



Siteringsforslag: "Frå bandsalat til bandbreidd: leksikografiske metodar i Revisjonsprosjektet for Bokmålsordboka og Nynorskordboka" (2019). Revisjonsprosjektet for Bokmålsordboka og Nynorskordboka. Universitetet i Bergen. Presentasjon gjeve av Gunn Inger Lyse på faglunsj ved LLE, Universitetet i Bergen, 23.05.2019."

Kontakt: ordbokene@uib.no.

Denne presentasjonen er lisensiert under ein Creative Commons Namngjeving 4.0 Internasjonal Lisens.

URL til lisens: <http://creativecommons.org/licenses/by/4.0/>

I dette innlegget: presentere kildematerialet, verktøy og metoder som brukes og utvikles i det norske Revisjonsprosjektet for Bokmålsordboka og Nynorskordboka. Seie litt om kva korpus vi brukar, kva søkjeverktøy vi har, Gje døme på når korpus er til god hjelp – men óg døme på at korpus kan vere misvisande.



UNIVERSITETET I BERGEN



Revisjonsprosjektet

for Bokmålsordboka og Nynorskordboka

- Hovudformål: oppdatere og forbetre Bokmålsordboka og Nynorskordboka (som Språkrådets viktigaste normeringsverktøy for bokmål og nynorsk).
- Tidsramme: 5½ år (oppstart haust 2018, prosjektslutt 2023)
 - Revidere eksisterande innhald;
 - Utjamne og samkøyre omfang og lemmautval mellom ordbøkene.



23.05.19

SIDE 2

Utg.punkt er kort sagt:

I dette prosjektet skal leksikografene revidere to eksisterende ordbøker parallelt, én på bokmål og én på nynorsk.

(3 siste kulepunkt)

Revidere eksisterande innhald (tydingsinndeling, definisjonar, etymologi, uttale);

UNIVERSITETET I BERGEN

Ordboksarbeidet i Revisjonsprosjektet

Ein balansejong mellom å arbeide



Effektivt Etterretteleg

23.05.19 SIDE 3



Spørsmålet vårt: Hvordan kan leksikografene jobbe effektivt, etterrettelig og klokt (med et godt empirisk grunnlag) for å peke ut et modernisert og relevant ordtilfang i begge målformene?

Nokre døme:

UNIVERSITETET I BERGEN



Frå bandsalat...

Begge Bokmål Nynorsk [Avansert søk](#)

Bokmålsordboka

Oppslagsord Ordbokartikkel

[båndsalat](#)
[bandsalat](#)

båndsalat m1; el. **bandsalat m1**
det at bånd (3) tyter ut slik at det er umulig å få det på plass igjen

Nynorskordboka

Vi har dessverre ingen informasjon om ordet 'b_ndsalat' i nynorskbasen.

Vanlege feilkjelder:

- Skrivefeil: Sjekk skrivemåten!
- Søk i feil ordbok: Bruk Begge-knappen!

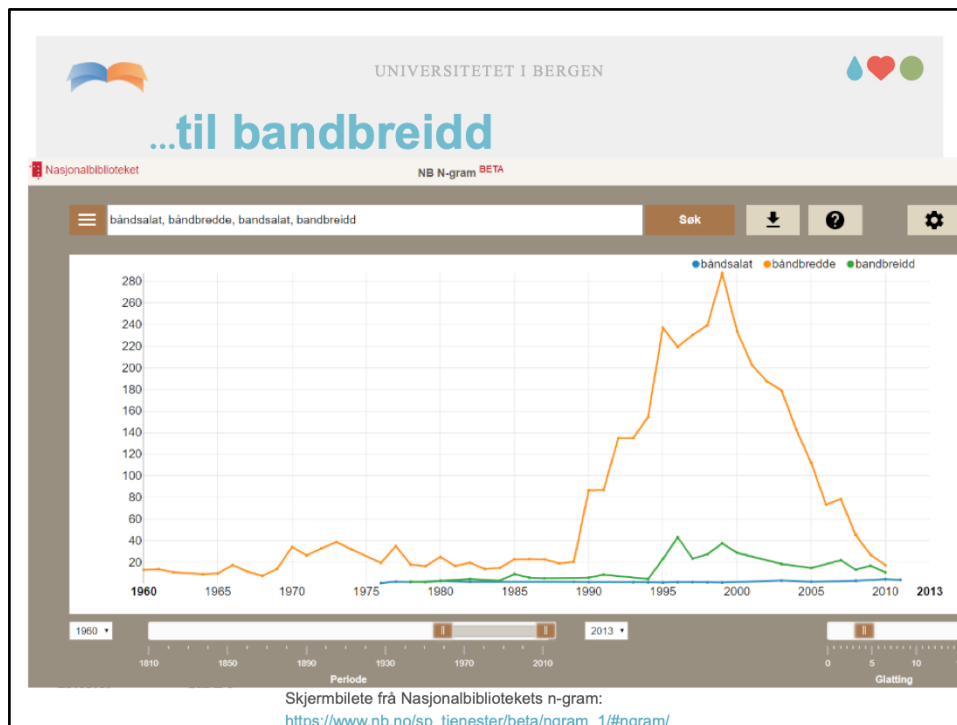


23.05.19 SIDE 4 FOTO: NTBScanpix

Bandsalat låg kun i BOB, men ikkje i NOB.

Bandsalat: som fenomen tona meir og meir ut då kassettspelaren vart erstatta av meir moderne måtar for å spele av musikk

Så er der til gjengjeld andre ord som fortener å kome inn, t.d. (klikk)



..bandbreidd.

Bandbreidd låg verken i BOB el NOB då prosjektet vårt starta opp.

UNIVERSITETET I BERGEN

Informasjonskjelder i Revisjonsprosjektet

- Språkrådet (normering)
- Språksamlingane ved UiB (usd.uib.no)
- tekstkorpus (lemmatilfang, tyding, syntaktisk åtferd, døme på typisk bruk, frasar og fleirordsuttrykk som orda inngår i)
- introspeksjon
- andre ordbøker

23.05.19 SIDE 6

Den empiriske forankringa i **leksikografiarbeidet er viktig.**

Vi brukar dagleg det vi kallar "korpus":

-Tekstkorpus –kolleksjon av elektronisk tekst som er søkbar, og der du óg kan analysere treffa.

Lemmutval: peike ut eit sentralt ordforråd

Identifisere i kva tydingar eit ord vert brukt ("skiballett" som eiga tyding av "ballett"?)

andenaud: vanskar med å dra anden (astma kan gje andenaud), men enda oftare: i overført tyding: mangel på naudsynte ressursar eller støtte (kommunen lir av økonomisk andenaud)

inf. om den syntaktiske "oppførsel"

Gode døme på typisk bruk av eit ord

Likevel nemne, for balansens skuld, at **ordbokaarbeidet ikkje kviler åleine på å sjå på data.** (liste)

-Språkrådet (slår opp i normeringsvedtak - normerer bøyning, ordklasse, stavemåte).

-Vitskaplege språksamlingar (som Metaordboka) – for å sjekke det vi ikkje finn belegg for i tekstkorpus, og/eller som allereie står i ordboka)

Anne: dlme på "corner" som har ei pokertyding som ikkje er lett å finne i korpus;



Informasjon vi treng frå korpus

- Lemmautval:
 - ordfrekvens, over tid og på tvers av domene
bandsalat, ballongbukse, barnehageplass,..
 - frekvens på bøyingsformer (t.d. partisipp, passiv)
betinget, begeistret
det var toppen plass til fem personar
 - ordkollokasjonar (fleirordsuttrykk)
få auge på, antikvarisk verdi,
lett/kjapp/rask til beins
- Tyding: konkordansar
- Syntaktisk åtfærd: konkordansar og kollokasjonar
- Gode bruksdøme: ordkollokasjonar
lovfestet rett til barnehageplass



Partisipp: INESS: lar oss lete spesifikt etter empirisk grunnlag for å se om et verb har kvaliteter som motiverer å legge dem inn som leksikaliserte adjektiver: At de brukes ofte attributivt (en avansert pianist), om de samsvarsbøyes: de er avanserte

UNIVERSITETET I BERGEN

Relevante verktøy for å søkje i korpus

- CLARINO Bergen Center:
 - Corpuscle (verktøy for søk på ordnivå)
 - INESS (verktøy for søk i setningsanalysar)
- Nasjonalbiblioteket: NB n-gram og ordgalaksar
- Retriever: Atekst (avistekst)
- Universitetet i Oslo: HaBiT (norsk webkorpus), LIA

23.05.19 SIDE 8

Prosjektet vårt har tilgang til forskjellig korpusmateriale, og her skal eg fokusere på dei to verktøya som er ramma inn øvst.

Dei svarar på to hovudbehov:

1. Vi treng å kunne søkje gjennom mykje tekst på ein enkel måte, mest mogleg fleksibelt. => kombinere etterretteleg + effektiv
2. Vi lenar oss på mest mogleg kvalitetssikra materiale (meir orientert mot skriftspråk enn talespråk)

Leksikografene bruker verktøy for språkanalyse som er tilgjengelig gjennom språkinfrastrukturen CLARINO (<https://clarin.w.uib.no/>) og som er distribuert gjennom UiBs CLARINO Bergen Cent

Corpuscle er et verktøy for å søke på ord og fraser. Gjennom revisjonsprosjektet er Corpuscle utvidet til å søke i flere korpus samtidig (til sammen ca. 2,4 milliarder ord). Infrastrukturen INESS lar oss søke etter syntaktiske konstruksjoner som ikke er så lett søkbare i et tradisjonelt tekstkorpus.

UNIVERSITETET I BERGEN

Corpuscle – søk og analyse (ordnivå)

eng | nob | Logg ut (Gunn Inger Lyse Samdal (gunn)) | Edit is av

CLARINO Korpuskel :: Aviskorpus (Bokmål) :: Konkordans

Avansert søk | bytt til Enkelt søk | Søkeshistorie ...

"b.ndsalat.*"

Søk | Raffiner vindu: document | Stop | Lagrede søk ... | Lagre søket som | felles

Done. Running time: 0.01 sec. (0.01 CPU sec.)

Type: kwic | Vis linjefilter | Attributter ... | Strukturer ... (vis i treff) | Linjer per side: | Kontekststørrelse: 300px

Treff 1 – 13 av 13 | Forrige | Neste | Last ned (Excel-modus) | Copy query PID

treff	source	year	date
imes Last har spunnet av gårde til den evige båndsalat. båndsalat. LAST JOURNEY: Kapteinen av or	VG	2015	2015-06-12
iar spunnet av gårde til den evige båndsalat. båndsalat. LAST JOURNEY: Kapteinen av orkestermuz	VG	2015	2015-06-12
issetten trenger oss, sier han og snurrer opp båndsalaten. Men han innrømmer at det er litt styr. Han	AP	2013	2013-07-21
et man aldri, det verste som kan skje er jo « båndsalat » og da er det tapt for alltid, forteller Hagen.	VG	2013	2013-02-16
« kjæreste May Irene Aasens mor. Frykter « båndsalat » DAT-taper, originale tegninger til Christo	VG	2013	2013-02-16
« jeg som enkelt og greit. Her slipper man å få båndsalat på DV-tapen, det er enkelt å skifte disk og e	AP	2008	2008-01-07
« nå. Dessverre vil noen si, med minner om en båndsalat som odela den perfekte milkstapen som den	VG	2008	2008-02-23
envisst til minnens dat, der vi kan mimre om båndsalat, susende lyd og kopiering av kassetter hjem	VG	2008	2008-02-23
« i produksjonsideer - det høres gjerne ut som båndsalat og unormal avspillingshastighet flere steder.	DB	2007	2007-01-16
« kvalitét og mang en kassettpiller odelagt av båndsalat, var kassetten et overlegent musikkmedium	BT	2006	2006-03-31
« å bekymre seg for at familievideoene ble til båndsalat, eller om videobåndet har nok igjen til å få i	DB	2002	2002-04-10
« bånd Av RUNE SKOGSETH Ray Charles laget båndsalat av NRK-kassetten som nesten stoppet den le	VG	1999	1999-07-17
« ir over. I baksetet moret han seg med å lage båndsalat av videotapen, forteller Svein Åge Johansen	VG	1999	1999-07-17

Treff 1 – 13 av 13 | Forrige | Neste

<http://clarino.uib.no/korpuskel/>

Corpuscle:



Grensesnitt for å tilgjengeleggjere, søkje i og analysere tekstkorpus.

Verktøyet er utvikla av Paul Meurer, som no jobber ved Universitetsbiblioteket (UBB). Verktøyet inngår no i UBBs Clarino Centre Bergen.

FØRST NOKO OM KVA DU FINN DER

- I dette grensesnittet får du tilgong til ulike typar korpus: nokon ligg med open tilgong, men nokre tekstkorpus har strengare tilgong grunna avtale med eigar av opphavsretten.
- både einspråklege og fleirspråklege korpus, på ulike språk.
- Der er korpus for ulike formål, med alt frå avistekst, Tekst skrivne av elevar, tli
- Transkriberte dialektopptak (prosjektet dialektendring)

Kva nyttar vi for revisjonsprosjektet? [klikk]

 UNIVERSITETET I BERGEN 						
"Corpuscle-Lex" (eigen server)						
Korpusnamn	Språk	Storleik (# ord)	Tidsrom	Sjanger	Lemma + ordklasse	Tilgang
Talk Of Norway	bokmål nynorsk	63,8 mill.	1999-2016	sakprosa	ja	open
Aviskorpus (bokmål)	bokmål	1509,1 mill.	1998-2015	sakprosa (avis)		open
NBs frie tekster (bokmål)	bokmål	516,4 mill.	1765-2013	blanda: sakprosa, skjønnlitteratur		open
Leksikografisk bokmålskorpus	bokmål	102,3 mill.	1985-2013	blanda: sakprosa, skjønnlitteratur	ja	avgrensa
Aviskorpus annotert	bokmål	29,0 mill.	2001-2009	sakprosa (avis)	ja	open
Forskning.no (2017)	bokmål	21,5 mill.	1998-2017	sakprosa (avis)	ja	avgrensa
Norsk Ordboks nynorskkorpus	nynorsk	107,8 mill.	1866-2012	blanda	ja	avgrensa
NBs frie tekstar (nynorsk)	nynorsk	46,2 mill.	1850-2010	blanda		open
Aviskorpus (nynorsk)	nynorsk	16,1 mill.	1998-2015	sakprosa (avis)		open

Vi har gjort eit utval som er lagd på ein dedikert server for leksikografane. For eksempel kan vi gjøre komplekse søk i Nynorskkorpuset, Leksikografisk Bokmålskorpus og Nasjonalbibliotekets frie tekster samtidig.

UTVAL:

For prosjektet vårt er det viktig med:

- Stort materiale frå ulike domene, (jf.kolonna Sjanger & Storleik)
- som dekkjer eit breitt tidsaspekt (jf. kolonna Tidsrom)
- kvalitetssikra materiale (mest mogleg) – utval av korpus
- Språkleg analysert materiale (jf kolonna lemma + ordklasse)
- Fordeling bm-nn: ca 90%-10%
- Norwegian Nynorsk (nno): 170 mill. ord til saman
- Norwegian Bokmål (nob): 2,2 milliard
- TIL SAMAN: 2,4 milliard ord

(vi skal vise døme på at balansen i dette korpuset kanskje ikkje er optimalt - bias)

FUNKSJONALITET

På vår eigen server kan vi søkje i fleire korpus på ein gong.

Verdi: vi sparar enormt med tid.

Treffa kan vi deretter sortere ut frå attributt som kva kjelde treffa kom frå, eller kva tidsrom treffa er fordelt over.


UNIVERSITETET I BERGEN


Lemmutval: trunkerte søk

Søk | Refine | window: 5 tokens | Stop | Lagrede søk ... | Lagre søket | som

Done. Running time: 0.04 sec. (0.05 CPU sec.)

Treff 1 - 30 av 6693 | Forrige | Neste | Gå til: | Last ned (Excel-modus) | Copy query URL | Vis linjefilter | Attributter ... | Linjer per side: | Kontekststørrelse:

corpus	treff	year
avis-nno	en mellom Voss og Medkila i Prestegardsmoen i dag. Ballen	2008
avis-nno	azeem Ahmed det meste sjølv då han skar inn frå høgre og	2011
avis-nno	azeem Ahmed det meste sjølv då han skar inn frå høgre og	2011
avis-nno	azeem Ahmed det meste sjølv då han skar inn frå høgre og	2011
avis-nno	azeem Ahmed det meste sjølv då han skar inn frå høgre og	2011
avis-nno	I sjå ut. TV2 har kjøpt OL-rettar. Altså risikerer vi i framtida	2011
avis-nno	istoria bak fotball, tennis, basketball, bordtennis, volleyball,	2015
avis-nno	bildeCredit> FOTO: ARNE S. GJONE ¶ Les også En skole for	2005
avis-nno	ire tiåringane har det, går det ut på eitt. Nokre har kalla dei	2006
avis-nno	r- Eirik Kval. Unge menneske i dag er ikkje kalla glasur- og	2015
avis-nno	får dei unge mange kallenamn, som «lydig», «finkis» og «	2014
avis-nno	274481.ece> < F ¶ ~ Norsk seier over Japan ¶ De norske	2006
avis-nno	leklasse, og at det på havets botn ligg ein tusen år gammal	2015
avis-plain	ui " på Hawaii, Paramount går snart i gang med " Girl in the	2001
avis-plain	og mars: ¶ Øvelse Sett repetisjoner ¶ Knebøy: 3 x 15 ¶ Leg	2006
avis-plain	ai og juni ¶ Øvelse Sett repetisjoner ¶ Knebøy: 3 x 10 ¶ Leg	2006
avis-plain	smith 3-4 x 15 (på hver fot) ¶ Leg extension 3-4 x 15 ¶ Leg	2006
avis-plain	Smith 3-4 x 10 (på hver fot) ¶ Leg extension 3-4 x 10 ¶ Leg	2006
avis-plain	Nordby synes OL-isen er rask. - Det er ikke så mye skru ("	2006
avis-plain) jorbanna på isen og forholdene. Vi er et lag som liker mye	2006

Kjelde for dette søket: Aviskorpus bm + nn i Copurscle-lex

Enkelt trunkert søk på "curl."

Dette er et eit søk vi ofte gjer for å sjekke om vi har fått med sentrale ord innanfor eit bokstavstrekk..

Søket vårt her:

Drøyt 6700 treff, der konkordansen gjev deg ordet i kontekst. Du kan velje ulike kolonner med informasjon knytta til kvart treff, t.d. kva år kjelda er frå (kolonna til høgre)

Framfor å sjå på konkordansane, kan ein også t.d. hente ut ordlister over kva ordformer som finst i treffa [klikk]

UNIVERSITETET I BERGEN

Lemmutval (forts.): ordlister

Documentation
FAQ
Links

Korpusliste

Søk
Konkordans
Kollokasjoner
Distribusjon
Ordlister
Diagram
Tekst
Metadata
Oversikt
Variabler
Korpusdok.

Ordbøker
Ordbank

"curl.*" Søk | Refine | window: 5 tokens | Stop | Lagrede søk ... | Done. Running time: 0.04 sec. (0.05 CPU sec.)

Antall treff: 6693, unique values: 429. Attributt: word | ignorerer storskriving | sorter: etter frekvens | relative to: -

Side 1 av 2. Previous Next

1514 (22,62%)	curling	21 (0,31%)	curlingpresident	7 (0,10%)	curling-	4 (0,06%)	curling-sporten
913 (13,64%)	curling-VM	20 (0,30%)	curlinglandslag	7 (0,10%)	curling-gull	4 (0,06%)	curlingeksperten
439 (6,56%)	curlinggutta	20 (0,30%)	curlingspilleren	7 (0,10%)	curlingbarn	4 (0,06%)	curlingfest
301 (4,50%)	curler	19 (0,28%)	curlingguttene	7 (0,10%)	curlingbarna	4 (0,06%)	curlingforbund
204 (3,05%)	curling-EM	18 (0,27%)	curlingdamene	7 (0,10%)	curlingdamer	4 (0,06%)	curlinggenerasjon
159 (2,38%)	curlingjentene	18 (0,27%)	curlingsteiner	7 (0,10%)	curlingforelder	4 (0,06%)	curlinginnsats
150 (2,24%)	curlinglaget	16 (0,24%)	curlingforbundets	7 (0,10%)	curlingkarrieren	4 (0,06%)	curlingklubben
129 (1,93%)	curlinglandslaget	16 (0,24%)	curlingklovnene	7 (0,10%)	curlingkolleger	4 (0,06%)	curlingkunstner
121 (1,81%)	curlinghallen	15 (0,22%)	curling-gutta	7 (0,10%)	curlinglagene	4 (0,06%)	curlinglegenden
116 (1,73%)	curlingherrene	15 (0,22%)	curlingklubb	7 (0,10%)	curlingnasjon	4 (0,06%)	curlingmamma
113 (1,69%)	curlinglag	14 (0,21%)	curlingboksene	6 (0,09%)	curling-finalen	4 (0,06%)	curlingmesterskap
112 (1,67%)	curler	14 (0,21%)	curlinggull	6 (0,09%)	curlingdronning	4 (0,06%)	curlingnasjonen
111 (1,66%)	curler	14 (0,21%)	curlingvinner	6 (0,09%)	curlingfans	4 (0,06%)	curlingproff
99 (1,48%)	curlingmillioner	14 (0,21%)	curlingmiljøet	6 (0,09%)	curlingfeberen	4 (0,06%)	curlingsenter
71 (1,06%)	curlingforbundet	14 (0,21%)	curlingseier	6 (0,09%)	curlingfolket	4 (0,06%)	curlingsjef
65 (0,97%)	curlingbanen	13 (0,19%)	curlere	6 (0,09%)	curlinghistorie	4 (0,06%)	curlingskipen
62 (0,93%)	curlingguttas	12 (0,18%)	curla	6 (0,09%)	curlinginteressene	4 (0,06%)	curlingsport
60 (0,90%)	curlingforeldre	12 (0,18%)	curlingbukser	6 (0,09%)	curlingjenter	4 (0,06%)	curlingsportens
54 (0,81%)	curling-seirer	12 (0,18%)	curlingfeber	6 (0,09%)	curlingkampene	4 (0,06%)	curlingsten
54 (0,81%)	curlingherrer	12 (0,18%)	curlingforeldrene	6 (0,09%)	curlingkonkurranse	3 (0,04%)	curlers
53 (0,79%)	curlinghall	12 (0,18%)	curlinglandslaget	6 (0,09%)	curlingleksjon	3 (0,04%)	curling-Norge
47 (0,70%)	curlingbane	12 (0,18%)	curlingmenn	6 (0,09%)	curls	3 (0,04%)	curling-aale

(ved å klikke på «ordliste» til venstre i menyen får du ut Kva visar ei slik liste oss?)

1. Peikar ut gode kandidatlar til oppslagsformer som bør vurderast med i ordboka, om dei ikkje allereie er med. (men seinare skal vi sjå døme på at høg frekvens ikkje fortel oss alt)
2. Slike lister viser noko om i kva grad eit ord er produktivt i samansetjingar. Kanskje er samansetjingane kandidatlar til å kome med, og det fortel uansett noko om i kva grad grunnordet er sentralt i ordforrådet vårt.



Distribusjon (t.d. over tid)



Korpuskel :: Aviskorpus (Bokmål) :: Distribusjon

eng | nob | Logg ut (Gunn Inger Lyse Samdal (gunn)) | Edit is av

- Hjem
- Komme i gang
- Dokumentasjon
- FAQ
- Publikasjoner
- Lenker
- Korpusliste
- Søk
- Konkordans
- Kollokasjoner
- Distribusjon**
- Ordlister
- Tekst
- Metadata
- Oversikt
- Variabler
- Korpusdok.
- Lokalisering

Avansert søk | bytt | Enkelt søk | Søkeshistorie ...

"b.nðbre..7ðð.7.7.7"

Søk | Refiner | vindu: 5 tokens | Stop | Lagrede søk ... | Lagresøket som felles

Done, Running time: 0.01 sec. (0.01 CPU sec.)

Vis distribusjon type: absolutt | counts only | include structures

av word
 relative to year
 grupper etter word
 og source
 data

Ignorer storskiving, Δ: 0 | filter:
 Ignorer storskiving, Δ: 0 | filter:
 Ignorer storskiving, Δ: 0 | filter:
 Ignorer storskiving, Δ: 0 | filter:

Fractions sum up to 1.0 in each row. Fractions in blue are unweighted means of group fractions. Fractions in green are distributions of total numbers.

Side 1/1 av 1 x 1. | Last ned

(sum)	bandbrede	båndbrede	båndbredden	båndbreddene	båndbredder
(sum) 756 (100,0)	1 (0,1)	521 (68,9)	214 (28,3)	1 (0,1)	19 (2,5)
756 (100,0)	1 (0,1)	521 (67,5)	214 (28,7)	1 (0,5)	19 (2,1)
1998 5 (100,0)		3 (50,0)	2 (40,0)		
1999 24 (100,0)		12 (50,0)	11 (45,8)		1 (4,2)
2000 45 (100,0)		36 (80,0)	7 (15,6)		2 (4,4)
2001 74 (100,0)		59 (79,7)	15 (20,3)		
2002 19 (100,0)		14 (73,7)	5 (26,3)		
2003 21 (100,0)		17 (81,0)	4 (19,0)		
2004 11 (100,0)		6 (54,5)	4 (36,4)	1 (9,1)	
2005 52 (100,0)	1 (1,9)	36 (69,2)	14 (26,9)		1 (1,9)
2006 71 (100,0)		38 (53,5)	31 (43,7)		2 (2,8)
2007 45 (100,0)		31 (68,9)	14 (31,1)		
2008 82 (100,0)		66 (80,5)	14 (17,1)		2 (2,4)
2009 53 (100,0)		35 (66,0)	18 (34,0)		
2010 47 (100,0)		29 (61,7)	17 (36,2)		1 (2,1)
2011 31 (100,0)		25 (80,6)	6 (19,4)		
2012 46 (100,0)		29 (63,0)	9 (19,6)		8 (17,4)
2013 71 (100,0)		49 (69,0)	20 (28,2)		2 (2,8)
2014 45 (100,0)		27 (60,0)	18 (40,0)		
2015 14 (100,0)		9 (64,3)	5 (35,7)		

23.05.19



UNIVERSITETET I BERGEN

Kollokasjonar: bruksdøme, underoppslag

Documentation
FAQ
Links

[lemma="hatt" & morph = ("subst")]

Søk Refine window: 5 tokens
Lagre søket som Done. Running time: 32.39 sec.

Vis kollokasjonar by word, venstrekontekst: 2, høyrekontekst: 2, combine context
MI * log(Freq) Last ned

11708 collocations calculated; page 1 of 235. Forrige Neste Show concordance for selection show freq. show MI show LL



Freq.	Delta	Collocate
74	-2	fjør _ hatten
1	2	<< hatt >> _ halitt
1	-2	defflasjonskrise _ Hattene
2	-2	høysåter _ Hattane
1	-1	SPARKA HATTEN
1	-1	DISSE HATTENE
48	-2	fjær _ hatten
4	-1	bredbremmete hatter
5	-1	bredbremmede hatter
2	-1	tresnutede hatter
2	-1	vidbremmede hatter
1	-2	HØY _ HATTEN
1	-1	trinsnuta hattane
1	2	hattane _ utfrunsa
1	-2	stiletthæl _ Hattane
7	-1	bredbremmete hatten
2	-1	SIN HATT

Kjelde for dette søket: LBK + nynorskkorpus i Copurscle-lex

Kollokasjonar er ord som har ein statistisk tendens til å bli funne saman. Corpuscle har eit knippe statistiske mål; fra veldig enkle (rein frekvens, relativ frekvens), til meir kompliserte statistiske mål.

Kollokasjonar er ofte til hjelp for å identifisere typiske bruksdøme i ordboka, men óg moglege underoppslag

Øvst: Fjør i hatten, nede på linje 7 fjær i hatten.


UNIVERSITETET I BERGEN


Søketemplat i Corpuscle

eng | nob | Sign out (Gunn Inger Lyse Samdal (gunn)) | Edit is off

CLARINO Corpuscle-Lex :: Aviskorpuss ann., Aviskorpuss (Nynorsk), Aviskorpuss (Bokmål), Forskning.no (2017), Leksikalsk bokmålskorpuss, NBs frie tekster (Nynorsk), NBs frie tekster (Bokmål), Nynorsk-korpuss, Talk Of Norway :: Query

Corpuscle Home
Documentation
FAQ
Links

Corpus list

Query

Concordance
Collocations
Distribution
Word List
Diagram
Text
Metadata
Overview
Variables
Corpus doc.

Ordbøker
Ordbøkerdigering
Ordbank

Advanced search | [switch to Basic search](#)

antikvarisk.*" [[]

Run Query | Reset | Build graphical query | Show query expression | **Saved queries** | Save Query as [] public

Saved queries for Aviskorpuss ann., Aviskorpuss (Nynorsk), Aviskorpuss (Bokmål), Forskning.no (2017), Leksikalsk bokmålskorpuss, NBs frie tekster (Nynorsk), NBs frie tekster (Bokmål), Nynorsk-korpuss, Talk Of Norway

Page 1 of 2 | [Previous](#) | [Next](#) | Go to page: [] | [Go](#)

Filter by name: [] | by query string: [] | public private

Click on a name to choose the query. (24 stored queries shown.) Click on a query to see its documentation.

Name	Query
* analysert: lemma-og-ordklasse_avis_forskning.no_LBK_Nynorskorpuset	[lemma="rett" & morph="(subst)"]
* analysert: lemma-og-ordklasse_avis_talkofnorway	[lemma="rett" & pos="subst"]
* analysert: lemmasøk (enkelt)	/hest/
* analysert: vilkårlig lemma-av bestemte-ordklasser + lemma (søk i: lbk-nnk-fn-new-avis)	[morph = ("adv" "adj")] [lemma = "cowboy" %c]
* streng: delstreng + [0 - 1 tegn]	"bygg.7e"
* streng: delstreng + [0 - 1 tegn] case-sensitiv	"bygg.7e" %c
* streng: delstreng + [0 - 4 tegn]	"croissant{.0,4}" [Delete]

Forklaring:
 "(.0,m)" gir mellom n og m tegn etter hverandre i en streng.
Bruk:
 Kan stå hvor som helst i strengen. F.eks. "bl{.0,2}sverm" gir *blsverm*, *blsverm*, *bl-sverm*, *bls-verm*.
Nyttig for:
 Finne variasjon i bøyingsmåter/stavemåter (Leks. flertall av "croissant": *croissant(s)*, *croissant(er)*, *croissant(ene)*, *croissant(en)* osv.
Alternativer/Ekvivalent med:
 ? "croissant.7.7.7.7" (0 til 4 tegn ordinært) eller "bl.7.7sverm" (0 til 2 tegn inni strengen)
 .+ "bl-sverm" (vil gi minst ett el. uendelig mange tegn inni ordet)

31 templer lagra



Frekvens fortel ikkje alt

- Ein høg frekvens er ikkje åleine nok:
 - Korleis er ordet distribuert over tid?
 - Korleis er ordet distribuert på tvers av domene?
- Døme: søk på "aor.*" i Aviskorpuset (bokmål)



For å grunngje at eit ord er "sentralt" i vokabularet vårt, er det ikkje nok i seg sjølv med høg frekvens.

Distribusjon (over tid, ulike domene)

To døme[klikk]

UNIVERSITETET I BERGEN

eng | nob | Logg ut (Gunn Inger Lyse Samdal (gunn))

CLARINO Korpuskel :: Aviskorpus (Bokmål) :: Ordliste

Avansert søk | bytt til Enkelt søk Query history ...

"aor:*" window: 5 tokens |
 felles
 Done. Running time: 0.03 sec. (0.03 CPU)

Antall treff: 113, unique values: 21. Attributt: | ignorer storskriving | sorter: etter frekvens
 Last ned include counts include fractions
 Side 1 av 1.

58 (51,33%)	aortastenose
22 (19,47%)	aorta
8 (7,08%)	aortaklaffen
4 (3,54%)	aorta-ventiler
3 (2,65%)	aorta-disseksjon
2 (1,77%)	aortaklaff
2 (1,77%)	aortic
1 (0,88%)	aorakelsteiner
1 (0,88%)	aorhhh
1 (0,88%)	aorin
1 (0,88%)	aorta-aneurisme
1 (0,88%)	aorta-blodåren
1 (0,88%)	aortaaneurisme
1 (0,88%)	aortabuen
1 (0,88%)	aortaklaffene
1 (0,88%)	aortaklaffeprotese
1 (0,88%)	aortaklaffer

Hjem
 Komme i gang
 Dokumentasjon
 FAQ
 Publikasjoner
 Lenker

Korpusliste

Søk
 Konkordans
 Kollokasjoner
 Distribusjon
 Ordliste
 Tekst
 Metadata
 Oversikt
 Variabler
 Korpusdok.

Lokalisering
 23.05.19

”aortastenose” ser kjempefrekvent ut! 58 treff, dobbelt så mange treff som ”aorta” (som er i ordbøkene våre)
 Men viss du ser på distribusjonen til ”aortastenose” over tid (eller nøyare på kollokasjonane)
 [klikk]

UNIVERSITETET I BERGEN

Vis distribusjon type: absolutt | counts only | include structures

av word | ignorer storskriving, Δ: 0 | filter:

relative to year | ignorer storskriving, Δ: 0 | filter:

grupper etter - | ignorer storskriving, Δ: 0 | filter:


og - | ignorer storskriving, Δ: 0 | filter:

Fractions sum up to 1.0 in each row. Fractions in blue are unweighted means of group frac

Side 1/1 av 1x1. | Last ned

	(sum)	aortastenose
(sum)	58 (100,0)	58 (100,0)
2005	57 (100,0)	57 (100,0)
2014	1 (100,0)	1 (100,0)

23.05.19 SIDE 18



...så ser du at så å seie alle treffa er frå 2005, faktisk frå mars/april 2005, då kong Harald vart innlagd på sjukehus med hjerteklaffproblem (=aortastenose)

Eit liknande døme er orda knytt til "askeskya" frå Island i 2010.

Corpuscle-Lex :: Aviskorpus (Nynorsk), Aviskorpus (Bokmål), Forskning.no (2017), Leksikalsk bokmålskorpus, NBs frie tekster (Nynorsk), NBs frie tekster (Bokmål), Nynorsk-korpus, Talk Of Norway :: Distribusjon

Avansert søk | bytt til Enkelt søk | Query history ...

"assistenttren[a]e]r"

Søk Refine window: 5 tokens | Stop | Lagrede søk
 ... | Lagre søket som felles
 Done. Running time: 0.21 sec. (0.34 CPU sec.)

Vis distribusjon type: absolutt | counts only | include structures

av word | ignorer storskriving, Δ: 0 filter:
 relative to corpus | ignorer storskriving, Δ: 0 filter:
 grupper etter - | ignorer storskriving, Δ: 0 filter:
 og - | ignorer storskriving, Δ: 0 filter:

Fractions sum up to 1.0 in each row. Fractions in blue are unweighted means of group fractions. Fractions in green are distributions of total numbers.

Side 1/1 av 1x1. | Last ned

(sum)	assistenttrenar	assistenttrener
(sum)	13264 (100,0)	163 (1,2) 13101 (98,8)
	13264 (100,0)	163 (31,3) 13101 (68,7)
avis-nno	29 (100,0)	29 (100,0)
avis-plain	12977 (100,0)	12977 (100,0)
fn-new	2 (100,0)	2 (100,0)
lbk	18 (100,0)	18 (100,0)
nink	238 (100,0)	134 (56,3) 104 (43,7)

Assistenttrenar/-trener

Hvis et ord forekommer mange ganger kun i eitt kildedokument eller kun i aviskorpuset, så viser dette kun at ordet er frekvent i eit domene.



INESS og NorGramBank: søk og analyse (setningsnivå)

- **Trebank:** eit syntaktisk analysert tekstkorpus der kvar setning har ein detaljert syntaktisk analyse.
- **INESS:** ein infrastruktur for å bevare og gjere trebankar tilgjengelege (søk og analyse). Del av CLARINO Bergen Centre.
- **NorGramBank:** ein trebank for norsk, utvikla i prosjektet INESS (2010-2017).



Trebank: Eit syntaktisk analysert tekstkorpus der kvar setning er forsynt med ei detaljert syntaktisk analyse.

INESS:

NorGramBank: syntaktisk analyse med NorGram, ein komputasjonell grammatikk på bm og nn.

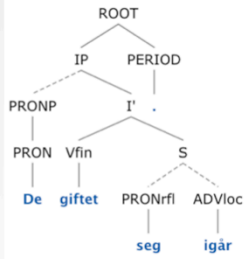
80 mill. ord (analysert tekst), eit subsett av det vi kan søkje på i corpuscle.

Fordelt på skjønnlitteratur, sakprosa og stortingstekst.

Bm-nn ratio: ca 90-10.



C-structure



F-structure

PRED	'gifte*seg<[8:de]>[4:seg]'
TNS-ASP	10 TENSE past, MOOD indicative
TOPIC	PRED 'de' NTTYPE 9 NSYN pronoun REF +, PRON-TYPE pers, PRON-FORM de, PERS 3, NUM pl, DEF +, CASE nom
ADJUNCT	8 1 { PRED 'i går' } 2 ADV-TYPE temp
OBJ	PRED 'seg' NTTYPE 7 NSYN pronoun 4 REF -, PRON-TYPE refl, PERS 3, NUM pl, CASE obl
SUBJ	[8]
VTYP	0 VTYP main, VFORM fin, STMT-TYPE decl





Søkjtemplat i INESS

Page 1 of 2 | [Previous](#) | [Next](#) | Go to page: | [Go](#) | public private

Filter by name: | by description: | by template:

Click on a name to choose a template. (31 stored templates shown.) | Show template expansion

[Select marked templates](#) | [Store marked templates as sketch ...](#)

Name	Description
<input type="checkbox"/> * ADJ-coord(@ADJ)	Adjectives coordinated with an adjective
<input type="checkbox"/> * ADJ-degreeadvs(@ADJ)	Degree adverbs modifying an adjective
<input type="checkbox"/> * ADJ-modifies(@ADJ)	Nouns modified by an adjective
<input type="checkbox"/> * ADJ-modnominadj(@ADJ)	Adjectives modifying a nominal head adjective
<input type="checkbox"/> * ADJ-suff(@SUFF)	Adjectives derived with a suffix
<input type="checkbox"/> * ADV-degmodifies(@ADV)	Adjectives modified by a degree adverb
<input type="checkbox"/> * ADV-types(@ADV)	The types of an adverb
<input type="checkbox"/> * N-adjmod(@N)	The adjectives modifying a noun
<input type="checkbox"/> * N-defmascorfem(@N)	Feminine vs. masculine inflection of a noun



31 templer lagra



- Main Page
- Knowledge center
- The project
- Documentation
- FAQ
- Publications
- Links
- Resources

Treebanks

- Treebank Selection
- Treebank Details
- Sentence Overview
- Sentence
- Query
- Sketch

Search in: nob-avis, nob-child, nob-fn, nob-lbk-av, nob-lbk-sa, nob-lbk-tv, nob-ndt-lfg, nob-novel, nob-novel_1, nob-novel_2, nob-novel_3, nob-novel_4, nob-novel_5, nob-novel_6, nob-novel_7, nob-sofie
max #: | fragments: none only | fully disamb.: none only
disambiguated: none only | unambiguous: none only

[Select query templates ...](#) | [Select sketch ...](#) | [Query history ...](#)

Template: * V-argframes(@V)

Description: Argument frames of a verb

Parameters:

@V:





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

UNIVERSITETET I BERGEN



Count	#a: atom	#arg1: value	#arg2: value	#arg3: value
1545	gifte*seg	pronoun		
866	gifte*seg*med	pronoun	pronoun	
409	gifte*seg*med	pronoun	common	
346	gifte*seg*med	pronoun	proper	
201	gifte*seg	common		
128	gifte			
119	gifte*seg	proper		
96	gifte*seg*med	common	common	
63	gifte*seg*med	common	pronoun	
59	gifte*seg*med	proper	common	
54	gifte*seg*med	proper	proper	
52	gifte*seg*med	proper	pronoun	
26	gifte*seg*med	common	proper	
25	gifte*bort		pronoun	
20	gifte*bort	pronoun	pronoun	
18	gifte*bort		common	
15	gifte*seg*til	pronoun	common	
10	gifte*bort	pronoun	common	
8	gifte*bort	common	pronoun	
5	gifte*bort	common	common	
5	gifte*seg*til	pronoun	pronoun	
4	gifte*bort		proper	
3	gifte*seg*til	common	common	
2	gifte*bort	proper	pronoun	
1	gifte*bort	pronoun	proper	
1	gifte*seg*til	common	proper	

23.05.19

SIDE 24



Count	#a: atom	#arg1: value	#arg2: value	#arg3: value
1545	gifte*seg	pronoun	UNIVERSITETET I BERGEN	
866	gifte*seg*med	pronoun	pronoun	
409	gifte*seg*med	pronoun	common	
346	gifte*seg*med	pronoun	proper	
201	gifte*seg	common		

Page 1 of 11 | [Previous](#) | [Next](#) | Go to page: [Go](#) | [Download](#)


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

Treebank	Document	Trans.	Id	Sentence	
nob-novel_7	oai:bibsys.no:biblio...	no	1739	En mann som vil gifte seg, er også i stand til slike handlinger, akkurat som sine forfedre.	Copy
nob-novel_7	oai:bibsys.no:biblio...	no	2135	- Saken er den at mor har tenkt å gifte seg!	Copy
nob-novel_7	oai:bibsys.no:biblio...	no	2194	Hun er en kvinne som gifter seg, akkurat som andre kvinner gifter seg hver dag, hver time på dagen!	Copy
nob-novel_7	oai:bibsys.no:biblio...	no	2416	Det var ikke noe galt i at en «kvinne» giftet seg igjen etter skilsmissen.	Copy
nob-novel_7	oai:bibsys.no:biblio...	no	2620	- Far giftet seg da han var på min alder.	Copy
nob-novel_7	oai:bibsys.no:biblio...	no	5153	Hvis du vil, skal jeg gi deg flere titalls eksempler på yngre søstre som har giftet seg før eldre, og det har ikke vært til hinder for at de eldste giftet seg med de beste ektemennene som tenkes kan.	Copy
nob-novel_7	oai:bibsys.no:biblio...	no	5892	Tanken på at datteren skulle gifte seg, gav ham en merkelig, ubehagelig følelse, enda det stred mot både fornuft og moral.	Copy


23.05.19

SIDE 25






UNIVERSITETET I BERGEN



Nytteverdi INESS

Særleg nyttig for

- funksjonsord (preposisjonar, adverb..)
- ord med mange syntaktiske samband (*gifte seg, falle i auge/få auge på/gjere store auge/ha auge for...*)
- ord med systematiske skilnader knytt til syntaktiske relasjonar
 - partisippformer og adjektiv



23.05.19 SIDE 26

Nå har Helge laget et søketemplat ved navn * P-prepobjpred(@PREP). Dette returnerer en oversikt over alle objekter som står til ulike preposisjoner. Hvis jeg for eksempel søker på nedover, gir den tilbake toppresultatene kinn, rygg og gate (se skjermdump). Dette kan være nyttig for å finne de mest prototypiske eksemplene på bruk av ulike preposisjoner.

Det jeg vil er å se på en bestemt partisipp, f.eks. «anvendt», og jeg vil se på:

1. hvor mange ganger er den samsvarsbøyd som predikativ? (vs. hvor mange ganger er den ikke det)
2. hvor mange ganger står den i attributiv stilling (og hvor ofte ikke)

Jf. hypotesen du skisserte i en epost i etterkant av fredagsseminaret (som Margunn og Mikkel for øvrig roste! Jeg kunne dessverre ikke være med og høre)

En hypotese er da at høye verdier for (1) Kongruensbøyning og Attributiv bruk markerer gode kandidater for leksikaliserte adjektiver i ordboken: De har to sentrale adjektiviske egenskaper.

(‘de er avanserte’, ‘en avansert pianist’)

(kanskje ikke alltid samsvarbøyning, men forekommer mest i attributiv stilling: ‘en etterlatt koffert’, ‘de ble etterlatt’)



Oppsummering

Korpus er til god hjelp for leksikografen når det gjeld:

- lemmatilfang (ord inn og ord ut av ordboka);
 - underoppslag;
 - ordtydingar;
 - syntaktisk åtferd;
 - bruksdøme;
-
- ..Men korpus må brukast klokt, og kan ikkje erstatte leksikografiske vurderingar
 - Behov for vidareutvikling av tekniske løysingar







UNIVERSITETET I BERGEN

