finner Bentzen totalt 188 relevante eksempler i NDC. I begge språka er *ikke-V2*-struktur klart vanligst, 122 av 126 for norsk og 60 av 62 for svensk, men det er tydelige forskjeller i de syntaktiske mønstrene: I de norske dataene er det klart vanligst med *kanskje* først og deretter subjektet eller en annen frase (*kanskje*+XP+V, 99 treff). I de svenske dataene er det derimot klart vanligst med subjektet eller en annen frase setningsinitialt før *kanskje* (XP+*kanskje*+V, 57 treff), f.eks. *Du kanske har lite frågor att komma med*. Spesielt interesssant er det at både grammatikalitetsvurdering (NSD) og spontantale (NDC) viser at når subjektet er en full DP (ikke pronomen), kan setninger med *kanskje* medføre en syntaktisk kontekst der ikke-V2 er eneste mulighet i norsk (2014, s. 237).

De fleste artiklene vi har sett på, viser data fra noen enkeltsøk over større språkområder. Stjernholm & Søfteland (2019) kan nevnes som et eksempel på det motsatte; her blir to målepunkt i NDC (Fredrikstad og Aremark) detaljanalysert på jakt etter nåtidens status for tradisjonelle målmerker. Her kombineres mange typer søk, ofte med fonetiske avgrensninger og studier av uttaledistribusjon – for bl.a. pronomen, nektingsadverb, tjukk l, senkning, monoftongering og førsteleddstrykk. En av konklusjonene er at en god del tradisjonelle dialekttrekk er i bruk, men at det er store individuelle variasjoner, også på tvers av metadata som alder og kjønn.

[6] OPPSUMMERING

I artikkelen har vi vist hvordan Glossa er utvikla med tanke på korpusbrukernes behov og ønsker. Vi har presentert statistikk for faktisk bruk av NDC og CANS, som viser at en stor andel av søka er etter enkeltord i den ortografiske transkripsjonen, men også at søk etter flerordsuttrykk eller fraser er frekvent. De fleste søka spesifiserer bare ordformer, men den tilgjengelige grammatiske informasjonen er også nokså mye i bruk, og de fleste søk blir gjort med henblikk på metadata. I presentasjonen av anvendelse i vitenskapelige publikasjoner har vi sett eksempler på studier av morfologi, syntaks og pragmatikk, i enkeltord og sammensatte strukturer, for store populasjoner og enkeltinformanter, på tvers av flere språk og i enkeltdialekter, med og uten fokus på uttalevariasjon og metadata. Konklusjonen må være at korpusene er velfungerende i bruk, både kvantitativt og kvalitativt.

TAKK

Denne artikkelen bygger på arbeidet til mange forskere, ingeniører og transkribører. Takk til alle som har bidratt til spennende forskning ved bruk av korpus og til dem som har vært med på å bygge dem.

Den største takken rettes til Janne. Uten henne hadde verken Tekstlaboratoriet, Glossa eller korpusene vært det de er i dag. Hun var en entusiastisk igangsetter, pådriver, inspirator, faglig veileder og god kollega. Takk!

REFERANSER

Amerikanordisk talespråkskorpus (CANS): Johannessen, Janne B. 2015. The Corpus of American Norwegian Speech (CANS). I Beata Megyesi (red.), *Proceedings of NODALIDA 2015. (NEALT Proceedings Series 23.)* (www.tekstlab.uio.no/norskiamerika/korpus.html)

Anderssen, Merete, Lundquist, Björn & Westergaard, Marit. 2018. Cross-linguistic similarities and differences in bilingual acquisition and attrition: Possessives and double definiteness in Norwegian heritage language. *Bilingualism: Language and Cognition* 21(4), 748–764.

Bentzen, Kristine. 2014. Verb placement in clauses with initial adverbial *maybe*. *Nordic Atlas of Language Structures Journal* 1, 225–239.

BigBrother-korpuset: Tekstlaboratoriet, ILN, Universitetet i Oslo. www.tekstlab.uio.no/nota/bigbrother/

CLARINO: https://clarin.w.uib.no/

Faarlund, Jan Terje, Lie, Svein & Vannebo, Kjell Ivar. 1997. *Norsk referansegrammatikk.* Oslo: Universitetsforlaget.

Garbacz, Piotr. 2014. Morphological dative in Norwegian dialects. *Nordic Atlas of Language Structures Journal* 1, 72–86.

Johannessen, Janne B. & Larsson, Ida. 2015. Complexity matters: On gender agreement in heritage Scandinavian. *Frontiers in Psychology* 6. (artikkel 1842)

Knooihuizen, Remco. 2014. Variation in Faroese and the development of a spoken standard: In search of corpus evidence. *Nordic Journal of Linguistics* 37(1), 87–105.

Lohndal, Terje & Westergaard, Marit. 2016. Grammatical gender in American Norwegian heritage language: Stability or Attrition? *Frontiers in Psychology* 7. (Artikkel 344)

Lohndal, Terje & Westergaard, Marit. 2019. Verb second word order in Norwegian heritage language: Syntax and pragmatics. I David W. Lightfoot & Jonathan Havenhill (red.), *Variable properties in language: Their nature and acquisition*, 91–102. Washington DC: Georgetown University Press.

Nordisk dialektkorpus (NDC): Johannessen, Janne B., Priestley, Joel J., Hagen, Kristin, Åfarli, Tor A. & Vangsnes, Øystein A. 2009. The Nordic dialect corpus: An advanced research tool. I Kristiina Jokinen & Eckhard Bick (red.), *Proceedings of NODALIDA 2009.* (*NEALT Proceedings Series Vol. 4.*) http://www.tekstlab.uio.no/nota/scandiasyn/index.html

Nordisk syntaksdatabase (NSD): Lindstad, Arne M., Nøklestad, Anders, Johannessen, Janne B, Vangsnes, Øystein A. 2009. The Nordic Dialect Database: Mapping microsyntactic variation in the Scandinavian languages. I Kristiina Jokinen & Eckhard Bick (red.), *Proceedings of NODALIDA 2009.* (*NEALT Proceedings Series Vol. 4.*) https://tekstlab.uio.no/nsd

NoTa: Norsk talespråkskorpus – Oslodelen, Tekstlaboratoriet, ILN, Universitetet i Oslo. http://www.tekstlab.uio.no/nota/oslo/index.html

Nygård, Mari. 2016. Talespråkssyntaks: En analyse av norske diskursellipser. *Norsk Lingvistisk Tidsskrift* 34(1), 5–24.

Nøklestad, Anders, Hagen, Kristin, Johannessen, Janne B., Kosek, Michal & Priestley, Joel. 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.

Riksem, Brita R. 2018. Language mixing in American Norwegian noun phrases. *Journal of Language Contact* 11(3), 481–524.

Spilling, Eivor F. & Haugen, Tor A. 2013. Gradbøying i norsk: En bruksbasert tilnærming. *Maal og Minne* 2013:2, 1–40.

Stjernholm, Karine & Søfteland, Åshild. 2019. Talemål i Sør-Østfold: Ideologi, struktur og praksis. *Målbryting* 10, 101–131.

Søfteland, Åshild & Borten, Kaja. 2018. 'Æ e trønder, æ, sjø': Den pragmatiske partikkelen 'sjø' i midt-norske dialektar. *Norsk Lingvistisk Tidsskrift* 36(2), 249–280.

SUMMARY

In this article we show how the search interface Glossa has been developed in step with the various corpora that have been built at the Text Laboratory. Furthermore, we present statistics on what kind of searches people do – single words or longer phrases, with or without specifications for phonetic form or grammatical features etc. – focusing on the Nordic Dialect Corpus and the Corpus of American Nordic Speech. Finally, we demonstrate how researchers have searched for data in these corpora and used them in published articles – both simple and extended search, in smaller or larger language areas – within several different branches of linguistics.

KONTAKT

Åshild Søfteland
Høgskolen i Østfold
ashild.softeland@hiof.no

Anders Nøklestad
Universitet i Oslo
anders.noklestad@iln.uio.no

Joel Priestley
Universitet i Oslo
joel.priestley@iln.uio.no

Kristin Hagen
Universitet i Oslo
Kristin.hagen@iln.uio.no