

# Korpuslingvistik (SV2119)

## Föreläsning 2: Språkbankens korpusar och sökverket Korp



Språk  
BANKEN



Richard Johansson

`richard.johansson@svenska.gu.se`

20 september 2013

# 1. introduktion



# dagens föreläsning

- ▶ Språkbankens korpusar: vilka finns och hur är de annoterade
- ▶ sökverktyget Korp



# något ytterligare om konkordanser

- ▶ konkordans: de olika sammanhangen för ett ord, för t.ex.
  - ▶ lexikografi, översikt över olika betydelser
  - ▶ undervisning och självstudier
- ▶ Korp (dagens ämne) är ett konkordansverktyg

DN 1987

Med åtta man på scenen kör de rak och jordnära <b>rock</b> , trots det himmelska budskapet.
1 640 meter: 1) <b>Rock</b> Garden -- Karl-Erik Nilsson 15,7, 2) Robis Lafayette
20.00 New Brighton <b>Rock</b> -- Tio rockband och soloartister uppträder vid E
TROLLKARLEN BLACK BERT framträder på Hard <b>Rock</b> Café, Sveavägen 75, kl 14.
Syftet med karavanen var inte bara att hedra <b>rockens</b> kanske störste sångare utan också att fästa uppr
Vi ser gärna att det anordnas <b>rock</b> för ungdomar, men vi vill inte ha dem inne i stai
s ganska experimentell och inte längre så sk (r) amlig <b>rock</b> .
<b>Rock</b> mot våld i Högsätrahallen mellan kl 18 och 01 m
Det populära <b>rock-</b> och rörelseparet Mora Träsk återkommer nu me
<b>Rockens</b> rebell och drottning
Mycket skratt, säger hon till statisterna i vita <b>rockar</b> .
ag betalar dubbel garderobsavgift bara ni inte rör min <b>rock</b> .
<b>Rock</b> och pop i Karlskoga

# konkordansverktyg

- ▶ Korp (inklusive Corpus Workbench) är gratis och kan installeras på din egen dator
- ▶ installationen kräver dock en del teknisk ansträngning och om du har en egen korpus kan det vara mer lämpligt att installera ett enklare konkordansverktyg
- ▶ exempel på fristående verktyg:
  - ▶ AntConc: <http://www.antlab.sci.waseda.ac.jp/software.html>
  - ▶ WordSmith: <http://www.lexically.net/wordsmith/> (Windows)



## 2. Språkbanken



## kort om Språkbanken

- ▶ ~1970: första svenska korpusen: Press-65
- ▶ 1972: professur i språkvetenskaplig databehandling (Sture Allén)
- ▶ 1975: Språkbanken (“Logoteket”)
- ▶ ...
- ▶ olika namn t.ex. Språkdata
- ▶ ...
- ▶ då: Språkbanken var synonymt med en samling texter för språkvetenskapliga studier
- ▶ idag: Språkbanken är en forskningsenhet med fokus på språkteknologi för det svenska språket genom tiderna



### 3. Korp





# sökverket Korp: inledning

- ▶ Språkbankens korpusar söks med hjälp av verktyget Korp
- ▶ Korp finns på <http://spraakbanken.gu.se/korp>
- ▶ användarhandledning <http://spraakbanken.gu.se/swe/forskning/infrastruktur/korp/anvandarhandledning>

DN 1987

Med åtta man på scenen kör de rak och jordnära	<b>rock</b>	, trots det himmelska budskapet.
1 640 meter: 1)	<b>Rock</b>	Garden -- Karl-Erik Nilsson 15,7, 2) Robis Lafayet
20.00 New Brighton	<b>Rock</b>	-- Tio rockband och soloartister uppträder vid E
TROLLKARLEN BLACK BERT framträder på Hard	<b>Rock</b>	Café, Sveavägen 75, kl 14.
Syftet med karavanen var inte bara att hedra	<b>rockens</b>	kanske störste sångare utan också att fästa uppr
Vi ser gärna att det anordnas	<b>rock</b>	för ungdomar, men vi vill inte ha dem inne i stadi
s ganska experimentell och inte längre så sk ( r ) amlig	<b>rock</b>	.
	<b>Rock</b>	mot våld i Högsåtrahallen mellan kl 18 och 01 m
Det populära	<b>rock-</b>	och rörelseparet Mora Träsk återkommer nu me
	<b>Rockens</b>	rebell och drottning
Mycket skratt, säger hon till statisterna i vita	<b>rockar</b>	.
ag betalar dubbel garderobsavgift bara ni inte rör min	<b>rock</b>	.
	<b>Rock</b>	och pop i Karlskoga

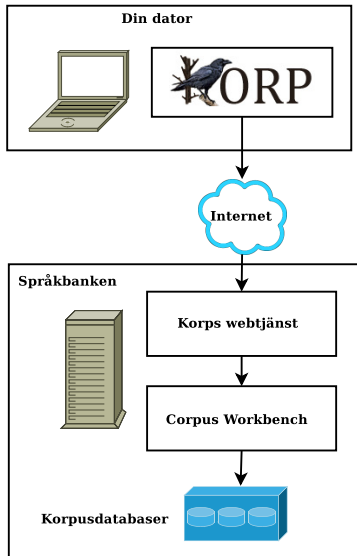


# Korps uppbyggnad

- ▶ Korp är ett webbaserat sökverktyg som kommunicerar med ett serverprogram (webbtjänst) på Språkbanken
- ▶ du kan också själv utveckla egna program som använder Korps webbtjänst
  - ▶ <http://spraakbanken.gu.se/eng/research/infrastructure/korp/ws>
- ▶ Korp bygger på ett underliggande verktyg, Corpus Workbench, utvecklat vid universitetet i Stuttgart:  
<http://cwb.sourceforge.net/>



# Korps uppbyggnad



# sökningar i Korp

- ▶ enkel sökning på enskilda ord
- ▶ utökad sökning med mer komplexa kriterier (grafiskt)
- ▶ avancerad sökning med sökspråket CQP



# enkla sökningar i Korp

- ▶ sökning på enskilt ord
- ▶ sökning på grundform
- ▶ välj korpusar att söka i
- ▶ tips: om det går väldigt långsamt, välj ett mindre antal korpusar
- ▶ resultatflikar: KWIC, statistik, ordbild



- ▶ sammanställning och rangordning
- ▶ exportera

- ▶ visar vanliga syntaktiska sammanhang
  - ▶ adjektiv: vanliga substantiv
  - ▶ substantiv: vanliga adjektiv, prepositioner och verb
  - ▶ verb: vanliga substantiv och adverbial

# korpusar i Språkbanken

<http://spraakbanken.gu.se/swe/resurser/corpus>

- ▶ modern dagstidningstext: GP, DN, ...
- ▶ modern romantext: Bonniers, Norstedts, ...
- ▶ populärvetenskap: Läkartidningen, F&F, ...
- ▶ sociala medier: bloggar, twitter
- ▶ 1800-talslitteratur (Litteraturbanken)
- ▶ medeltida text (fornsvenska)
- ▶ färöisk textkorpus (dagstidningstext)
- ▶ parallella korpusar
- ▶ inlärarkorpusar
- ▶ ... och en hel rad andra

Nedladdningsbart: <http://spraakbanken.gu.se/eng/node/1587>





# fördelning

- ▶ Bloggar: 392M
- ▶ Tidningar: 304M
- ▶ Twitter: 250M
- ▶ Wikipedia: 122M
- ▶ Web: 115M
- ▶ Finlandssvenskt: 66M
- ▶ Vetenskapligt: 46M
- ▶ Riksdag och EU-parlament: 38M
- ▶ Romaner: 22M
- ▶ ...
- ▶ **Totalt: 1402M**



# exempel: bloggkorpusar

BLOGGHR2001 (stödjer ej utökad kontext)

Man **ba**; get a room for godsake!

Bara så ni inte blir glada och sedan **ba**, ahmen vad är det här för skit?

BLOGGHR2002 (stödjer ej utökad kontext)

Carmen Ly kom in till mig en liten stund sedan och **ba** att jag skulle rädda henne från en tvestjärt som hade nuddat h

BLOGGHR2003 (stödjer ej utökad kontext)

PÅ PIPPI OCH SÅ SOMNADE HON I MIN KONTOR EFTER HON BÖNADE O **BA** O TITTA PÅ PIPPI 20 GÅNGEN I GÅR.

Mannen bönade o **ba** om låv o gå till jobbet.

BLOGGHR2004 (stödjer ej utökad kontext)

**ba** vafan ere.

alla i bilen ber o **ba** vi kommer do o allt fan.

**ba** hoppa in hoppa in.

BLOGGHR2005 (stödjer ej utökad kontext)

ddade lättemaskinen och de sprutade såklart överallt, o där står ja o **ba** "hjäälp deva inte meningen, de började av sej själv ", suck.

Alla andra får **Ba**, B eller U.

Jag har **ba** "sååååå mycket att blogga, i don 't even know where to start!  
att ha tjej just nu ... Vaddå, ska han ha pojkvän i stället (haha, skojar **ba** ) inte för att det är nå 't problem för mig ... Jag frågade henr  
ssen att inte ge en tid ... Om de försöker föregå dig o säger: " Jag ska **ba** 'ut en stund, kommer om ett tag ...! " Så betyder det troligen:  
annare efter undermeningar, konst-pauser och antydningar ... " Ska **ba** 'till en kompis ... " är alldeles för vagt ... Det betyder antaglig

Asså, alla kompisar ringde mig o ja **ba** "; asså, jag e på stan med Dogge ... " Nu verkar ryktet nått alla,  
ra CHILLA resten av dagen, skriva ut lite skolgrejer, fixa naglarna och **ba** döööö.

Ögonen **ba** plopp så trött.

Min livmoder **ba** : Livmoder: Hej, allt är klappat och klart för bebisen \* glad, kärll

Jag **ba** HAHA FUCKERS GENOMSKINLIGA NI ÄR DÅ NÅGON LÄSTE JU SP

Plötsligt blir det grönt och dom står där och **ba** " VÄNTA SKA BARA LÄSA UT DEN HÄR SIDAN!!!



## 4. annoteringsmodell i Språkbankens korpusar



## om annotering

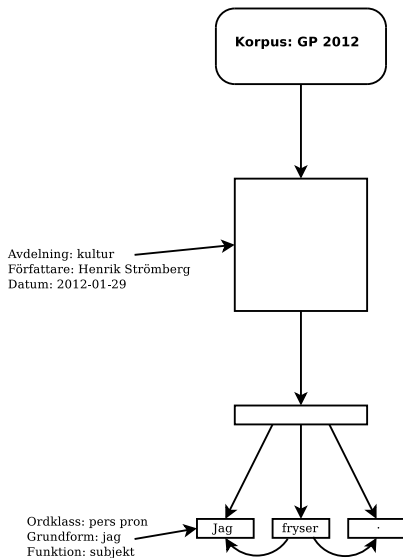
- ▶ för att göra intressantare analyser behöver vi mer information än bara texten i sig
- ▶ **annoteringsmodell**: hur beskriver vi vad som finns i korpusen?
- ▶ vilka enheter finns?
  - ▶ texter, meningar, ord, ...?
- ▶ hur är enheterna relaterade?
- ▶ vilka attribut har respektive enhet?
  - ▶ textens författare, publiceringsdatum, ...?
  - ▶ ordets böjningsform, ...?
- ▶ mer i föreläsning 3!

# annotering i Språkbankens korpusar

- ▶ vilka enheter finns?
  - ▶ **korpusar, dokument, meningar, ord**
- ▶ hur hänger enheterna ihop?
  - ▶ en korpus består av dokument
  - ▶ ett dokument består av meningar
  - ▶ en mening består av ord
  - ▶ ett ord består av fritext
  - ▶ ett ord kan vara länkat till ett annat ord genom **dependensrelation** (stöds dock inte av Korp)
- ▶ vilken information finns om enheterna?
  - ▶ ett dokument har **textattribut** (korpusberoende)
  - ▶ ett ord har **ordattribut**



# exempel



# textattribut

- ▶ textattributen beror på vilken korpus vi använder.
- ▶ exempel GP 2012:
  - ▶ avdelning i GP
  - ▶ författarnamn
  - ▶ datum
- ▶ exempel Strindbergs brev:
  - ▶ författarnamn
  - ▶ mottagarnamn
  - ▶ år
  - ▶ band i brevsamlingen
  - ▶ ...
- ▶ exempel bloggkorpora:
  - ▶ författarens namn, ålder, hemort, ...
  - ▶ bloggens teman

# ordattribut

- ▶ ordet i sig
- ▶ grundform
- ▶ ordklass, t.ex. “verb”
- ▶ formbeskrivning (msd), t.ex. “verb presens aktiv”
- ▶ förled och efterled i sammansättning
- ▶ ...





## varifrån kommer annoteringen?

- ▶ korpusar och dokument är givna (eller manuellt indelade)
- ▶ texterna är antingen elektroniska i ursprungsformen (t.ex. GP) eller digitaliserade (ibland med OCR)
- ▶ i de flesta fall automatiskt indelade i ord och meningar
- ▶ i de flesta fall automatiskt lingvistiskt analyserade



## 5. utökad sökning



# utökade sökningar i Korp

- ▶ sök på ordattribut
- ▶ kombination av villkor: och, eller
- ▶ sökning på en kombination
- ▶ samma resultatflikar: KWIC, statistik, ordbild



## exempel

- ▶ verb som följs av “Göteborg”?
- ▶ vanligaste substantiv i partiprogrammen inför valet 2002?



# begränsningar i Korp

- ▶ begränsade möjligheter för t.ex. syntaktisk sökning
- ▶ t.ex. “vilka objekt är vanligast för verbet *köpa*”?
- ▶ ordentliga syntaktiska sökverktyg beskrivs i föreläsning 5



# experimentella korpusar: Korplabbet

- ▶ under **Korplabbet** (välj kugghjulet till höger) finns korpusar under utveckling, framför allt historiskt material
  - ▶ lag och rätt, t.ex. landskapslagar, Tänkeböckerna
  - ▶ tidningstext från 1700- och 1800-talet
  - ▶ biblar
  - ▶ äldre romaner



# exempel: Tänkeböckerna (under Lagrummet)

KWIC Statistik

Antal träffar: 89

Förslagslista 1 2 3 4 Nästa Visa kontext

STOCKHOLMS STADS FÄREBÖCKER

Han swarade sigh inthet hafua föränt **döden** myckett mindre blifua mesterman.

is Anderson i Upsala klagar öfuer D. Johan Bothvidi att hans barn Anna Månsdötter är på hans arbete **döden** blifwen, i thet at när hon stod och arbetade i boden föll en bräike neder och slogh huffudet i k

Jagh Hanns Olofzonn swär widh Gudh och hans h. ewangelium, at thet dråp som på den **döde** pigan skett är, är med wåda skedt, och emot mitt wett och willia.

Efter som timbermannen på sin eed här betygade om omständigheten, efter dan **döde** pigann, i sådant aff tydel wåda war tilkommet, såsom och fadrenn här for retten silaf det beb

item bär alle schrifter fram ifrå Blasij **död** 1621 på sin öed tilgbrände.

weson i Nyland i Perno sochn i Embom by som goringen gjorde, huad han hade med den andre som **döder** är, at beställa.

råbbade den i sängen lågh, mån om han fick hugg eller ej, seger han sigh inthet weta, utan när thän **döde** föll på glåfwet, sade Marcus slå watn på honom.

När han slap up wille han göma sigh bak om sengen, då slogh han den **döde** .

Och flere som hafua sedtt den **döde** .

til **döden** framfarat så kann rätten ej annars beslutat än at Clas Allert betalr huad han efter underrätter

Likwal efter han i samma schuldforndan hafuer till **döden** framfarat så kann rätten ej annars beslutat än at Clas Allert betalr huad han efter underrätter

Stathåldaren gaf tilkenna at een **död** man är funnen ett stycke ifrå hans glird halsen afskuren om någon wet hwar han är hemma.

Jören Behms skräddares mester swänn Jören be d bekänner at den **döde** är hans landzman född i Tyningen och är een schredrare be: d m. Hanns och någre dagar för är

Hon begärer inthet blifua hos Hendrich uthan hos Jacob där will hon blifua till sin **dödz** stund.

Hendrich bekänner och när han lågh på sin sotesångh och inthet annat förwäntade ahn **döden** kallade han Erich Larson till sigh bad honom till förmyndere så wäll öfwer hans hustru såsom i

Han seger och at den **döde** låfuede komma här i staden till påschoa.

Han seger och att den **döde** fick tw par nye sko af honom bekomet.

Han seger och at capitenerne mötte på Horns stön där togo du den **döde** med sigh på slådan.

kum han ifrå Callmar med den **döde** på skap till Stockholm är 16 åhr wid pass.

Den **döde** fick honom peningar i förwaring i Callmar haf.

Een lärnt red byndell hade den **döde** .

De 4 r. daller som den **döde** lårnte honom haf.

Jören Behms mester swän som nu är hoos Arent Hysings bekæmer at den **döde** haf.

Huru wdth han följde den **döde** ?

taget r. daller till låns af den **döde** .



# exempel: Digidaily

KWC Statistik

Antal träffar: 26 905

Första sida 1 2 3 4 5 6 7 8 9 10 11 - 1076 1077 Nästa Visa kontext

Resultat

Kesjar **Napoleon** har varit nära att bli tillfångalagen af prussarne

**Napoleon**, uppfallande det till folje af haudeusera vacklande i sin ställning, har till en början ombytt s

Kesjar **Napoleon** liar dragt i fall med spar-| tank enkelhet.

nu för tiden uppdiktas, anføres: "Berliner liorsen-Ztg" har på fullt allvar utspriddt ryktet, att kesjar **Napoleon** blifvit galen och i hopplost tillstend fört till S. t Cloud.

Kesjar **Napoleon** var sedan i går arméens överbefälhavare

Prins **Napoleon** följer kesjarens stab, men utan särskildt befäl.

\* Den nye heliges (den förste **Napoleon**) bronson och blodsforvand understodjar genom sitt namn hvalfågar och tak | på byggnad

**Napoleon** talar i sin deposesch allt manligt re signerad - språk och uppkallar alla patriotismens | ansträng

Man erfar vid denna prokla, mation tydligt att **Napoleon** djupt känner huru | han sett allt på ett kort och all det är sitt största-väggel han uför

När prins **Napoleon** med sin fader öfvervart Saarbricksens intagande, afsändes följande deposesch, från kesjaren t

Kesjar **Napoleon**, som i Thursdags afred till är mken har till trupperna utfärdat en proklamation full af lif och

Det franska örlogsskepp, som enligt telegram skulle strandat, visar sig nu varit lustjakten " Jérôme **Napoleon** ", som varit sånd att hmtla prins **Napoleon** från sin afbrutna Spetsbergs-expedition.

n skulle strandat, visar sig nu varit lustjakten " Jérôme **Napoleon** ", som varit sånd att hmtla prins **Napoleon** från sin afbrutna Spetsbergs-expedition.

M : Flotta Acngissons son, **Napoleon**, 4 år, ass.

anlog titeln af Duc d' Aumont efter sin farbror, lian innehade ett generals-befäl i Normandien, då **Napoleon** återkom frånEiba; han begaf sig då öfver till England, samlade med modor och besvar en frik

Både borddansen, Ericssons latoric = ma = chiii och tjelfwa Ludwig

Ludwig **Napoleon** roar sig med sin Kesjarinna i all hussig-het, och formar icke längre lidraga sig wcl = dms up

wal knappt n igot twincel alt "ous **Napoleon** är lik- angelagen om förbund med Österrike, som ett stocstätt wänslapigt förhållande till Ro

Detta måste war-ra af stort värde för Vobis **Napoleon**, ty det ger honom tillfälle att genom en storartad hög

Genom bisättning af Reichstadt stoft wid sidan af **Napoleon** de = l i i les kissa, fylfts ben lucka, som ända till denna stund står öppen emellan **Napoleon** ds

sidan af Napoleon de = l i i les kissa, fylfts ben lucka, som ända till denna stund står öppen emellan **Napoleon** den 3-die och legitmiteten.

Men från den stunden är han = verkliggen Kesjar Nikolaus icke wille erkänna honom, och om han låter rista

1 tredje, såsom hwilken Kesjar Nikolaus icke wille erkänna honom, och om han låter rista namnet." **Napoleon** li " på Hertigens af Reichstadt kista, så ger han det ögenstenligaste bewis på saken.

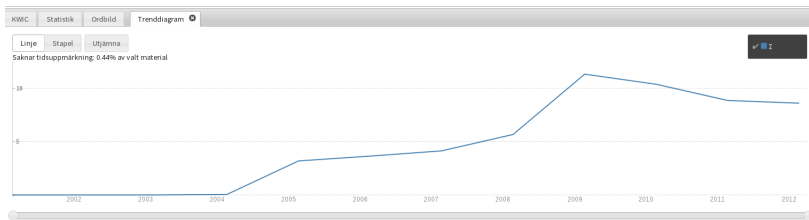
Ke s, r **Napoleon** tros hysa den afwigcu all af

Kesjar = loms **Napoleon** har uppbjudt hela sin

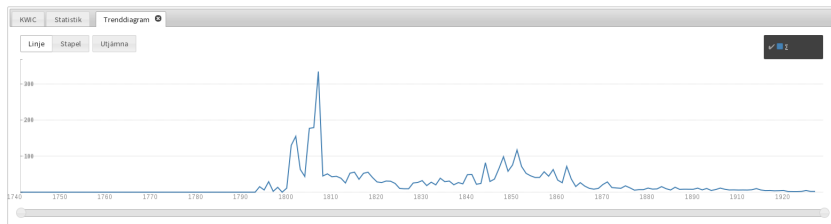


# trenddiagram: exempel på en neologism

- ▶ välj “visa trenddiagram” under statistikfliken



# exempel på variation pga historiska omständigheter



# avancerad sökning: sökspråket CQP

- ▶ prova att växla mellan utökad och avancerad!
- ▶ `[(word = "köttbulle") & (pos = "NN")] [(pos = "VB")]`
- ▶ `[((word = "köttbulle" | word = "hamburgare"))] [(pos = "VB")]`
- ▶ exempel på sökning som enbart stöds i CQP-läget:  
`[(word = "kö.*") & (pos = "NN")] [(pos = "VB")]`

<http://cwb.sourceforge.net/documentation.php>



## 6. information



## nästa föreläsning: annotering

- ▶ annoteringsmodell: hur beskriver vi språkliga fenomen systematiskt?
- ▶ format: hur lagrar vi annoteringen i filer?
- ▶ fallstudier
- ▶ annoteringsprocessen
- ▶ verktyg



# uppgift 1

- ▶ sökning i Korp
- ▶ utförs på egen hand
- ▶ kommer upp på kursens websida den 20/9
- ▶ **deadline 18/10**

