



# Vad är en korpus och vad kan man använda den till?

Lars Borin

Språkdata, Inst. för svenska språket

Göteborgs universitet



# vad är en korpus?

- korpusar är (stora) textsamlingar (typiskt tiotals miljoner ord),
- sammanställda
- och annoterade
- med ett syfte i åtanke,
- (för vilket man utvecklar [dator]verktyg)



# sammansatta ...

- korpussammansättning är precis som opinionsundersökningar:
  - man tar ett representativt och tillräckligt stort stickprov/urval ur en väldefinierad population
  - för att kunna ställa frågor och få svar som ger (statistiskt) signifikant information om populationen



## ... och annoterade

- strukturmärkning och textmetadata
- 'ordklass'-taggning
- lemmatisering
- syntexanalys ('trädbanker')
- länkning (av parallellkorpora)
- länkning (av modaliteter)
- textbindning, dialogakter, m.m.



# korpusklassificering

- tal ~ skrift
- balanserade
- flerspråkiga: parallella ~ jämförbara
- inlärarkorpusar
- 'domänkorpusar' (fackspråk, m.m.)
- monitorkorpusar
- små ~ stora

# Kinnaurikorpusen

\ref 07/007a/01

\tx @ma r@N boa lo shigyO //

\mrep @ma r@N bOba lo -sh-i -gyO

\gl mother with father say-?-?-D.PST

\tr Mother and father said:

\ref 07/007a/02

\tx jO tshEtsats-u nam@N ch@ tate //

\mrep jO tshEtsats-u nam@N ch@d ta -te

\gl this girl-POSS name(N) what keep-LET'S

\tr "what should we name this girl?"

\ref 07/007a/03

\tx nam@N t@ sOthlets tate //

\mrep nam@N t@ sOthlets ta -te

\gl name(N) EMP name keep-LET'S

\tr Let's keep the name (=name her) Sothlets."

# Stockholm Umeå Corpus: textmetadata

```
<tei.2>
<teiheader id=h.kl01>
<fileDesc>
<titleStmt>
<title>suc-kl01</title>
</titleStmt>
<extent words=2001>2001 word
  tokens</extent>
<publicationStmt>
<distributor>SUC</distributor>
</publicationStmt>
<sourceDesc>
<listBibl id="file-kl01">
<biblFull id="bib-kl01">
<titleStmt>
  <title level=m>Akilles hä</title>
  <author>Jan Mårtenson</author>
</titleStmt>
<extent words=2001>pp 99-106</extent>
```

```
<publicationStmt>
  <publisher>Wahlström &
    Widstrand</publisher>
  <pubPlace>Stockholm</pubPlace>
  <idno type="isbn">91-46-15872-3</idno>
  <date>1990</date>
</publicationStmt>
</biblFull>
</listBibl>
</sourceDesc>
</fileDesc>
<profileDesc>
<textclass>
<catRef target=SUC.KL>
</textclass>
</profileDesc>
</teiheader>
```

# Stockholm Umeå Corpus: SUC-taggar

```
<text id=k101>
<body>
<p>
<s id=k101-001>
<d n=1>-<ana><ps>MID<b>-</d>
<w n=2>Vilka<ana><ps>HD<m>UTR/NEU PLU IND<b>vilken</w>
<w n=3>djävla<ana><ps>JJ<m>POS UTR/NEU SIN/PLU IND/DEF...
<w n=4>optimister<ana><ps>NN<m>UTR PLU IND NOM<b>opti...
<d n=5>,<ana><ps>MID<b>,</d>
<w n=6>frustade<ana><ps>VB<m>PRT AKT<b>frusta</w>
<name type=person>
<w n=7>Lasse<ana><ps>PM<m>NOM<b>Lasse</w>
</name>
<d n=8>.<ana><ps>MAD<b>.</d>
</s>
```

# Stockholm Umeå Corpus: PAROLE-taggar

```
<text id=k101>
<body>
<p>
<s id=k101-001>
<c lem='-' msd='FI' n=1>-</c>
<w lem='vilken' msd='DH@0P@S' n=2>Vilka</w>
<w lem='djävla' msd='AQP00N0S' n=3>djävla</w>
<w lem='optimist' msd='NCUPN@IS' n=4>optimister</w>
<c lem=',' msd='FI' n=5>,</c>
<w lem='frusta' msd='V@IIAS' n=6>frustade</w>
<name type=person>
<w lem='Lasse' msd='NP00N@0S' n=7>Lasse</w>
</name>
<c lem='.' msd='FE' n=8>.</c>
</s>
```

# Talbanken: Bruksprosa

P10234081001	0000	<<	GM	086
P10234081002	*DE	PODPHH	SS	086
P10234081003	1000	RC	SSET	086
P10234081004	10002SOM	PORPHH	SS	086
P10234081005	10002TAGITS	VVSN	PAIV	086
P10234081006	10002UT	ABZA	PL	086
P10234081007	10002TILL	PR	OAPR	086
P10234081008	10002UNDEROFFICERSUTBILDNING	VN	SS OA	086
P10234081009	KOMMER	KVPS	FV	086
P10234081010	ATT	ID	IM	086
P10234081011	FÅ	FVIV	IV	086
P10234081012	DET	PODP	OO DT	086
P10234081013	STÖRSTA	AJSU	OO AT	086
P10234081014	ANSVARET	NNDD	OO	086
P10234081015	OCH	++OC	OO++	086
P10234081016	DE	PODP	OO DT	086
P10234081017	SVÅRASTE	AJSU	OO AT	086
P10234081018	UPPGIFTERNA	NNDD	OO	086
P10234081019	AV	PR	OOETPR	086
P10234081020	ALLA	POTP	OOETDT	086
P10234081021	VÄRNPLIKTIGA	AJ	HS OOET	086
P10234081022		IP	IP	086



# korpusar och språkvetenskap

- om inte den empiriskaste, så i alla fall mycket objektiv lingvistik
- enda sättet att undersöka frekvens [se N. Ellis, 'Frequency effects in language processing', *Studies in Second Language Acquisition* 24 (2002): 142-188, + fler i samma nummer]

# konkreta exempel [1]

Konkordanser

Använd korpus:

Press 65

Kontext i tecken:

120 tecken

Kontextbalans:

50%-50%

Antal träffar:

20 träffar

Språk  
BANKEN

Grad på typsnitt:

standard

Söksträng:

elektronhjärn\*

Sök i:

konkordans

frekvens

frekv.tabell

Sök

Återställ

Språkbanken 2003

[Info](#) | [Nyheter](#) | [Språkbanken](#) | [Inst. för svenska språket](#) | [SAOB](#) | [In English](#) | [Kommentera!](#)

Söksträng: **elektronhjärn\*** [Material: **p65**] [Types: **3**] [Tokens: **6**]

oskadliggöra vetenskapsmannen Vonbraun, uppfinnare av den **elektronhjärna** -- Alpha 60 -- som styr staden. + Alphaville  
ch lönsamt sätt. + Till slut en liten kommentar till orden ' **elektronhjärna** ', 'elektronautomatiserad' m. fl., som Mattsso  
väl kunna mäta sig med premiärministern, dennes omvittnade ' **elektronhjärna** ' till trots. + Till hösten kommer Brown att l  
att förmedla, en orationell insikt som den ultrarationella **elektronhjärnan** måste utrota för att kunna fungera. + Han rä  
trött. + Som Svenska Dagbladet mycket riktigt skrev en gång: **Elektronhjärnan** är död -- leve datamaskinen!+  
senare kom ett provisoriskt körtillstånd i awaktan på att **elektronhjärnorna** i Albany skulle sköta om resten av byråkra

SdsKo [Art](#) [Ko](#)

SvdNä [Art](#) [Ko](#)

Dn\_Nä [Art](#) [Ko](#)

SvdKo [Art](#) [Ko](#)

SvdNä [Art](#) [Ko](#)

SvdMä [Art](#) [Ko](#)

# konkreta exempel [2]

## Konkordanser



Använd korpus:

Press 65

Kontext i tecken:

120 tecken

Kontextbalans:

50%-50%

Antal träffar:

20 träffar

Grad på typsnitt:

standard

Söksträng:

järna datamaskin dator

Sök i:

- konkordans
- frekvens
- frekv.tabell

Sök

Återställ

© Språkbanken 2003

[Info](#) | [Nyheter](#) | [Språkbanken](#) | [Inst. för svenska språket](#) | [SADB](#) | [In English](#) | [Kommentera!](#)

Söksträng: **elektronhjärna datamaskin dator** [Types: 3(Visar 1-20)]

p65	p76	dn	p95	p96	p97	p98	svd00	p01	p02	p03	romi	romii	Totalt	Ord
<a href="#">3</a>	-	-	-	-	-	-	<a href="#">1</a>	-	-	-	-	-	4	ELEKTRONHJÄRNA
<a href="#">12</a>	<a href="#">8</a>	<a href="#">5</a>	<a href="#">2</a>	<a href="#">1</a>	-	<a href="#">1</a>	<a href="#">4</a>	<a href="#">2</a>	-	-	<a href="#">1</a>	<a href="#">5</a>	41	DATAMASKIN
-	<a href="#">5</a>	<a href="#">48</a>	<a href="#">159</a>	<a href="#">197</a>	<a href="#">337</a>	<a href="#">290</a>	<a href="#">414</a>	<a href="#">419</a>	<a href="#">394</a>	<a href="#">188</a>	<a href="#">9</a>	<a href="#">7</a>	2467	DATOR

# konkreta exempel [3]

## Konkordanser



Använd korpus:

SVD 00

Kontext i tecken:

120 tecken

Kontextbalans:

50%-50%

Antal träffar:

20 träffar

Grad på typsnitt:

standard

Söksträng:

han hon

Sök i:

- konkordans
- frekvens
- frekv.tabell

Sök

Återställ

© Språkbanken 2003

[Info](#) | [Nyheter](#) | [Språkbanken](#) | [Inst. för svenska språket](#) | [SAOB](#) | [In English](#) | [Kommentera!](#)

[Söksträng: **han hon**] [Types: 2(Visar 1-20)]

p65	p76	dn	p95	p96	p97	p98	svd00	p01	p02	p03	romi	romii	Totalt	Ord
<a href="#">6441</a>	<a href="#">7392</a>	<a href="#">25930</a>	<a href="#">38888</a>	<a href="#">33437</a>	<a href="#">68086</a>	<a href="#">52035</a>	<a href="#">72572</a>	<a href="#">79514</a>	<a href="#">96604</a>	<a href="#">43056</a>	<a href="#">112327</a>	<a href="#">56252</a>	692534	HAN
<a href="#">1173</a>	<a href="#">1578</a>	<a href="#">6818</a>	<a href="#">14364</a>	<a href="#">12725</a>	<a href="#">26334</a>	<a href="#">18535</a>	<a href="#">26645</a>	<a href="#">32753</a>	<a href="#">43389</a>	<a href="#">20738</a>	<a href="#">70371</a>	<a href="#">40728</a>	316151	HON

[\[Ingångsidan\]](#) | [\[Upp\]](#).

# konkreta exempel [4]

**Språk**  
**BANKEN**

Kontextlängd

50+50 tecken

Sortering

osorterat

Hämta max

1000 träffar

Sortera

även  skiljetecken

Max träfflängd

10 st ord

Visa

20 rader

Sök ".+re" "än" [msd="PF@..S@S"]

Info

<<

>>

Tot : 614

Kvar : 594

s de värsta skulderna var betalda . Stockholm var **dyrare än hon**  
isar på precis samma villkor . Anita har det inte **sänre än vi**  
te bryta den mulna tystnaden i bilen . Hon jobbar **bättre än jag**  
nte jag , sa Ann-Charlotte bestämt . Du är mycket **friare än jag**  
vad beträffar politiken så förstår han den mycket **bättre än jag**  
r du varit så länge ? frågade han och rösten blev **skarpare än han**  
berätta fantastiska historier . Trots att jag var **äldre än han**  
sjukhusskjortan . De hade väntat på honom mycket **längre än vi**  
om jag umgicks med lite grand fast hon var mycket **äldre än jag**  
grep jag hur svårt hon måste ha haft det , mycket **svårare än jag**  
också , men jag är inte säker . Hon var ju tre år **äldre än jag**  
e hon varie tiugoförsta mai . Hon var bara tio år **äldre än jag**

tänkt sej . Ändå knappade hon in på maten utom när  
och förstår jag hennes planer rätt har hon det sna  
trodde så liten och klen som hon är och så fint kl  
och mycket yngre . Jag skulle inte kunna leva så .  
. Mycket bättre ! Men jag menar inget illa med det  
menat . --- På stan , svarade hon kort och hon vän  
och högskoleutbildad tyckte jag ofta att jag inte  
andra . Han kunde vara välvillig och charmerande ,  
. Engla var sjukvårdsbiträde , en fränskild barnlö  
som hatade henne för att hon skämde ut mej i skola  
och räknade sej som halvt vuxen . Vi pratade inte  
Eftersom hon var i konstant behov av tröst och m

# konkreta exempel [5]

**Språk-**  
**BANKEN**

Kontextlängd

50+50 tecken

Sortering

osorterat

Hämta max

1000 träffar

Sortera

även  skiljetecken

Max träfflängd

10 st ord

Visa

20 rader

Sök ".+re" "än" [msd="PF@..O@S"]

Info

<<

>>

Tot : 37

Kvar : 17

minne medan däremot Sven Erik , som är så mycket **yngre än oss**  
r alltid varit sån . Vet du att hon är fjorton år **äldre än mej**  
pojkar och två flickor . Min syster är tretton år **äldre än mej**  
rid , denna kvinna som jag sätter högst av alla , **högre än dem**  
om skulle ha otaliga stridsefanter , större och **farligare än dem**  
at , vi tar ned underskotten . Och vår politik är **rättvisare än er**  
entorna minst lika bra som de betygsintagna , och **bättre än dem**  
artyg som går i trafik på Medelhavet är äldre och **mindre än dem**  
artyg som går i trafik på Medelhavet är äldre och **mindre än dem**  
aste är inte att vinna , han har hållit på mycket **längre än mig**  
aste är inte att vinna , han har hållit på mycket **längre än mig**  
ter . Han sade bara : - Korta spelare har i regel **lättare än oss**

båda två , kan svara på hur svåra frågor som helst  
. Vi har samma mamma , men Gudruns pappa gick ifrå  
 , så vi har aldrig haft särskilt mycket kontakt .  
jag delat bädd med , var outröttlig som en bäver  
de dittills stött på . Vid Chenabfloden , som de k  
. De som har bäst ställt får också vara med och be  
som kommit in enbart på högskoleprovet . Studieupp  
i Västindien . promenaddäcket som löper i en oval  
i Västindien . promenaddäcket som löper i en oval  
och skall hyllas lika mycket , tyckte Mika . Att v  
och skall hyllas lika mycket , tyckte Mika . Att v  
långa när det blåser . Jag hade chansen att få hem