Korpus og leksikografi

Leksikografane i
revisjonsprosjektet for *Bokmålsordboka* og *Nynorskordboka*
https://ordbok.uib.no

Språkrådet

UNIVERSITETET I BERGEN

Informasjonskjelder i revisjonsprosjektet

- tekstkorpus (lemmautval, tyding, syntaktisk åtferd, bruksdøme)
- Språkrådet (normering)
- Språksamlingane ved UiB (særleg *Metaordboka*)
- introspeksjon
- andre ordbøker

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)

Corpuscle – søk og analyse (ordnivå)

- http://clarino.uib.no/korpuskel/
- Grensesnitt for å tilgjengeleggjere, søkje i og analysere tekstkorpus.
- Inngår i CLARINO Centre Bergen (Universitetsbiblioteket ved UiB).

”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
Nynorskkorpus	nynorsk	107,8 mill	1866-2012	blanda	ja	avgrensa
NBs frie tekster (Nynorsk)	nynorsk	46,2 mill	1850-2010	blanda		open
Aviskorpus (nynorsk)	nynorsk	16,1 mill	1998-2015	sakprosa (avis)		open

UNIVERSITETET I BERGEN

Lemmautval: trunkerte søk

"curl.*"

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

felles

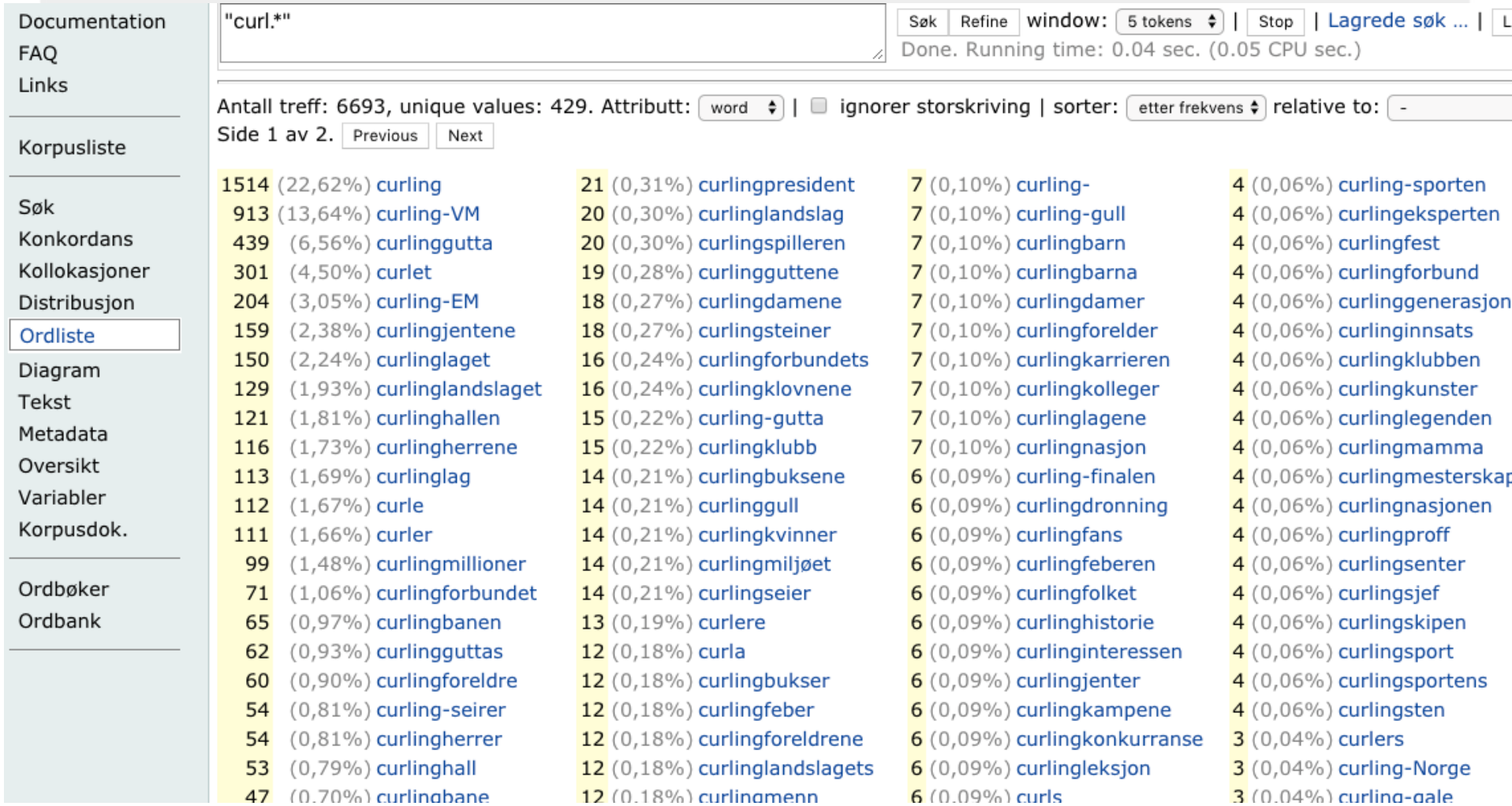
Done. Running time: 0.04 sec. (0.05 CPU sec.)

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

Kontekststørrelse: 400px

corpus		treff		year
avis-nno	en mellom Voss og Medkila i Prestegardsmoen i dag. Ballen	curla	inn i lengste hjørnet utan sjanse for målvakta. Fortuna vanı	2008
avis-nno	:azeem Ahmed det meste sjølv då han skar inn frå høgre og	curla	ballen i motsett hjørne. 0-2, og alt låg til rette for Hødds føı	2011
avis-nno	:azeem Ahmed det meste sjølv då han skar inn frå høgre og	curla	ballen i motsett hjørne. 0-2, og alt låg til rette for Hødds føı	2011
avis-nno	:azeem Ahmed det meste sjølv då han skar inn frå høgre og	curla	ballen i motsett hjørne. 0-2, og alt låg til rette for Hødds føı	2011
avis-nno	:azeem Ahmed det meste sjølv då han skar inn frå høgre og	curla	ballen i motsett hjørne. 0-2, og alt låg til rette for Hødds føı	2011
avis-nno	l sjå ut. TV2 har kjøpt OL-rettar. Altså risikerer vi i framtida	curling	med reklameavbrot. Samtidig har NRK vist at det finst alter	2011
avis-nno	istoria bak fotball, tennis, basketball, bordtennis, volleyball,	curling	, golf og synkronsymjing. Kvifor akkurat desse? Eg ville fort	2015
avis-nno	bildeCredit> FOTO: ARNE S. GJONE ¶ Les også En skole for	curlingbarna	- 06.02.2005 Svikt i lærebøkene, ¶ svikt hos lærerne? - 04.	2005
avis-nno	lre tiåringane har det, går det ut på eitt. Nokre har kalla dei	curlingbarna	, barna som får kosta banen sin fri av foreldra. Dei lærer fo	2006
avis-nno	rl- Eirik Kval. Unge menneske i dag er ikkje kalla glasur- og	curlinggenerasjon	for ingenting, men det er det vaksne som har lagt opp til. M	2015
avis-nno	får dei unge mange kallenamn, som «lydig», «flinkis» og «	curlinggenerasjonen	». Angrepa kjem gjerne når nokon går ut i media og taler fo	2014
avis-nno	274481.ece> < F ¶ ~ Norsk seier over Japan ¶ De norske	curlinggutta	slo Japan 8-5 i torsdagens første VM-kamp. ¶ NTB Publisert	2006
avis-nno	leklasse, og at det på havets botn ligg ein tusen år gammal	curlingstein	. - Nei, nei - han er ikkje tusen - han er 500 år gaı	2015
avis-plain	ui " på Hawaii, Paramount går snart i gang med " Girl in the	curl	", og Warner Bros skal lage " Lisa Andersen story ", for å ne	2001
avis-plain	ui " på Hawaii, Paramount går snart i gang med " Girl in the	curl	", og Warner Bros skal lage " Lisa Andersen story ", for å ne	2001
avis-plain	og mars: ¶ Øvelse Sett repetisjoner ¶ Knebøy: 3 x 15 ¶ Leg	curl	3 x15 ¶ Leg extension 3 x15 ¶ (ca. 30-60 sek. pause melloı	2006
avis-plain	ai og juni ¶ Øvelse Sett repetisjoner ¶ Knebøy: 3 x10 ¶ Leg	curl	3 x 10 ¶ Leg extension 3 x10 ¶ (ca. 30-60 sek.pause melloı	2006
avis-plain	Smith 3-4 x 15 (på hver fot) ¶ Leg extension 3-4 x 15 ¶ Leg	curl	3-4 x 15 ¶ (ca. 30-60 sek.pause mellom hver serie) ¶ April,	2006
avis-plain	Smith 3-4 x10 (på hver fot) ¶ Leg extension 3-4 x 10 ¶ Leg	curl	3-4 x 10 ¶ (ca. 30-60 sek.pause mellom hver serie) ¶ Effek	2006
avis-plain	Nordby synes OL-isen er rask. - Det er ikke så mye skru ("	curl	") i isen heller som hun hadde trodd. - Jeg må innrømme at	2006
avis-plain	) forbanna på isen og forholdene. Vi er et lag som liker mye	curl	og dragninger, med mange steiner i spill. Det er dumt å sky	2006

Lemmautval (forts.): ordlister



UNIVERSITETET I BERGEN

Kollokasjonar: bruksdøme, underoppslag

Documentation
FAQ
Links

Korpusliste

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

Ordbøker
Ordbank

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

Søk | Refine | window: 5 tokens | Lagre søket som

Done. Running time: 32.39 sec. (

Vis kollokasjoner 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	deflasjonskrise _ **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	trisnuta **hattane**
1	2	**hattane** _ utfrunsa
1	-2	stiletthæl _ **Hattane**
7	-1	bredbremmete **hatten**
2	-1	SIN **HATT**

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)

UNIVERSITETET I BERGEN

CLARINO

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

Korpuskel :: Aviskorpus (Bokmål) :: Ordliste

- Hjem
- Komme i gang
- Dokumentasjon
- FAQ
- Publikasjoner
- Lenker

- Korpusliste

- Søk
- Konkordans
- Kollokasjoner
- Distribusjon
- Ordliste
- Tekst
- Metadata
- Oversikt
- Variabler
- Korpusdok.

- Lokalisering

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

"aor.*"

Søk | Refine | window: 5 tokens | Stop

som felles

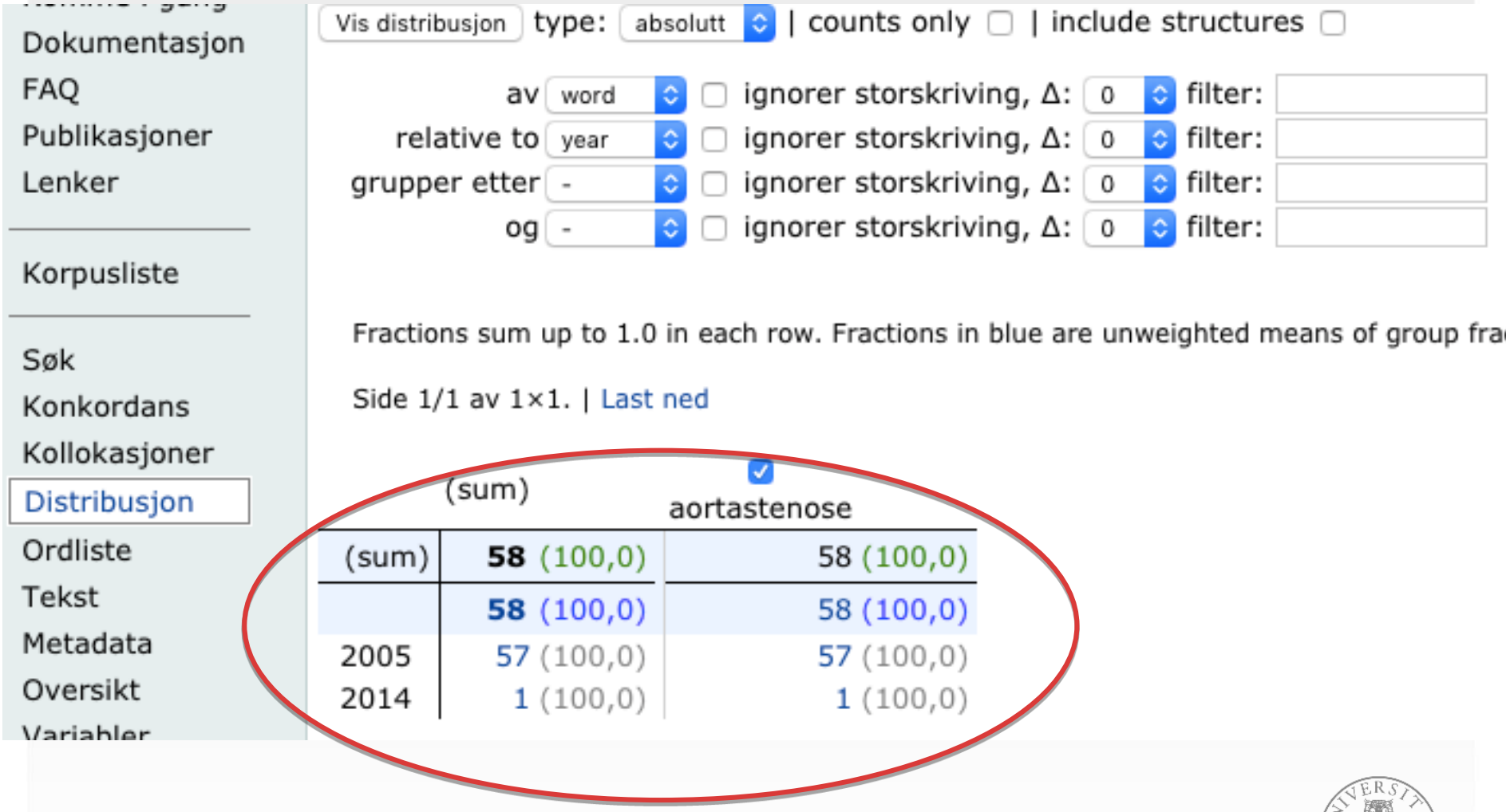
Done. Running time: 0.03 sec. (0.03 CPU s

Antall treff: 113, unique values: 21. Attributt: word | ignorer storskriving | sorter: etter frekvens relative to:

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


Dokumentasjon
FAQ
Publikasjoner
Lenker
Korpusliste
Søk
Konkordans
Kollokasjoner
Distribusjon
Ordliste
Tekst
Metadata
Oversikt
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 1×1. | Last ned
(sum)
aortastenose
(sum) 58 (100,0) 58 (100,0)
58 (100,0) 58 (100,0)
2005 57 (100,0) 57 (100,0)
2014 1 (100,0) 1 (100,0)

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.

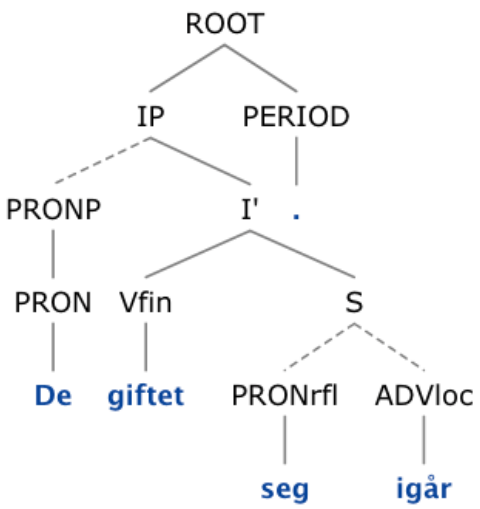
Side 1/1 av 1×1. | Last ned

	(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)
nnk	238 (100,0)	134 (56,3)	104 (43,7)

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).

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 |
          | 8 REF +, PRON-TYPE pers, PRON-FORM de, PERS 3, NUM pl, DEF +,
          | CASE nom |
ADJUNCT   1 { 2 | PRED 'i går'
              | 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
```

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

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

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

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

Miljøfyrtårn®