Oslo-korpuset av taggede, norske tekster

Janne Bondi Johannessen og Anders Nøklestad
Universitetet i Oslo

1. Innledning

Det har etter hvert blitt klart for mange språkforskere at et elektronisk tekstkorpus kan være et uvurderlig hjelpemiddel. Ved å foreta noen få tastetrykk kan man få ut store mengder eksempler på bruken av et ord, en bøying, en sekvens av ord osv. på en svært oversiktlig måte (hvorav den ypperste er en énlinjes konkordans). Dersom korpuset i tillegg er grammatisk tagget, dvs. at hvert ord er forsynt med ordklasse og morfosyntaktiske trekk, har man enda flere muligheter; man kan søke etter alle ord av en bestemt ordklasse, alle varianter av et grammatisk flertydig ord som har en bestemt type trekk, alle tilfeller av en sekvens der første ordet er et bestemt leksem og det andre en bestemt ordklasse osv. Oslo-korpuset er det første store taggede tekstkorpuset for norsk, og vil bli presentert i det følgende.

Et problem med mange taggede tekstkorpus er at søkingen blir komplisert. Det er vanlig å tilby brukerne å benytte regulære uttrykk, som består av ikke-selvforklarende tegn som klammer, parenteser, loddrette streker, et-tegn, prikker og stjerner, for å nevne noe. Terskelen for å bruke korpuset blir på denne måten nokså høy for dem som ikke er vant til slike uttrykk fra før. For Oslo-korpuset har vi valgt å gå en annen vei når det gjelder brukergrensesnitt. Et tekstkorpus skal vanligvis brukes av lingvister og filologer. De fleste i denne gruppen har ikke spesielle kunnskaper i informatikk eller formaliserte systemer. Det er vår erfaring at de kvier seg for å benytte korpus pga den høye terskelen for å skrive inn søkeuttrykkene, eller for å finne ut hva slags grammatisk merking som finnes og hvordan taggene spesifikt ser ut eller hvordan de ulike tekstene forkortes, dersom man bare ønsker et mindre tekstutvalg. Ut fra denne erfaringen har vårt mål vært at brukergrensesnittet må være enkelt.

2 Selvforklarende søkegrensesnitt

2.1 Enkle ord

Det aller viktigste må være at de aller enkleste søkene er totalt selvforklarende. Man skal ikke behøve å skrive inn anførselstegn, punktumer, stjerner eller lignende bare for å søke på et enkelt ord - eller for to ord etter hverandre. Betrakt følgende utsnitt fra hovedsøkesiden:

(1)

Øverst til venstre kan brukeren skrive inn et helt ord. Øverst til høyre (ikke synlig på dette utsnittet) kan brukeren skrive inn et ord til, slik at det kan søkes på to ord i sekvens. Mellom disse to er det en rute hvor man kan angi opptil hvor mange ord som kan komme mellom de to ordene. Under hvert ord kan man krysse av for om det man har skrevet inn, er begynnelse eller endelse av ordet (kan f.eks. brukes for prefikser og suffikser), eventuelt om det er ordets oppslagsform.

Det er altså lekende lett å søke etter for eksempel sekvensen *glad...lingvist* (hvor man ønsker både entall og flertall av disse ordene, og opptil fem ord mellom), man skriver simpelthen inn ordene i vinduene, fyller inn 5 i ruten mellom, og krysser av for grunnform. Alternativet ville vært å komponere et regulært uttrykk - i dette tilfellet som nedenfor (med systemet til CQP fra IMS Stuttgart):[1]

---

[1]For å gi leseren en idé om hva som må kunnes for å skrive et slikt uttrykk som i (2) (som man altså ikke behøver å skrive for Oslo-korpuset), forsøker vi oss på følgende beskrivelse av det aktuelle regulære uttrykket: Husk hakeparentes rundt søkeuttrykket. For å få alle forekomster av et leksem, be om at taggen skal være det som er omsluttet av anførselstegn: et punktum for å antyde en hvilken som helst bokstav fulgt av en stjerne for å antyde at det kan komme uendelig mange tegn av denne typen foran det vi strengt tatt er interessert i, fulgt av en bakoverskråstrek som sier fra om at neste tegn skal tydes som et teksttegn og ikke en del av språket til det regulære uttrykket. Neste tegn er

(2) [tagg=".*\"glad\".*"] []{0,5} [tagg=".*\"lingvist\".*"]

Det sier seg selv at man trenger mer enn fem minutter for å lære seg å skrive inn et slikt uttrykk. Man må jo lære seg syntaksen til regulære uttrykk, med de spesifikke deler som finnes i akkurat den versjonen man bruker.

2.2 Grammatiske tagger

Ofte er det også viktig å søke bare etter ord av en bestemt ordklasse. I vår nåværende versjon er det en liste over alle ordklassene, samt et par andre kategorier, på hovedsøkesiden:

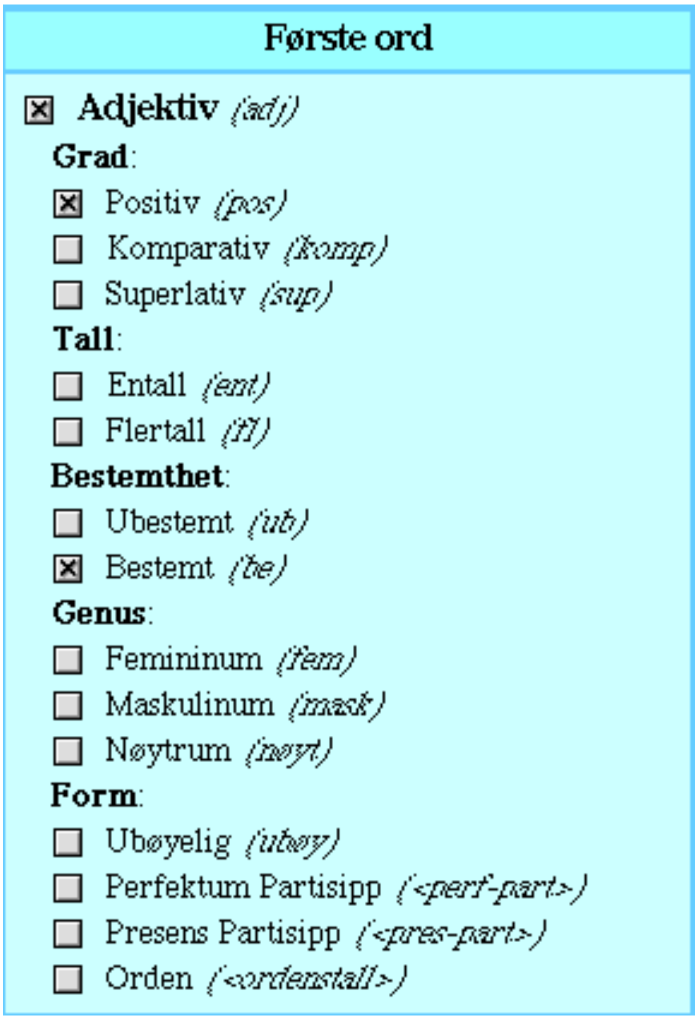
(3)

Ved å klikke på en av ordklassene, kommer man inn i et vindu hvor man kan krysse av mer spesifikt for akkurat de trekkene man er interessert i. Er man bare interessert i ord som er adjektiver i positiv form, og bare bestemt form, klikker man bare i de ønskede bokser:

---

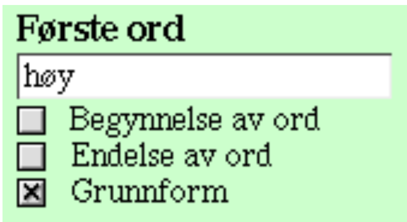
anførselsetegn, som betyr at vi ønsker oss noe fra et bestemt leksem, nemlig det følgende, som i dette tilfellet er *glad*. Så følger en ny bakoverskråstrek for å angi at neste tegn, altså anførselstegnet, skal tydes bokstavelig, så et punktum for et hvilket som helst tegn, som kan forekomme uendelig mange ganger, slik stjernen antyder. Anførselstegnet angir slutten på denne taggen, og hakeparentesen slutten på hele første del av uttrykket. Så kommer to hakeparenteser uten noe imellom, fulgt av en klammeparentes, og tallene 0,5 og slutt på klammeparentes. Dette antyder at det kan komme fra null til fem ord mellom det første vi søker på og det neste. Deretter kommer samme prosedyre for *lingvist* som for *glad*. Merk at dette er et enkelt søk, hvor vi ikke ber om spesielle grammatiske tagger, eller et bestemt tekstutvalg. Når man benytter det enkle klikk-og-skriv-grensesnittet, blir det vi skriver inn, oversatt til et slikt regulært uttrykk.

(4)



Er man bare interessert i disse adjektivformene av et bestemt adjektiv, som f.eks. *høy*, skriver man bare inn ordet, krysser av for grunnform, og så sender man søket av gårde:

(5)



Skulle man selv komponere et regulært uttrykk for dette, ville det bli slik:

(6)
[tagg=".*\"høy\".*" & tagg=".* adj.*"&tagg=".* pos.*"&tagg=".* be.*"]

Ikke bare er det nødvendig å lære hvordan det regulære uttrykket skulle skrives. Man må også lære hvordan de forskjellige grammatiske klassene og trekkene forkortes. Når man isteden anvender klikk-og-skriv-grensesnittet, behøver man ingen forkunnskaper. Her er et utsnitt av resultatet av søket:

(7)
AV/Ad96/01: fyller 95 år mandag, men den **høye** alderen til tross er han
AV/Ad96/01: om går rundt med angst. - De **høye** tallene kan i neste
AV/Ad96/01: merkbart ned som følge av de **høye** kraftprisene. I oktober
AV/Ad96/01: r, Terje Borgos mener at den **høye** fellingsprosenten er et
AV/Ad96/01: 2. Rut Tellefsen er blant de **høytstående** fruene som

Man kan også søke på grammatiske tagger uten å angi et bestemt ord. Denne muligheten ser vi på som viktig, og gjør faktisk Oslo-korpuset mer fleksibelt å bruke enn British National Corpus, hvor man bare kan søke på grammatisk tagg knyttet til et bestemt ord (Reference to the SARA Windows Client, ch.3.5). Med de samme kriterier for søk som ovenfor, men uten at søket er knyttet til adjektivet *høy*, får vi et slikt resultat:

(8)
Søkestreng: [tagg=".* adj.*" & tagg=".* pos.*" & tagg=".* be.*"]

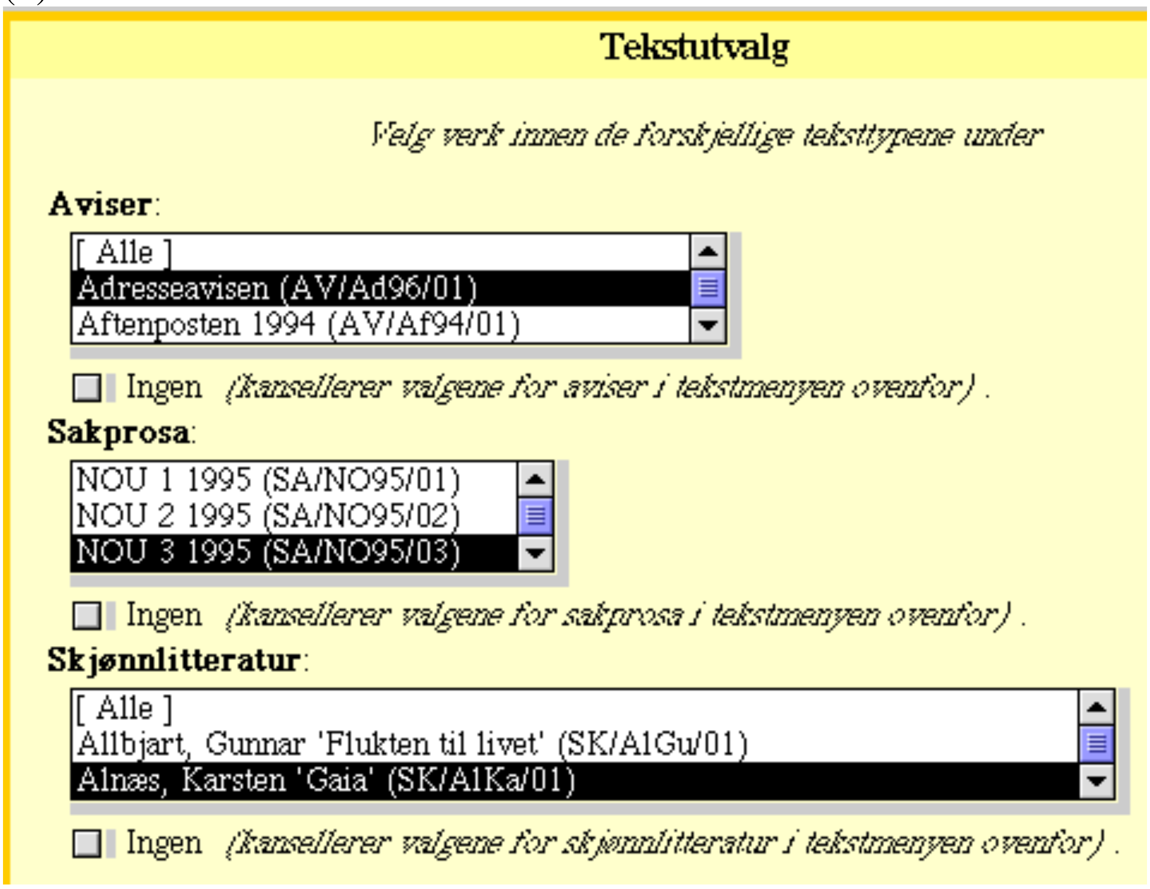
SK/AlGu/01: ham der han står - bare dette **fredfulle** suset av trær som
SK/AlGu/01: kestemmer - han stirrer inn i **brustne**
SK/AlGu/01: . Hans egne hender brenner i **blodige** flammer - og det
SK/AlGu/01: ikke vann som kan tvette dem **rene**. Han ser sig selv vandre
SK/AlGu/01: er i ham - den har hugget sig **fast** og vil ikke slippe byttet
SK/AlGu/01: ham - det blev svidd inn den **forferdelige** natten da han
SK/AlGu/01: fulgt, men siden - efter det **forferdelige** øieblikket ved
SK/AlGu/01: den døde, så øinene som var **brustne** og blodet som var

2.3 Tekstutvalg

Det er velkjent at enkelte korpus er satt sammen etter svært fastlagte kriterier, hvor man skal ha en bestemt prosentsats av hver genre, for å lage et såkalt "representativt" korpus. (Brown)) Det finnes også andre korpus, hvor målet tvert i mot er å ikke dekke en rekke genrer, men å oppnå f.eks. maksimal størrelse (Bank of English). Vi mener at det er en fordel at brukerne kan bruke et korpus til å sammenligne genrer. Dessuten vet vi jo alle at det er stor forskjell på språkbruken i ulike typer tekster. Dersom man er på jakt etter eksempler på en bestemt grammatisk konstruksjon, bør derfor korpuset ha en viss genrevariasjon. Samtidig ser vi ikke noe stort poeng i at det skal være like mye tekst fra hver genre. Ut fra disse premissene har vi konstruert Oslo-korpuset, som dermed består av flere genrer. Vi har valgt ut tre hovedtyper: avis (aviser og ukeblader), sakprosa (særlig offentlige rapporter og lovtekster) og skjønnlitteratur (romaner). Det er nærmere 19 millioner ord i bokmålsdelen, men fordelingen er ikke slik at det er en tredjedel av hver genre. Vi har forsøkt å ha en viss mengde av hver type, men fordi vi i stor grad brukte hva vi hadde, er den endelige mengden slik: aviser: ca. 10 mill. ord, sakprosa: ca. 7 mill. ord, skjønnlitteratur: ca. 2 mill. ord

I visse tilfeller ønsker brukeren bare å benytte en liten del av tekstene. F.eks. kan det være interessant å studere bare én avis, én forfatter, eller bare skjønnlitteratur. Igjen skal det ikke være nødvendig for brukeren å lære seg alle forkortelser for alle teksttyper. Det er bare å klikke seg inn på Tekstutvalgs-siden, og velge eller velge bort tekster. Standardutvalget er rett og slett alle tekstene. Nedenfor har vi valgt Adresseavisa, en NOU-rapport og romanen *Gaia* av Karsten Alnæs.

(9)



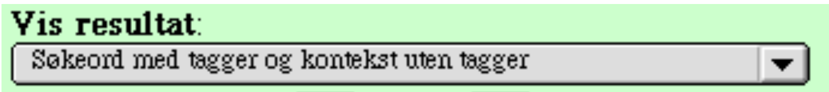
Uten et klikk-og skriv-grensesnitt måtte vi funnet ut hva slags koder disse tekstene var representert som, og skrevet inn en søkestreng av typen nedenfor (hvor søkeordet rett og slett er "ord"):

(10)
[word="ord"&(src="AV/Ad96/01"|src="SA/NO95/03"|
src="SK/AlKa/01")]

3. Visning av søkeresultatet
3. 1 Visning med og uten tagger

Når et korpus er tagget, kan man bruke det til å søke på bestemte tagger. Det er én måte, og den er vist ovenfor, (8). Har man søkt på én bestemt tagg, er det som regel unødvendig å se taggene også i søkeresultatet. Likeledes er det unødvendig å se taggene hvis man bare er ute etter eksempler på et gitt ord. Det er nok denne tankegangen som ligger bak det faktum at de fleste søkbare korpus ikke lar en få se taggene overhodet (et par eksempler er LDC Online og det svenske Parole-korpuset).

Men vi mener at det godt kan være tilfeller der det er viktig for brukeren å få se taggene. Det kan f.eks. dreie seg om en undersøkelse av når et bestemt ord forekommer som den ene eller andre ordklassen. I Oslo-korpusets grensesnitt har vi lagt inn en meny for hvordan søkeresultatet skal vises. Ovenfor, i (1), så vi standardsettingen, hvor ingen tagger vises. Nedenfor ser vi et valg hvor selve søkeordet vises med tagger, mens konteksten er uten tagger:

(11)


Søkeresultatet ser da slik ut (ved siden av søkeordet er ordets grunnform, ordklasse og morfosyntaktiske trekk, og til slutt syntaktisk funksjon angitt etter en krøllalfa):

(12)
Søkestreng: [word="stemme" & (src="AV/Ad96/01")]

AV/Ad96/01: re hele det norske folk ved å **stemme "stemme" verb inf @IV** mot, eller utfordre NATO-kolleger ved

AV/Ad96/01: utfordre NATO-kolleger ved å stemme **"stemme" verb inf @IV** for. I slike saker gjelder diplomatiet

AV/Ad96/01: lte Dusan Tadic. Med usikker stemme **"stemme" subst mask appell ent ub @<P-UTFYLL** forteller hun om drap, tortur og
AV/Ad96/01: re hele det norske folk ved å stemme **"stemme" verb inf @IV** mot, eller utfordre NATO-kolleger ved

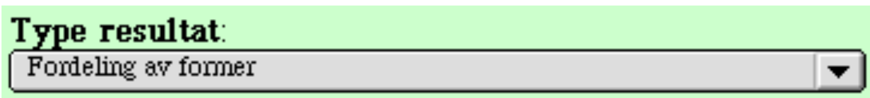
AV/Ad96/01: nt læresveinene og sa med høy **stemme "stemme" subst mask appell ent ub @<P-UTFYLL** : " Kong Inge er død. Håkon

På samme måte kan man be om at både søkeord og kontekst vises med tagger. Selvsagt blir søkeresultatet da litt mer rotete å se på, men for noen typer søk er slik informasjon svært verdifull. F.eks. kan man ønske å finne ut hva slags grammatiske omgivelser et bestemt verb har. Da er det uvurderlig med muligheten til å få vist tagger.

3.2 Type søkeresultat

Søkesystemet vi bruker i Oslo-korpuset er bygget over CQP-systemet fra IMS Stuttgart, som vi først overførte til web i vårt bosniske korpus. I dette systemet gis det muligheten til å få vist søkeresultatene på flere måter. Vi har valgt å presentere mulighetene i et menysystem. Ovenfor, i (1), kan man se at det er KWIC-konkordanse som er standardsettingen. Men det er også mulig å få vist resultatet i form av fordeling av former, eller fordeling i forhold til kilde, eller en kombinasjon av flere muligheter. Nedenfor kan man se at fordeling i forhold til former er valgt:

(13)



Resultatet (med søking på *be*- som forledd) ser slik ut:

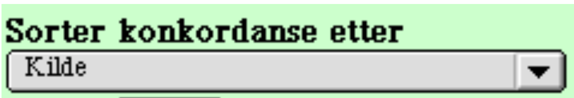
(14)
Søkestreng: [word="be.*"]
Søk etter: Fordeling av former
Fordeling

beste	59
bedre	43
betyr	34
begge	25
begynte	24
behov	24
bedrifter	19
besøk	19
best	18
bedriften	17
betydelig	17
betale	16
beskjed	13
bestemt	13
ber	12

3.3 Sortering av søkeresultatet

I visse tilfeller kan man ønske seg å se søkeresultatet sortert etter bestemte kriterier. Det vanlige standardkriteriet er sortering etter kilde:

(15)

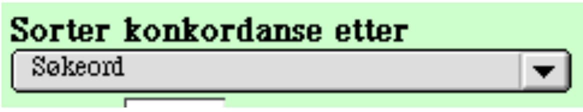


Søkeresultatet blir vist slik at alle resultater fra én kilde vises før resultatene fra neste osv.:

(16)
Søkestreng:
[word="svensk" & (src="AV/Ad96/01" | src="SA/Lo81/01")]
Sortert etter kilde

AV/Ad96/01: gene i sin debutsesong og ble **svensk** toppscorer med klar
AV/Ad96/01: aland har vært en sensasjon i **svensk** seriefotball etter
AV/Ad96/01: a Ullevi og skrøt uhemmet til **svensk** presse. En nytent
AV/Ad96/01: spiller som ble solgt til en **svensk** klubb. Bent
AV/Ad96/01: tland i vinter. Av en antatt **svensk** bestand på 1100 gauper
SA/Lo81/01: ansk, finsk, islandsk eller **svensk** domstol gjelder også i
SA/Lo81/01: ansk, finsk, islandsk eller **svensk** dom eller fastsetting fra
SA/Lo81/01: s gjennom dansk, finsk eller **svensk** apotek eller av enkelte

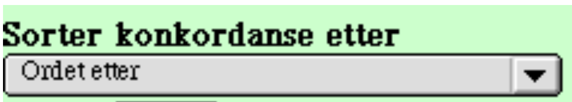
Det er også mulig å sortere søkeresultatene etter søkeordet selv. Det kan være nyttig når man søker på et variabelt søkeord, f.eks. alle ord som begynner med *svensk-*:

(17) 

(18)
Søkestreng:
[word="svensk.*" & (src="AV/Ad96/01" | src="SA/Lo81/01")]
Sortert etter søkeord

AV/Ad96/01: gene i sin debutsesong og ble **svensk** toppscorer med klar
AV/Ad96/01: aland har vært en sensasjon i **svensk** seriefotball etter
AV/Ad96/01: dsstillende. PALME-FILM | En **svensk** film om mordet på
AV/Ad96/01: sjef : Tore Gullen trener det **svenske** landslaget, for andre
AV/Ad96/01: r tiden har treningssamling i **svenske** Bruksvallarna, åtte
AV/Ad96/01: rningsmannen, slik en gruppe **svenske** eksperter mener
AV/Ad96/01: med at det nå er to år siden **svenskene** sa ja til EU, brakte
AV/Ad96/01: treken av Hellas. Så selv om **svenskene** produserer Volvo
AV/Ad96/01: treken av Hellas. Så selv om **svenskene** produserer Volvo
AV/Ad96/01: før i tiden, mener de fleste **svensker**. Men mens « før i
AV/Ad96/01: oregått i samarbeid med flere **svensker**, har ifølge
AV/Ad96/01: i Norge var lite tilfreds med **svenskproduserte** dyner
AV/Ad96/01: Verdal er på utkikk etter en **svenskregistrert** blå Audi

Videre kan man sortere søkeresultatene etter ord foran eller etter søkeordet. Det kan for eksempel være nyttig når man er på jakt etter betydningen til forskjellige preposisjoner som følger etter et gitt verb. Da kan det være kjekt å få alle eksempler på samme preposisjon sammen:

(19) 

Resultatet blir da slik:

(20)
Søkestreng: [word="skriver"]
Sortert etter ordet etter

AV/Ad96/01: e vinskjønnerne i avisene som **skriver for** medlemmer av
AV/Ad96/01: en i motsetning til andre som **skriver for** ulike

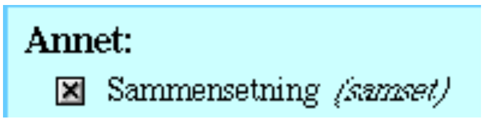
AV/Ad96/01: feller i Oslo, så selv om vi **skriver fra** bygdemiljø,
AV/Ad96/01: isten lavere. EU-kommisjonen **skriver i** en rapport at
AV/Ad96/01: rand, Øya Eiendoms advokat, **skriver i** et brev til
AV/Ad96/01: oreningsleder Bjørn Andersson **skriver i** et brev til
AV/Ad96/01: til NTB. Bioteknologinemnda **skriver i** sin uttalelse at
AV/Ad96/01: erk kontrast til det Huntford **skriver om** «Fram»-ferden.
AV/Ad96/01: språkformen. Ola Svein Stugu **skriver om** strukturell
AV/Ad96/01: usikkskolene, og Aftenposten **skriver om** lærerflukt fra
AV/Ad96/01: en. Forfatteren Leo Oterhals **skriver om** forliset i sin siste
AV/Ad96/01: rårets bok «Utror». Oterhals **skriver om** havets helter og

Vi har også lagt inn mange andre muligheter som gjør søkingen fleksibel. For eksempel kan man velge hvor mye kontekst som skal vises (opptil noen hundre tegn på hver side) for hvert søkeord. For å begrense søketiden, har vi også lagt inn en standard på at de første 1000 forekomster av søkeordet vises. Men dette tallet kan man endre både opp og ned ved å skrive noe annet i ruten. Vi har også lagt inn muligheten for å hente ut et tilfeldig utvalg av forekomster - tilfeldig i forhold til kilder. Det gir muligheten til f.eks. å begrense søket til 500 forekomster, men likevel få forekomster fra forskjellige deler av korpuset.

4. Uortodokse søkemuligheter med interessante resultater
4.1 Sammensetninger

Taggeren som er brukt, har selvsagt en sammensetningsanalysator som analyserer og bestemmer grammatisk tagg til produktivt dannede sammensetninger. Alle slike ord gis også en tagg, slik at man kan se at de er nyanalysert. Dette gir også muligheten til å søke på alle produktivt dannede sammensetninger i korpuset:

(21



En morsom bieffekt av dette, er at slike sammensetninger viser mye om forfatteren og/eller verket. Det er for eksempel lett å se at følgende utdrag er fra en mannlig forfatter, som skriver om politikk og krig (nærmere bestemt Nordahl Grieg).

(22)
Søkestreng: [tagg=".* samset.*" & (src="SK/GrNo/01")]

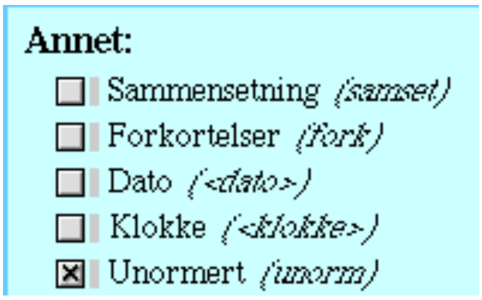
SK/GrNo/01: ilkjempet sig viktige goder, **firti-timers-dagen**. Tirret
SK/GrNo/01: vilde bli opfattet som en ny **klasseaksjon**, en
SK/GrNo/01: t som en ny klasseaksjon, **en maktutvidelse**. Det vilde bli
SK/GrNo/01: en. Blodige og åndeløse blev **kommunardflokkene**
SK/GrNo/01: terkere enn en ide. Idag kom **pariserarbeiderne** rolige, i
SK/GrNo/01: rolig omtanke for alle. I et **granathull** ved Brunete fikk
SK/GrNo/01: er over trekkene. **Krumrygget** bøide han sig over bordet,
SK/GrNo/01: r sommer, og deltok selv som **kampflyver**. Noe av
SK/GrNo/01: st. " Brevkortene var merket **Madrid-divisjonen**, og
SK/GrNo/01: , patronbelter, ammunisjon, **uniformsplagg** og m

Slik kan man sammenligne forskjellige tekster med tanke på ordforrådet. Man kan også sammenligne antallet sammensetninger i flere tekster, og slik si noe om forfatterens kreativitet.

4.2 Unormerte ord

I arbeidet med taggeren oppdaget vi at svært mange tekster har ord og bøying som ikke er i samsvar med normen fra Norsk Språkråd. Vi valgte å legge inn en god del slike ord i ordlistene våre, men med en ekstra tagg om at de er unormerte. Dette kan man så søke på, og slik se hva slags tekster og hva slags ord som avviker, og på hvilken måte:

(23)



De fleste av ordene nedenfor tilhører gruppen av "konservative" ord, altså ord som ligner svært på dansk.

(24)
Søkestreng: [tagg=".* unorm.*" & (src="SK/GrNo/01")]

SK/GrNo/01: elge. Det var papirflagg med **billedet** av kommunardene
SK/GrNo/01: m fortrolig aktelse, når han **bragte** glassene med
SK/GrNo/01: m gatene, endeløst. De gikk **seks-syv** i bredden, menn,
SK/GrNo/01: egi var sterkere enn en ide. **Idag** kom pariserarbeiderne
SK/GrNo/01: lgedagsklær, en halv million **mann**, med innflytelse i
SK/GrNo/01: sitt. Kirkegården var som en **kolonihave** av små stenhus;
SK/GrNo/01: n den var sperret av soldater **efter** flere attentatforsøk. Det

5. Avslutning

Det har vist seg at denne typen brukergrensesnitt slår an hos folk. For øyeblikket har vi ca. 150 registrerte brukere fra 15 land, og jevnlige tilbakemeldinger viser at folk bruker korpuset. Selv om vi har lagt vekt på at grensesnittet skal være brukervennlig og fleksibelt, har vi fått noen tilbakemeldinger som gjør at vi er i ferd med å endre det. Blant annet skal vi legge inn muligheten for å søke etter tre ord, vi skal ha muligheten til å søke negativt, og til å gi kollokasjoner som søkeresultat. Allerede nå er det mulig å gjøre en god del slike ting, fordi vi også har en søkeside hvor man kan skrive inn regulære uttrykk. Men som ventet er det bare ekspertbrukere som benytter seg av den, og det er faktisk tilbakemeldingen fra noen av disse som gjør at vi ønsker å utvide søkemulighetene i det enkle klikk-og-skriv-grensesnittet.

Web-referanser

Brown-korpuset:
http://www.ldc.upenn.edu/lol/docs/BROWN.html
Bank of English:
http://titania.cobuild.collins.co.uk/boe_info.html
CQP IMS:
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/
LDC Online:
http://www.ldc.upenn.edu/
Oslo-korpuset av taggede norske tekster:
http://www.tekstlab.uio.no/norsk/bokmaal/
http://www.tekstlab.uio.no/norsk/nynorsk/
Svenske korpus:
http://spraakbanken.gu.se/
The Oslo Corpus of Bosnian Texts
http://www.tekstlab.uio.no/Bosnian/Corpus.html
The SARA Windows Client:
http://info.ox.ac.uk/bnc/getting/chap4.htm

Janne Bondi Johannessen j.m.b.johannessen@ilf.uio.no
og Anders Nøklestad anders.noklestad@ilf.uio.no
Tekstlaboratoriet, ILF
Universitetet i Oslo
P.b. 1102 Blindern
0317 Oslo