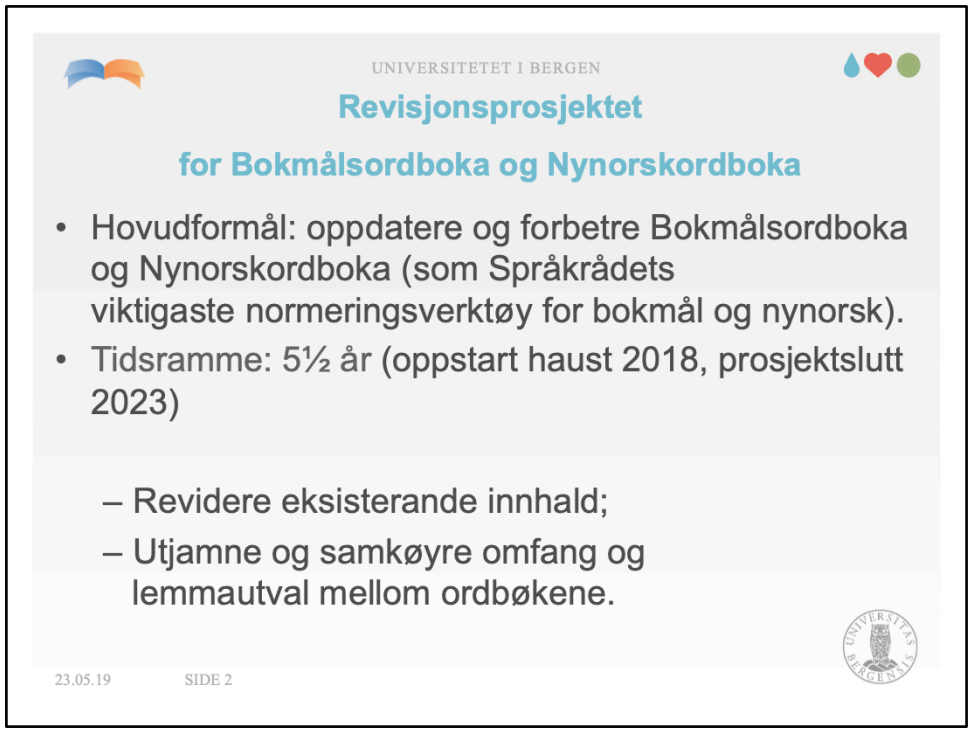
Siteringsforslag: ”Frå bandsalat til bandbreidd: leksikografiske metodar i Revisjonsprosjektet for Bokmålsordboka og Nynorskordboka” (2019). Revisjonsprosjektet for Bokmålsordboka og Nynorskordboka. Universitetet i Bergen. Presentasjon gjeve av Gunn Inger Lyse på faglunsj ved LLE, Universitetet i Bergen,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

Kontakt: ordbokene@uib.no.

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.



Utg.punkt er kort sagt:
I dette prosjektet skal leksikografene revidere to eksisterende ordbøker parallelt, én på bokmål og én på nynorsk.

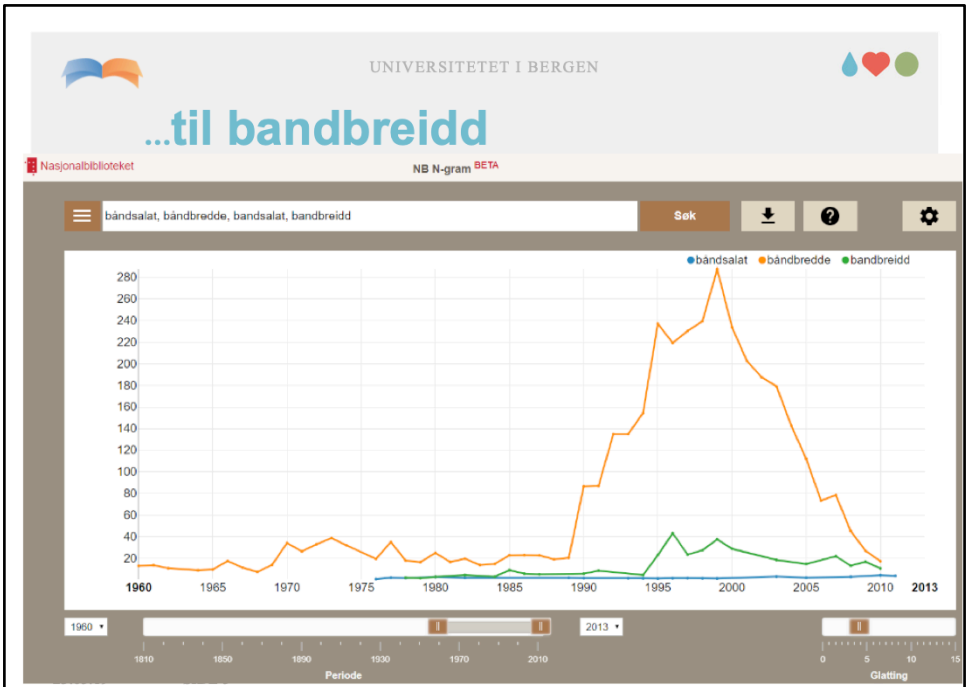
(3 siste kulepunkt)
Revidere eksisterande innhald (tydingsinndeling, definisjonar, etymologi, uttale);

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:

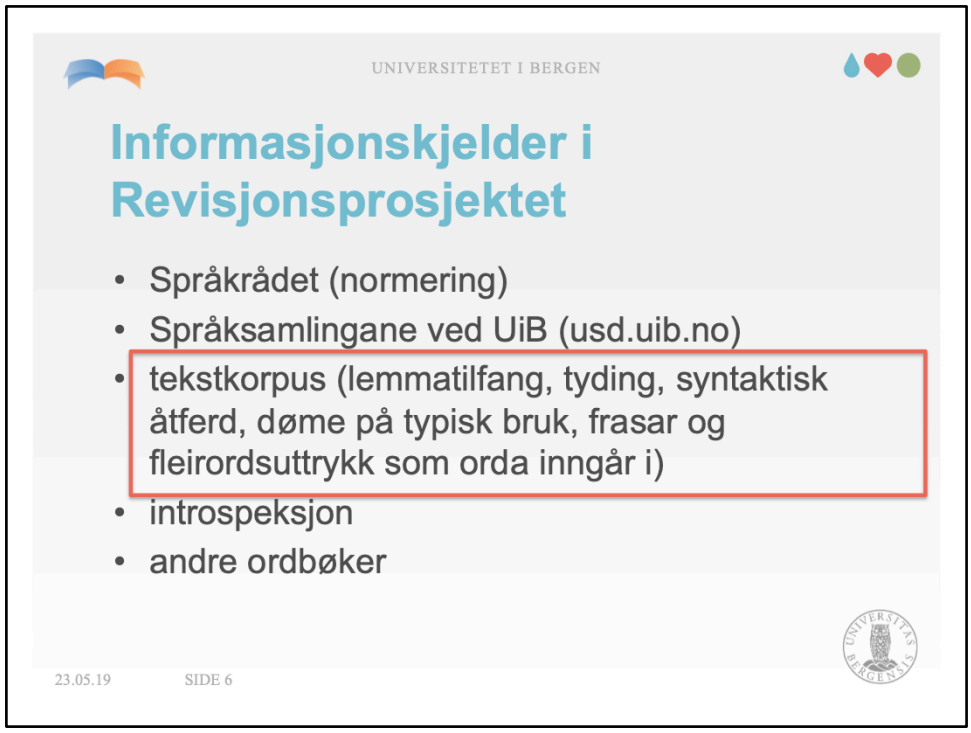
Bandsalat låg kun i BOB, men ikkje i NOB.
Bandsalat: som fenomen tona meir og meir ut då kassettspelaren vart erstatta av meir moderne måtar for å spele av musikk

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



Skjermbilete frå Nasjonalbibliotekets n-gram:
https://www.nb.no/sp_tjenester/beta/ngram_1/#ngram/

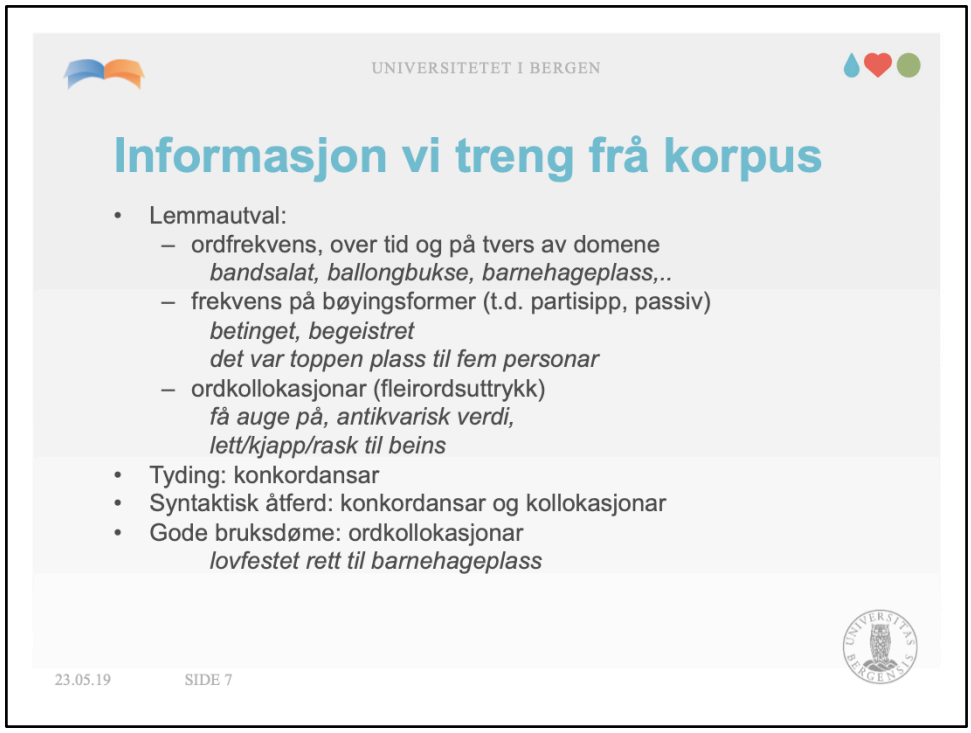
..bandbreidd.
Bandbreidd låg verken i BOB el NOB då prosjektet vårt starta opp.



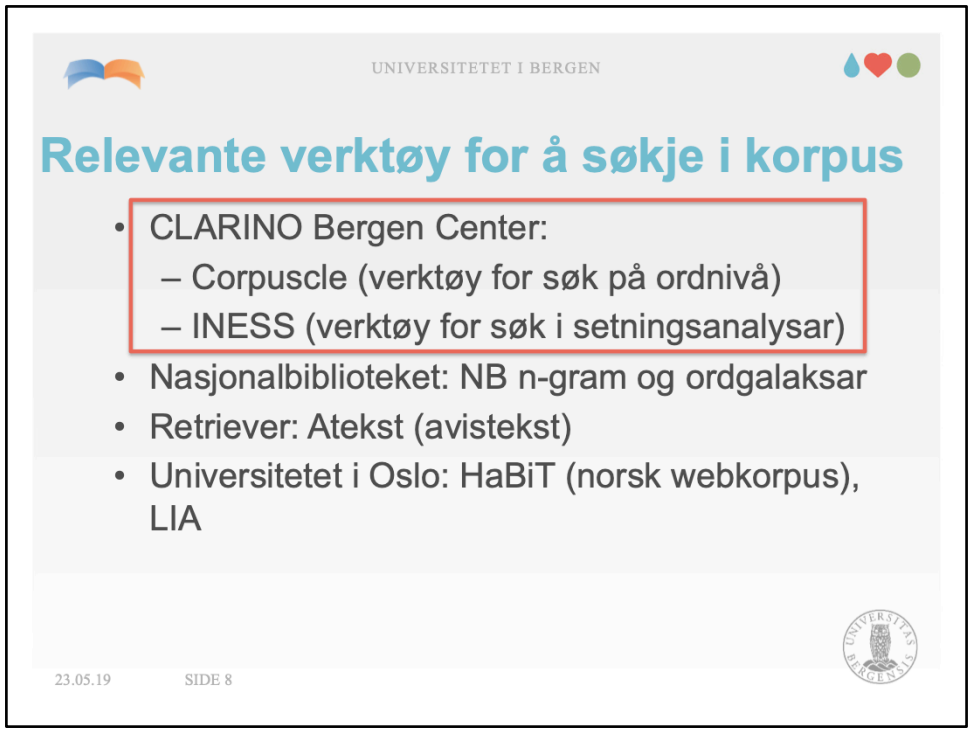
Den empiriske forankringa i **leksikografiarbeidet er viktig.**
Vi brukar dagleg det vi kallar ”korpus”:
-Tekstkorpus −kolleksjon av elektronisk tekst som er søkbar, og der du óg kan analysere treffa.
Lemmautval: peike ut eit sentralt ordforråd
Identifisere i kva tydingar eit ord vert brukt (”skiballett” 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 **ordbokarbeidet ikkje kviler åleine på å sjå på data**. (liste)

-Språkrådet (slår opp i normeringsvedtak - normerer bøying, 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;



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



Prosjektet vårt har tilgong til forkjellig 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 etteretteleg + 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.

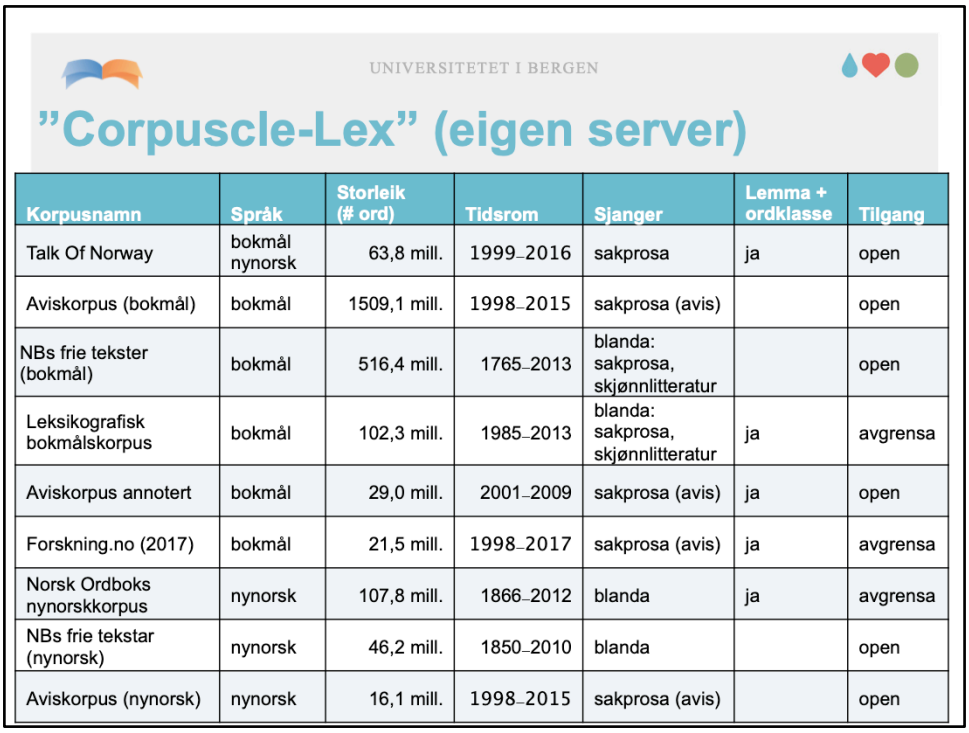
Corpuscle:

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 tekstkorpsus 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]



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:

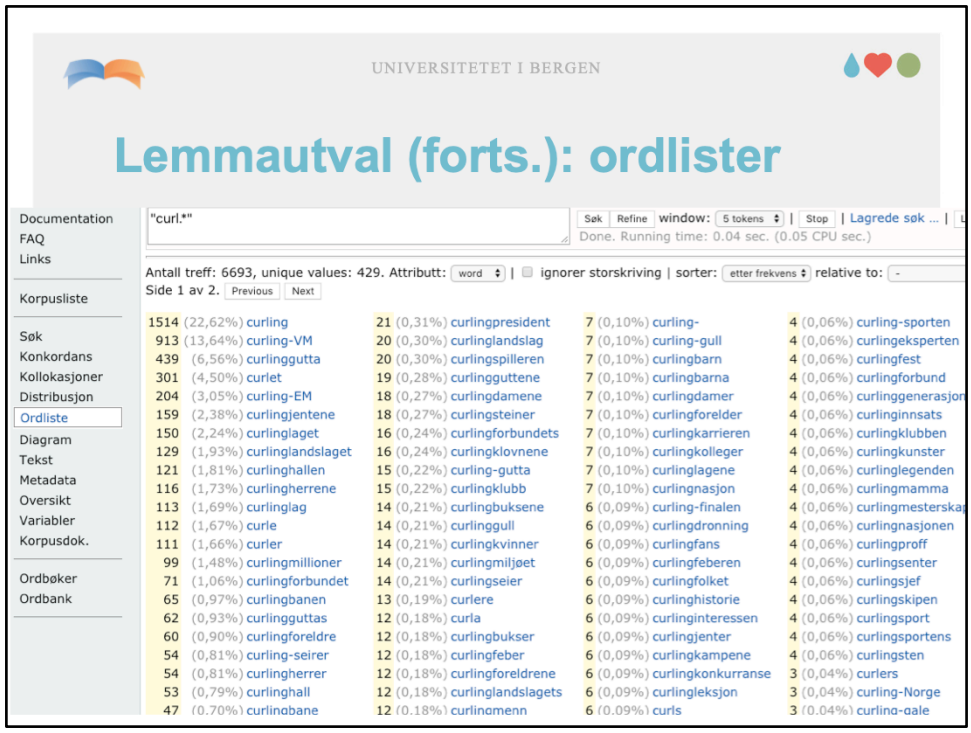
- Stort materiale frå ulike domene, (jf.kolonna Sjanger & Storleik)
- som dekkjer eit breitt tidsaspekt (jf. kolonna Tidsrom)
- kvalitetssikra materiale (mest mogleg) – utval av korpus
- Språkleg analysert materiale (jf kolonna lemma + ordklasse)
- Fordeling bm-nn: ca 90%-10%
- Norwegian Nynorsk (nno): 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.

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]

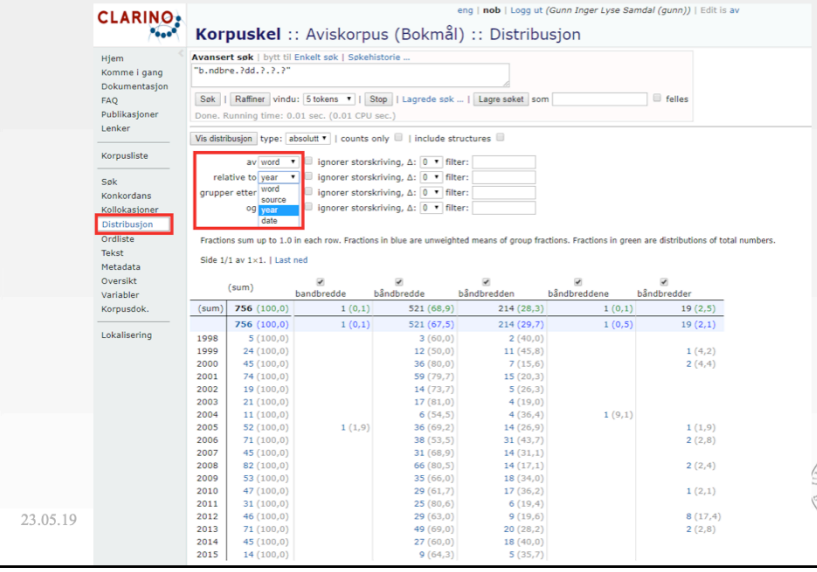


(ved å klikke på «ordliste» til venstre i menyen får du ut Kva visar ei slik liste oss?

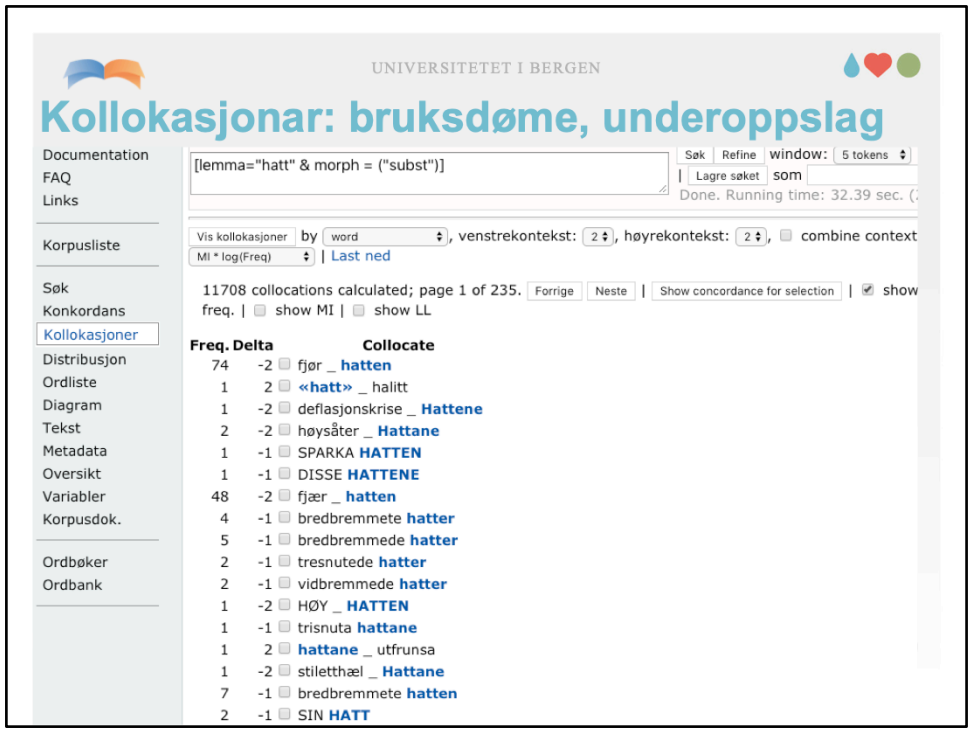
1. Peikar ut gode kandidatar til oppslagsformer som bør vurderast med i ordboka, om dei ikkje allereie er med. (men seinare skal vi sjå døme på at høg frekvens ikkje fortel oss alt)
2. Slike lister viser noko om i kva grad eit ord er produktivt i samansetjingar. Kanskje er samansetjingane kandidatar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

UNIVERSITETET I BERGEN

Distribusjon (t.d. over tid)



23.05.19



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

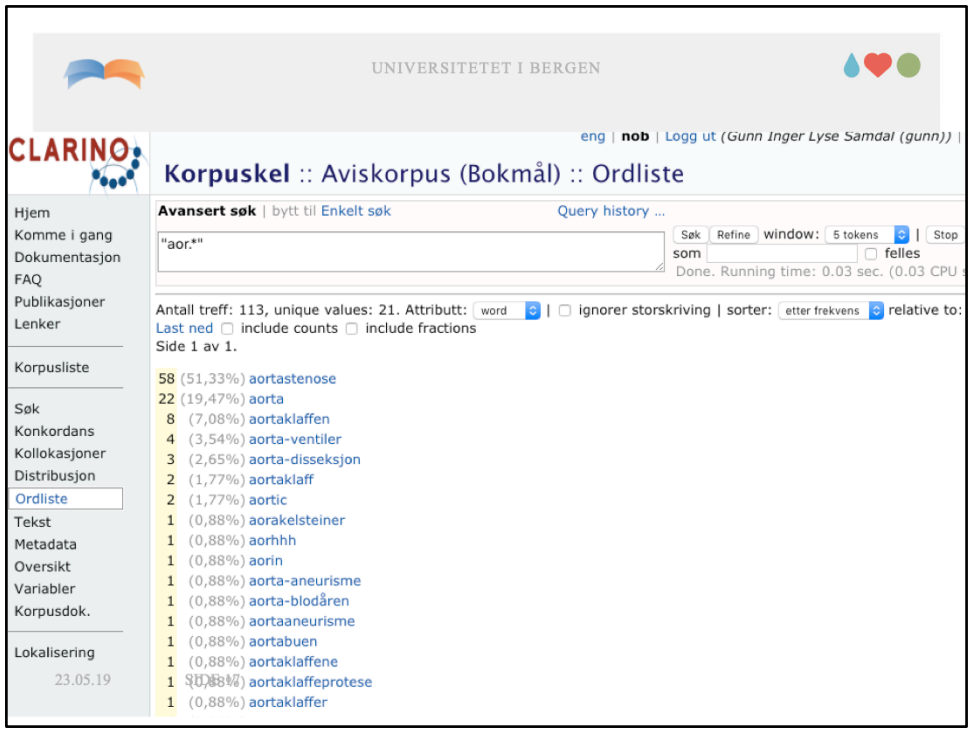
Kollokasjonar er ord som har ein statistisk tendens til å bli funne saman.
Corpuscle har eit knippe statistiske mål; fra veldig enkle (rein frekvens, relativ frekvens), til meir kompliserte statistiske mål.

Kollokasjonar er ofte til hjelp for å identifsere typiske bruksdøme i ordboka, men óg moglege underoppslag
Øvst: Fjør i hatten, nede på linje 7 fjær i hatten.

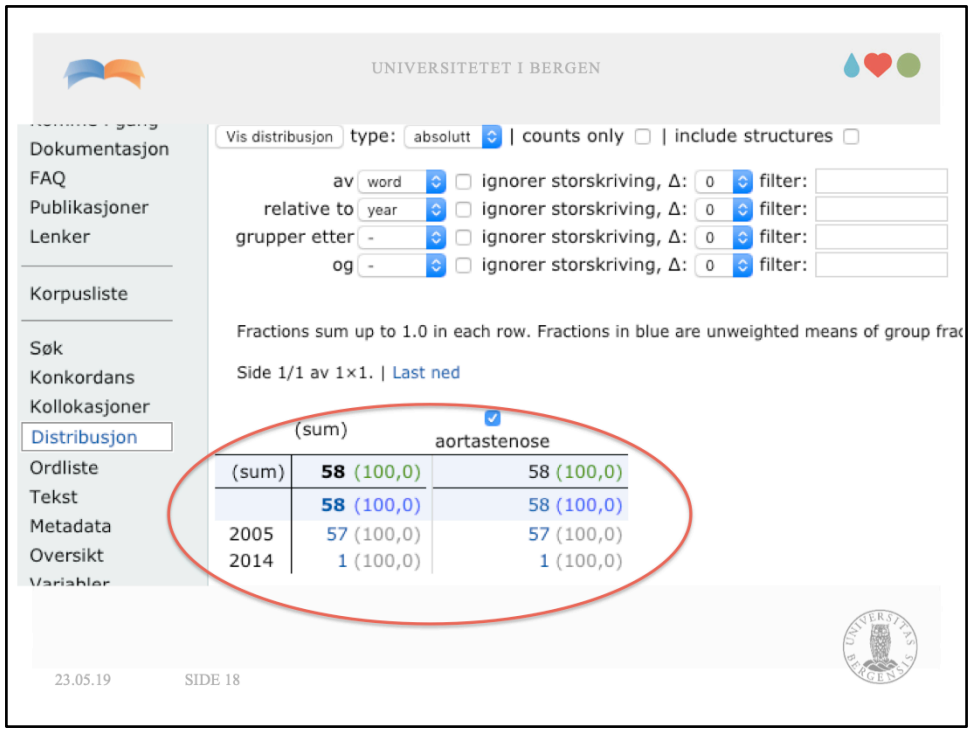
31 templater lagra

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]

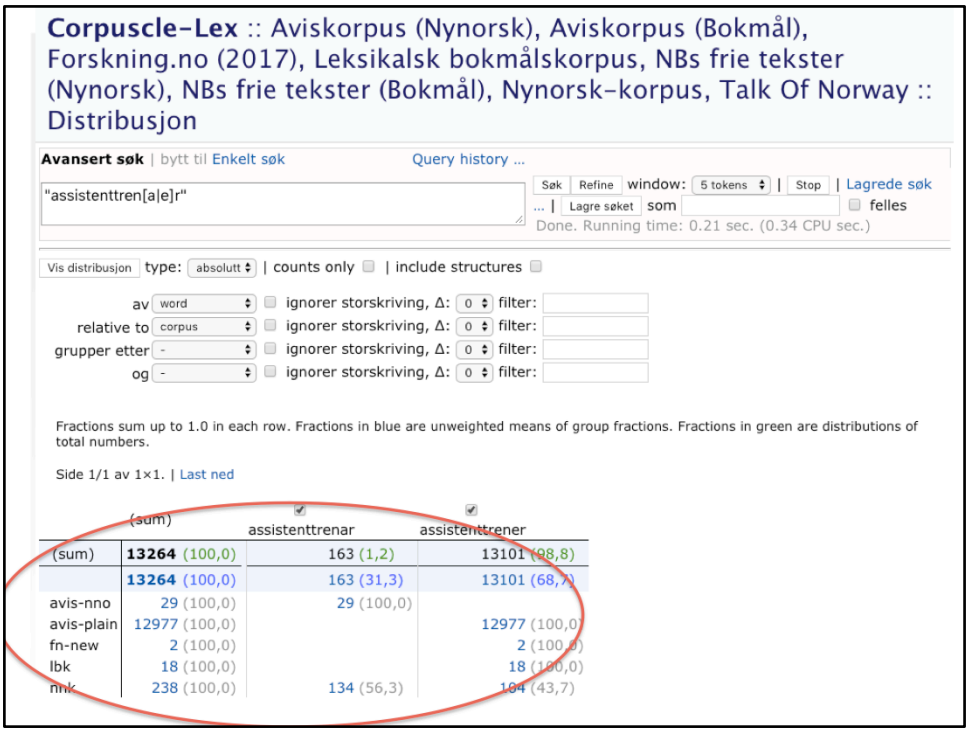


”aortastenose” ser kjempefrekvent ut! 58 treff, dobbelt så mange treff som ”aorta” (som er i ordbøkene våre)
Men viss du ser på distribusjonen til ”aortastenose” over tid (eller nøyare på kollokasjonane)
[klikk]

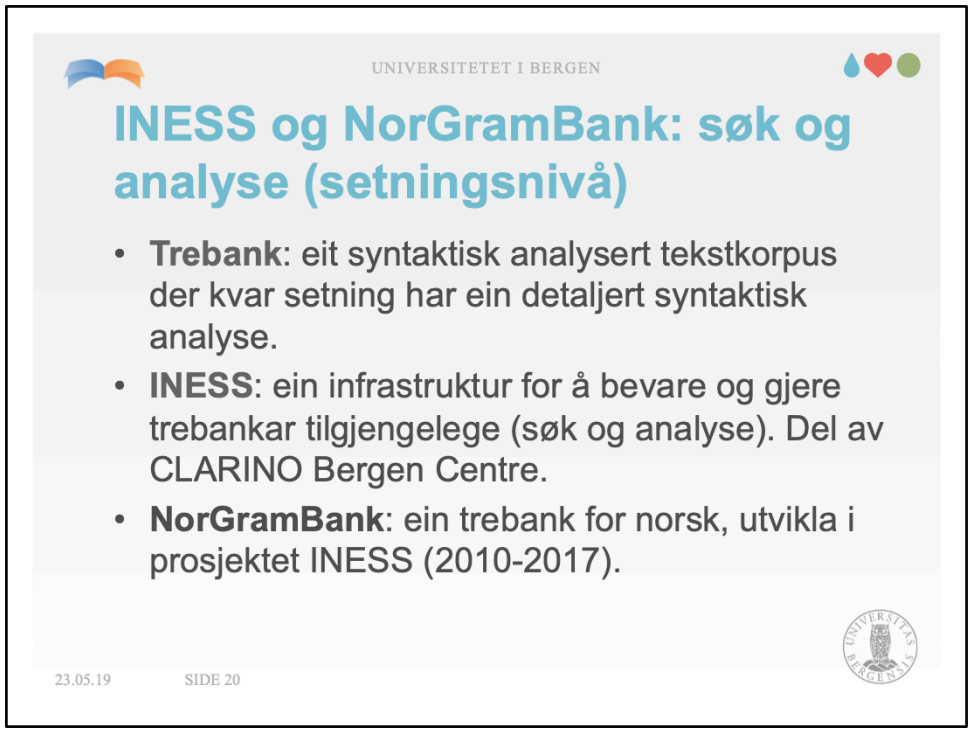


…så ser du at så å seie alle treffa er frå 2005, faktisk frå mars/april 2005, då kong Harald vart innlagd på sjukehus med hjerteklaffproblem (=aortastenose)

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

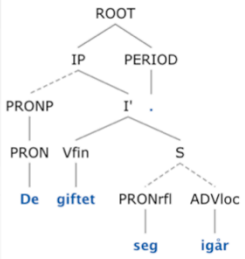


Assistenttrenar/-trener
Hvis et ord forekommer mange gonger kun i eitt kildedokument eller kun i aviskorpuset, så viser dette kun at ordet er frekvent i eit domene.



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.

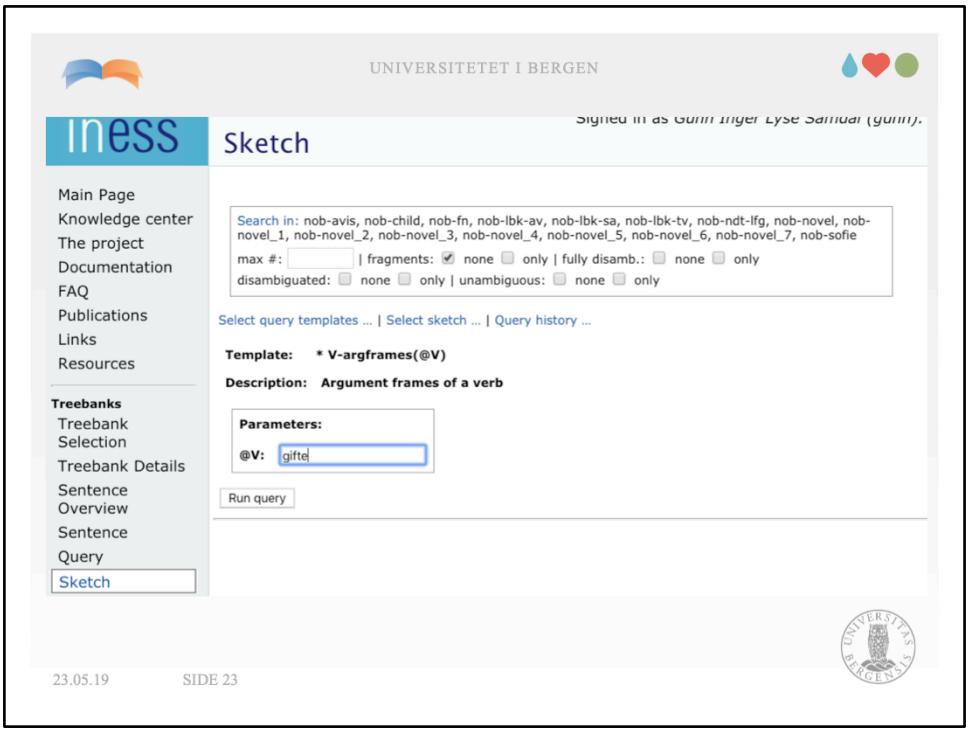
C-structure

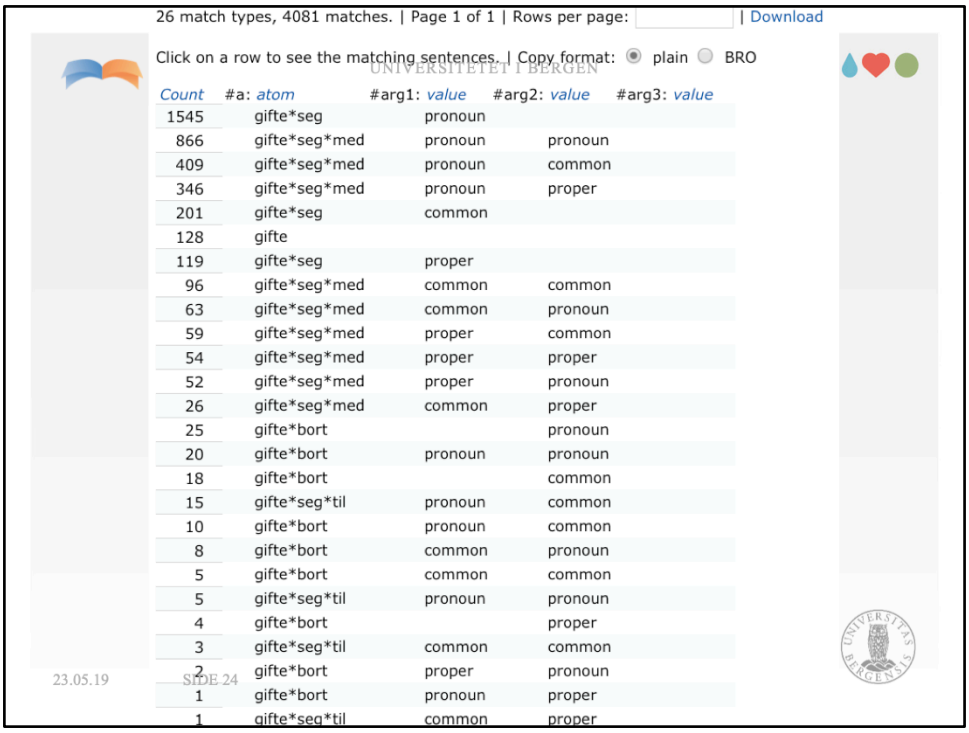


F-structure

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

31 templater lagra

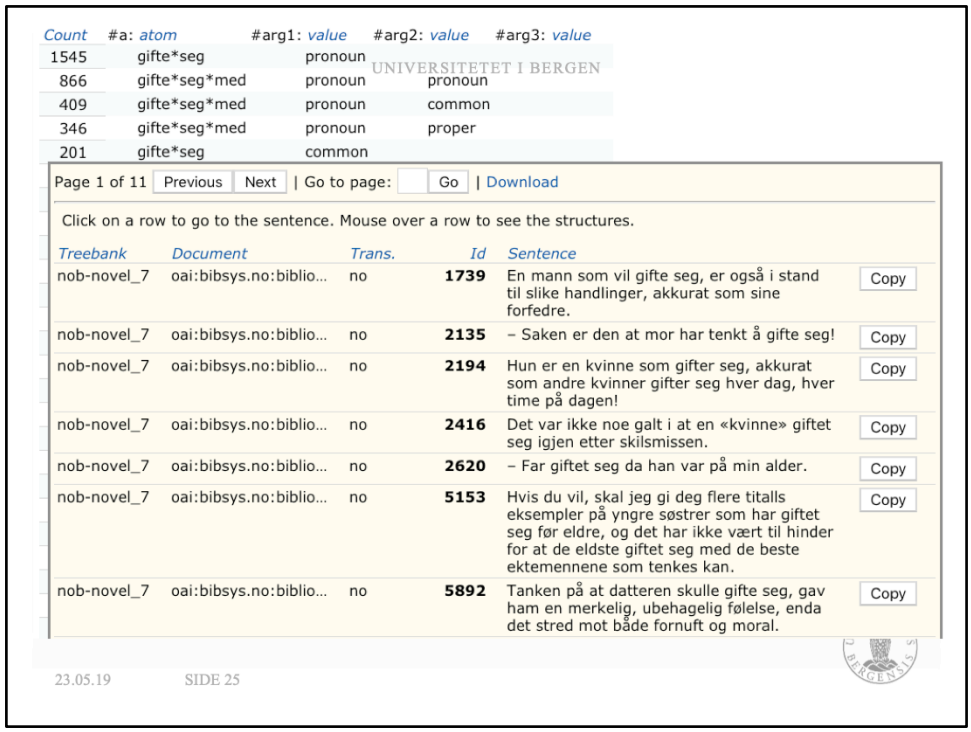

UNIVERSITETET I BERGEN
Signed in as Gunn Inger Lyse Sandal (gunn).
INESS
Sketch
Main Page
Knowledge center
The project
Documentation
FAQ
Publications
Links
Resources
Treebanks
Treebank Selection
Treebank Details
Sentence Overview
Sentence
Query
Sketch
Search in: nob-avis, nob-child, nob-fn, nob-lbk-av, 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: gifte
Run query
23.05.19 SIDE 23



26 match types, 4081 matches. | Page 1 of 1 | Rows per page: | Download

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

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	

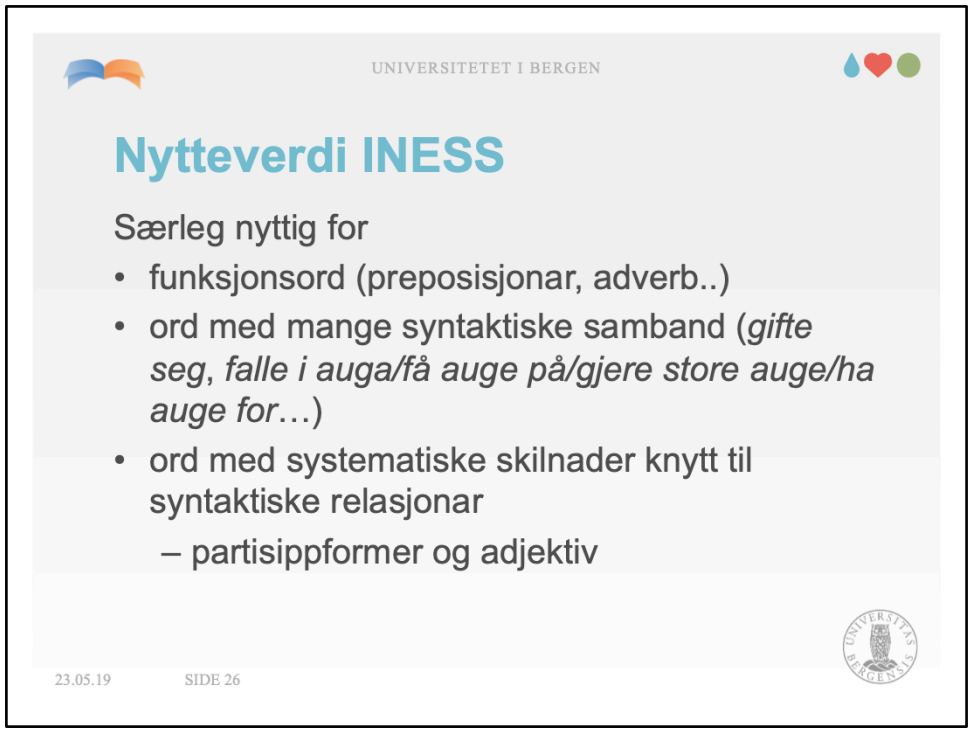


Count	#a: atom	#arg1: value	#arg2: value	#arg3: value
1545	gifte*seg	pronoun		
866	gifte*seg*med	pronoun	pronoun	
409	gifte*seg*med	pronoun	common	
346	gifte*seg*med	pronoun	proper	
201	gifte*seg	common		

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 søstrer 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



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


UNIVERSITAS BERGENSIS
UNIVERSITETET I BERGEN
Miljøfyrtårn®