Korpuslingvistik (SV2119)
Föreläsning 2: Språkbankens korpusar och sökverktyget Korp

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

sökverktyget 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

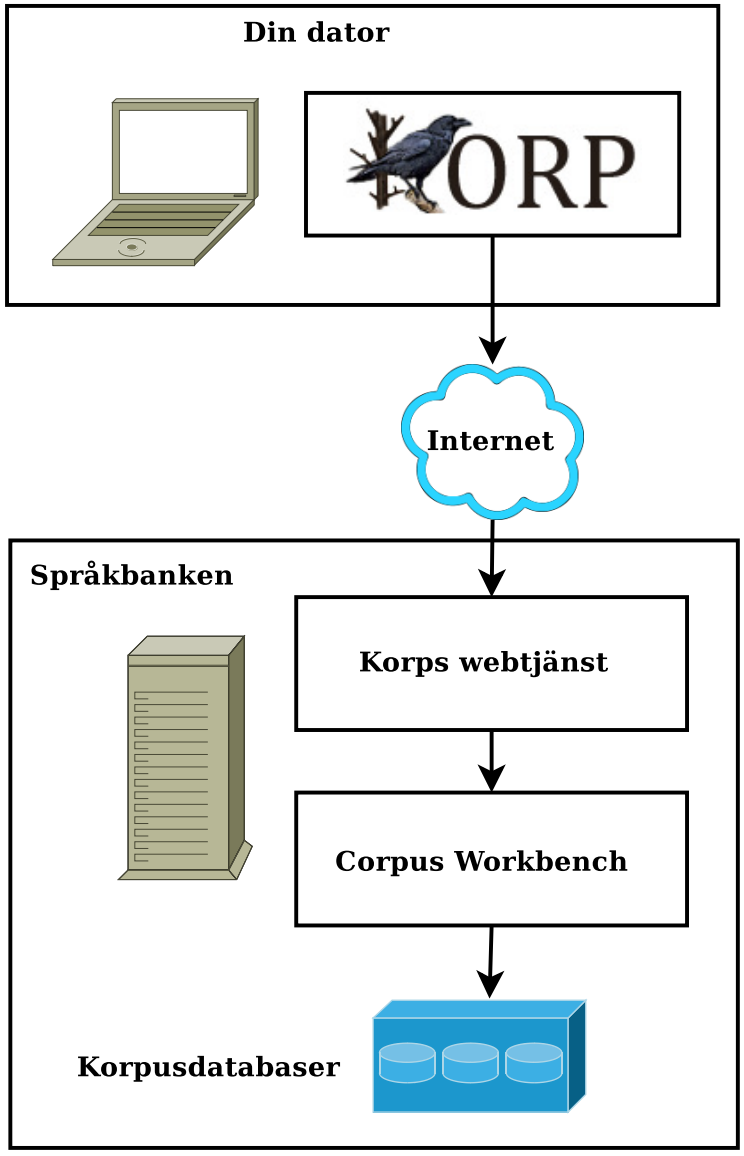
Med åtta man på scenen kör de rak och jordnära	**rock**	, trots det himmelska budskapet.
1 640 meter: 1)	**Rock**	Garden -- Karl-Erik Nilsson 15,7, 2) Robis Lafayet
20.00 New Brighton	**Rock**	-- Tio rockband och soloartister uppträder vid E
TROLLKARLEN BLACK BERT framträder på Hard	**Rock**	Café, Sveavägen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

statistik

- sammanställning och rangordning
- exportera

ordbild

- 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

Bloggmix 2001 (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?	
Bloggmix 2002 (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	
Bloggmix 2003 (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.	
Bloggmix 2004 (stödjer ej utökad kontext)	
ba vafan ere.	
alla i bilen ber o **ba** vi kommer do o allt fan.	
ba hoppa in hoppa in.	
Bloggmix 2005 (stödjer ej utökad kontext)	
ddade lattemaskinen och de sprutade såklart överallt, o där står ja o **ba** " hjääälp de va 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 henn	
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, kärl	
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

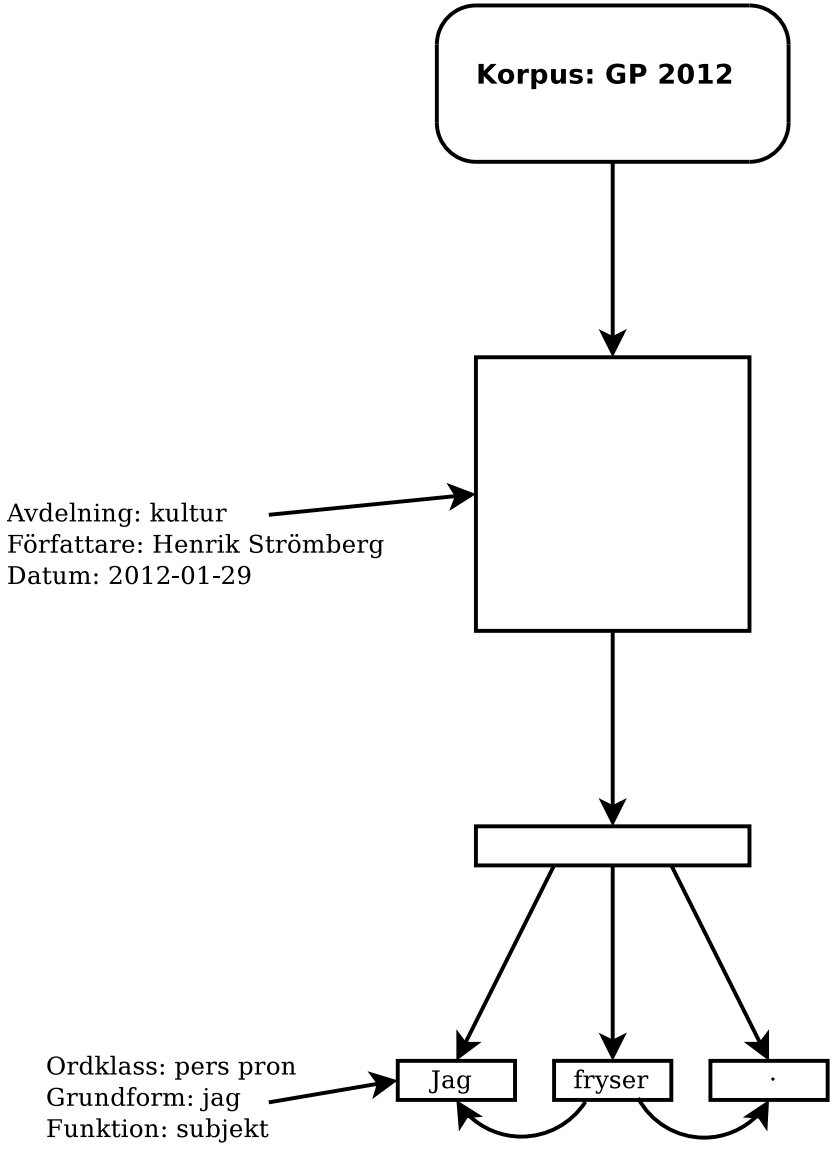
om annotering

- 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


Korpus: GP 2012
Avdelning: kultur
Författare: Henrik Strömberg
Datum: 2012-01-29
Ordklass: pers pron
Grundform: jag
Funktion: subjekt
Jag
fryser
.

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 bloggkorpusar:
  - 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?

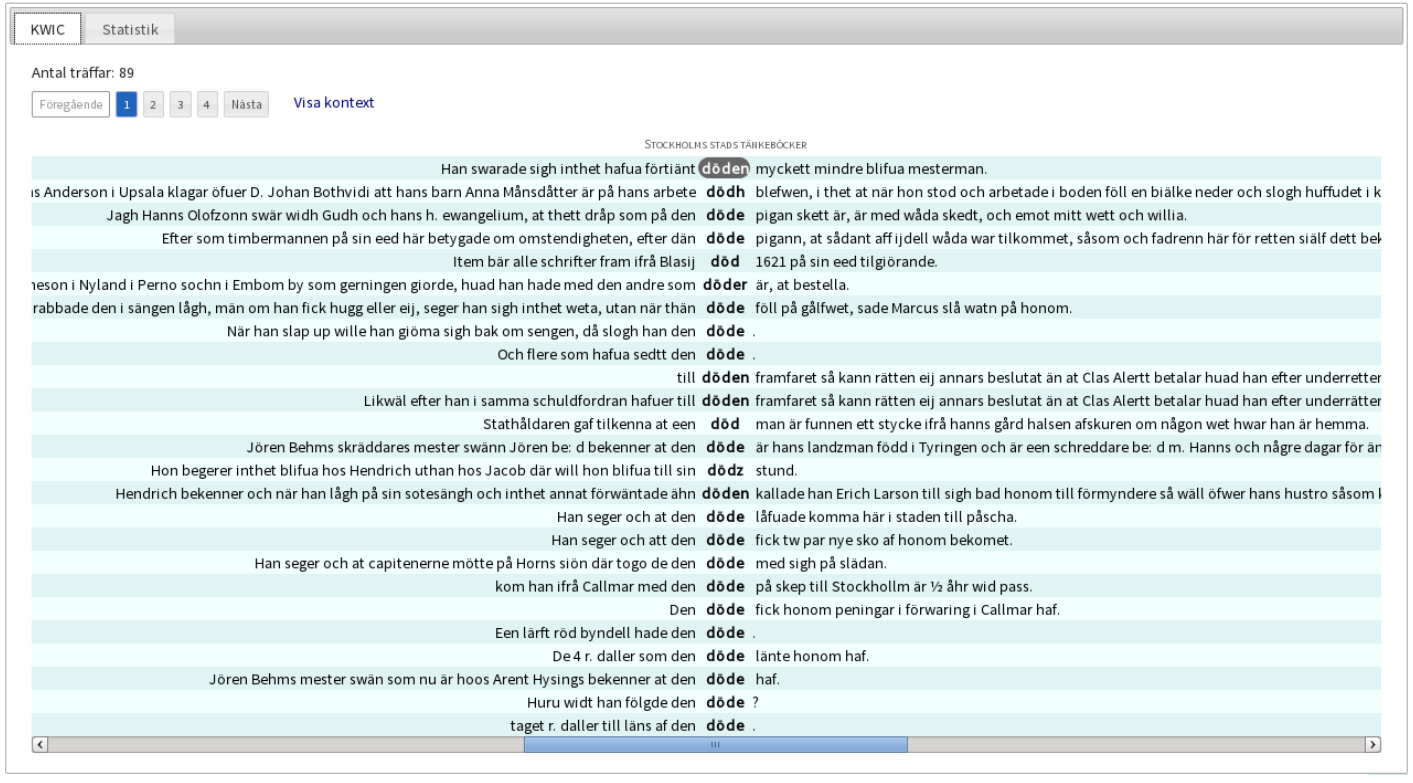
begränsningar i Korp

- 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

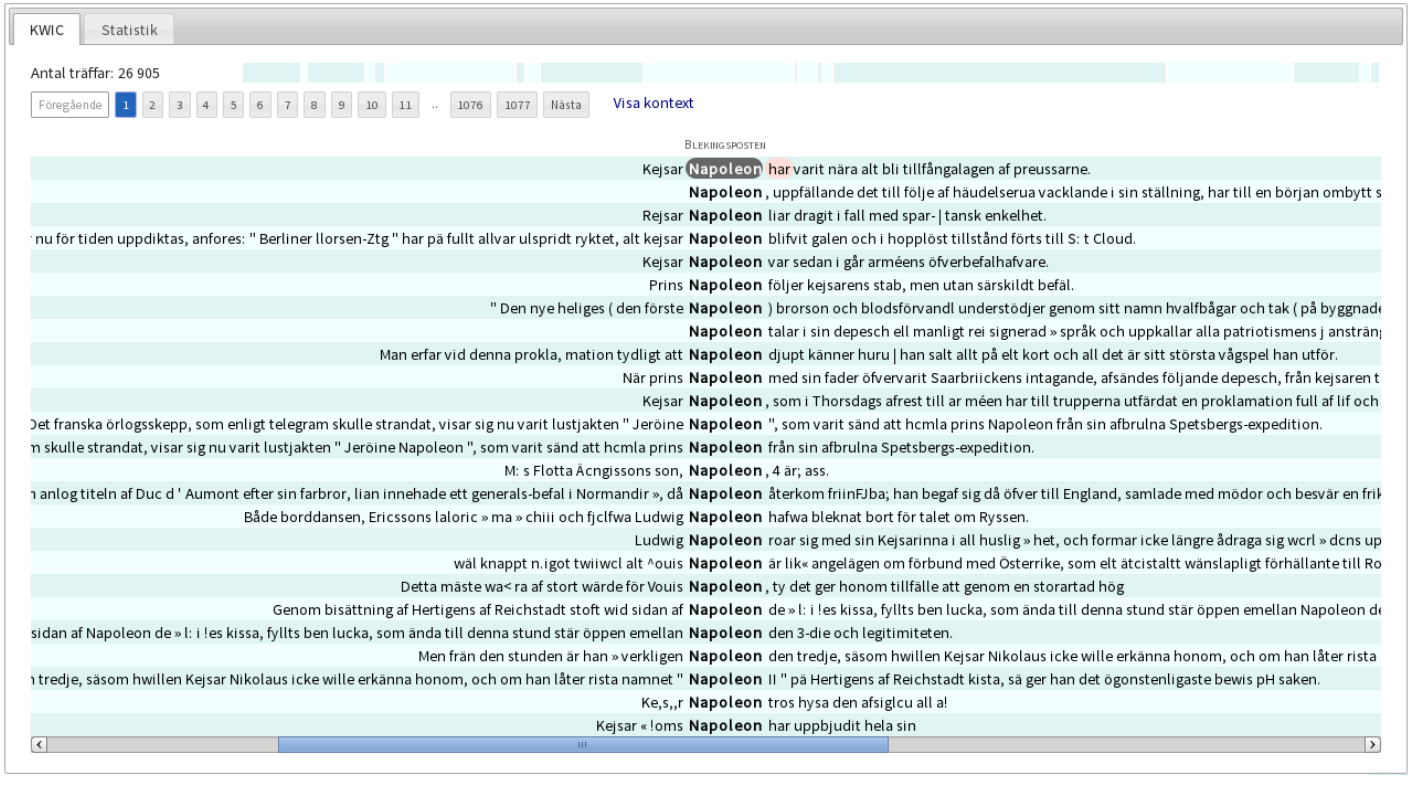
exempel: Tänkeböckerna (under Lagrummet)



GÖTEBORGS UNIVERSITET

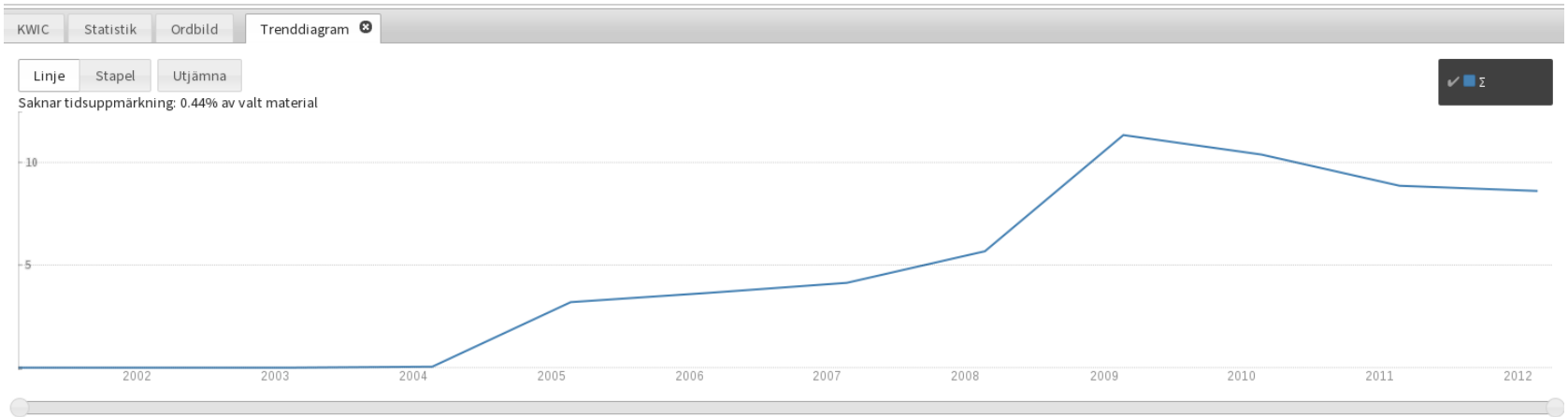
Språk-BANKEN

exempel: Digidaily

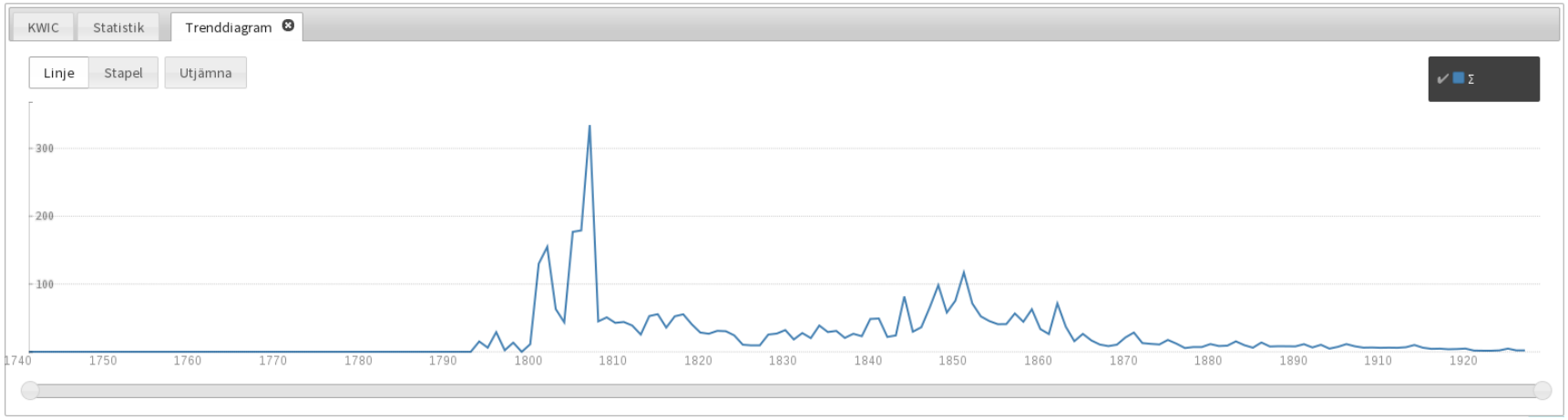


trenddiagram: exempel på en neologism

- 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**