

Korpuslingvistik (SV2119)

Föreläsning 1



Richard Johansson

`richard.johansson@svenska.gu.se`

6 september 2013

vad är korpusar och korpuslingvistik?

- ▶ korpus: en samling av datoriserad text
- ▶ korpuslingvistik: att undersöka språkvetenskapliga frågor med hjälp av korpusar

(SAOL sg. *en korpus*, pl. *korpusar*; engelska sg. *corpus*, pl. *corpora*)

- ▶ kurshemsida:
<http://spraakbanken.gu.se/personal/richard/k12013>
- ▶ 7 föreläsningar klockan 10.15 varannan vecka
- ▶ föreläsningarna i L308 eller K235
- ▶ datorlaboration i november (datum meddelas senare)

- ▶ kursbok: Wynne, M. (ed). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford, 2005.

<http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm>

- ▶ artiklar länkas från kurshemsidan

uppgifter

- ▶ två kursuppgifter:
 - ▶ sökningar i Språkbankens korpussamlingar (oktober)
 - ▶ användning av syntaktiska korpusar (november)
- ▶ lös uppgifterna individuellt, skriv kort rapport
- ▶ instruktioner till uppgifterna länkas från hemsidan



projektuppgift

- ▶ välj ett korpusrelaterat ämne av eget forskningsintresse
- ▶ formulera en frågeställning
- ▶ gör en undersökning individuellt eller i par
- ▶ skriv en uppsats (december)
- ▶ presentera den inför kurskamraterna (december)



1. Översikt



vad är det här?

- ▶ korpusar är **samlingar** med autentiska texter,
 - ▶ **utvalda**,
 - ▶ **datoriserade**,
 - ▶ och **annoterade** (lingvistiskt analyserade),
 - ▶ med ett **syfte** i åtanke,
- ▶ korpuslingvistik: undersöka språkvetenskapliga frågor med hjälp av korpusar
 - ▶ metoder
 - ▶ verktyg

(stora) textsamlingar ...

- ▶ text =
 - ▶ skriftspråk,
 - ▶ (transkriberat) talspråk
 - ▶ ...
- ▶ typiskt miljoner ord:
 - ▶ SUC (Stockholm Umeå Corpus): 1 miljon ord [publicerat skriftspråk]
 - ▶ Språkbankens moderna korpustexter >1 miljard ord, mest skriftspråk
 - ▶ BNC (British National Corpus) och andra “nationella korpusar”
~100 miljoner ord [skrift/talspråk]

... utvalda ...

- ▶ korpussammanställning är 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 ...

- ▶ “metadata” på olika nivåer
 - ▶ tid
 - ▶ författare (och information om författaren)
 - ▶ kategori
 - ▶ ...
- ▶ strukturmärkning
 - ▶ morfologisk analys och disambiguering / “ordklass-tagging”
 - ▶ syntaxanalys (“trädbanker”)
 - ▶ länkning (av parallellkorpora)
 - ▶ länkning (av modaliteter)
 - ▶ språkfel
 - ▶ dialogakter
 - ▶ ...

... med ett syfte i åtanke

- ▶ korpusar skapas alltid med ett syfte i åtanke, fast syftet kan vara vitt och vagt
- ▶ korpusarna kan ofta vara användbara för andra syften också
- ▶ syften:
 - ▶ språkvetenskapligt: empirisk lingvistisk forskning (belägg, frekvens)
 - ▶ språkteknologiskt: utveckling och utvärdering av automatiska datorprogram som analyserar språk
 - ▶ kulturvårdande: bevarande och tillgängliggörande av äldre texter, döda eller döende språk

så: varför korpuslingvistik?

- ▶ om inte den empiriskaste, så i alla fall mycket objektiv lingvistik
- ▶ enda sättet att undersöka frekvens (se N. Ellis, Frequency effects in lg processing, Studies in SLA 24 (2002): 143-188, + fler i samma nummer)

men Google då?

- ▶ det finns ingen större korpus (åtminstone för vissa sorters texter) än webben,
- ▶ ... men en grundprincip i all slags forskning är reproducerbarhet,
- ▶ ... och Googlesökningar är inte reproducerbara



varför inte Google?

- ▶ Kilgarriff ("Googleology is bad science", 2007)
 - ▶ resultaten inte konsekventa (webben förändras)
 - ▶ träffarna är antal dokument, inte antal förekomster
 - ▶ Google är inte avsett för språkliga undersökningar, och sökmotorn gör hemliga och oförutsägbara saker med sökorden (t.ex. normalisering)
 - ▶ vi vet inget om urvalet

varför Google?

- ▶ man kan söka efter belägg (“bevis för förekomsten”) på
 - ▶ ord
 - ▶ ordformer
 - ▶ (fasta) fraser
- ▶ man kan få en informell uppfattning om den relativa brukligheten av alternativa sätt att uttrycka samma sak (jfr Lingvistbloggen och Language Log)

kort korpushistoria

- ▶ Index Thomisticus (1946–2005)
 - ▶ <http://www.corpusthomisticum.org>
- ▶ Brown Corpus (1967)
- ▶ Press-65 (1970)
- ▶ Talbanken / MAMBA (1976–78)
- ▶ LOB, London-Lund (1978, 1980)
- ▶ HANSARD (~1990)
- ▶ BNC, Penn Treebank (1995)
- ▶ SUC (1996)
- ▶ “Web as Corpus” (~2003) (SweWAC 2010)



2. Typer av korpusar



typer av korpusar

- ▶ modalitet:
 - ▶ skrivet, talat, tal, tecknat, multimodal
- ▶ språktyp, genre, etc.
- ▶ språk:
 - ▶ ett, två, många; relation mellan språken (parallell, jämförbar)
- ▶ storlek
- ▶ typ av annotering: ingen, morfologisk, syntaktisk, ...

några korpusar, exempel: typisk blandad

- ▶ SUC (allmänsvenskt skriftspråk):
 - ▶ 500 texter om c:a 2000 ord
 - ▶ 9 huvudgenrer, med undergenrer:
 - ▶ K – imaginative prose
 - ▶ KK – general fiction
 - ▶ KL – science fiction and mystery
 - ▶ KN – light reading
 - ▶ KR – humour
- ▶ Brown corpus (Francis & Kučera, 1964)

några korpusar, exempel: begränsad mängd

- ▶ Skriven finsk romani (Borin et al):
 - ▶ c:a 110.000 ord
 - ▶ en betydande del av den totala skrivna produktionen på finsk romani

några korpusar, exempel: parallella

- ▶ Europarl: <http://www.statmt.org/euoparl>
 - ▶ Europaparlamentsprotokoll 1996–2006
 - ▶ Protokollen översätts till alla medlemsspråk
 - ▶ 20 språk, ~10–50 miljoner ord/språk
 - ▶ Alla språk länkade meningsvis till engelska

EN: Technical requirements for inland waterway vessels (vote)

ET: Siseveelaevade tehnilised nõuded (hääletus)

SV: Tekniska föreskrifter för fartyg i inlandssjöfart (omröstning)

- ▶ Föregångaren: Hansard (engelska–franska)
- ▶ Bibeln (eller delar av den) på 100 språk:
<http://homepages.inf.ed.ac.uk/s0787820/bible/>

några korpusar, exempel: jämförbar

- ▶ Wikipedia (<http://sv.wikipedia.org>, ...)
- ▶ länkad artikel–artikel (287 språk)
- ▶ texterna likartade men inte översatta
- ▶ innehåller också användbar halvstrukturerad information

Sverige, officiellt Konungariket Sverige (info), är ett nordiskt land på Skandinaviska halvön i Nordeuropa. Sverige har landgräns i väst med Norge (svensk-norska gränsen) och i nordost med Finland, samt en fast förbindelse över havet med Danmark via Öresundsbron i sydväst. Med en area på 449 964 km² är Sverige det storleksmässigt femte största landet i Europa och har en befolkning på 9,5 miljoner människor. Sverige har en låg befolkningstäthet med 21 invånare per km² men med en betydligt högre täthet i södra halvan av landet. ...

La Svèsia (449 964 km²; 9 082 995) la xe un Stado del nord de l'Eoropa inte la Penisola Scandinava. La so cavedal la xe Stocolma che La gà 765 044 abitanti. Altre cità importante le xe Göteborg, Uppsala, Malmö e Lund. La confina co la Norveja a nord-ovest e co la Finlandia a est; la xe coligada co la Danimarca traverso del Ponte del Øresund. La xe bagnà dal Mar Baltego. La ga na densità demografega pitosto bassa e despensada iregolarmente. El so teritorio el xe sior de legne, fero e aqua. I Svedesi i gode de un bon livel de vita ...



några korpusar, exempel: talat barnspråk

► CHILDES (<http://childes.psy.cmu.edu>)

*NOR: did you wash your hands Jinny ?
*JIN: yeah .
*NOR: you did already ?
*JIN: already .
*NOR: you forgot the candy off from around your mouth huh ?
<n is wiping J's face>
*JIN: yeah .
*NOR: <laughs> okay <% N goes back to sit down> .
*JIN: do I look dirty ?



några korpusar, exempel: inlärarspråk

- ▶ ICLE (International Corpus of Learner English)
- ▶ ASU (Andraspråkets StrukturUtveckling)
- ▶ USE (Uppsala Student English corpus)
- ▶ SFI och SSM

korpusar i Språkbanken

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

<http://spraakbanken.gu.se/korp>

- ▶ modern dagstidningstext
- ▶ modern romantext
- ▶ populärvetenskap
- ▶ sociala medier (bloggar, twitter)
- ▶ 1800-talslitteratur (Litteraturbanken)
- ▶ medeltida text (fornsvenska)
- ▶ färöisk textkorpus (dagstidningstext)
- ▶ parallella korpusar (skönlitteratur)
- ▶ inlärarkorpusar
- ▶ ... och en hel rad andra

3. Urval av korpusar



- ▶ TOP-DOWN:
 - ▶ syfte
 - ▶ urval (vad? hur mycket?)
- ▶ BOTTOM-UP:
 - ▶ varifrån?
 - ▶ hur?
- ▶ undersökning av (eller undervisning om/i):
 - ▶ allmänspråk → stora balanserade korpusar
 - ▶ genre/delspråk → små (specialiserade) korpusar
 - ▶ inlärarspråk → inlärarkorpusar
 - ▶ översättning/språkkontakt/språktypologi → flerspråkiga korpusar

korpusurval och -representativitet

- ▶ en korpus är ett urval
- ▶ urvalet är **representativt** om det som gäller för urvalet också gäller i allmänhet
 - ▶ är Göteborgsposten representativ för “svenska språket”? För svensk tidningstext?
- ▶ omöjligt att objektivet avgöra om en korpus är representativ
 - ▶ Clear (“Corpus sampling”, 1992)
 - ▶ Biber (“Representativeness in corpus design”, 1993)

exempel, BNC:s skriftspråksdel

<http://www.natcorp.ox.ac.uk>

DOMAIN	%	TIME	%
Imaginative	21.91	1960-74	2.26
Arts	8.08	1975-93	89.23
Belief and thought	3.40	Unclassified	8.49
Commerce/finance	7.93	MEDIUM	%
Leisure	11.13	Book	58.58
Natural/pure science	4.18	Periodical	31.08
Applied science	8.21	Misc. published	4.38
Social science	14.80	Misc. unpublished	4.00
World affairs	18.39	To-be-spoken	1.52
Unclassified	1.93	Unclassified	0.40

exempel, BNC:s talspråksdel

REGION	%	CONTEXT-GOVERNED	%
South	45.61	Educational	20.56
Midlands	23.33	Business	21.47
North	25.43	Institutional	21.86
Unclassified	5.61	Leisure	23.71
		Unclassified	12.38
INTERACTION	%		
Monologue	18.64		
Dialogue	74.87		
Unclassified	6.48		

tillgänglighet

- ▶ hur mycket material kan vi komma åt? (t.ex. finsk romani, runsvenska)
- ▶ licenser och copyright...
 - ▶ publicerat på nätet $\neq$ fritt tillgängligt!
 - ▶ det finns en risk att ditt arbete blir oanvändbart! (t.ex. ESPC)
 - ▶ Språkbankens lösning: korpusarna kan laddas ned gratis, men meningarna är omkastade.

4. Korpusbearbetning



att få tag i texten och göra den användbar

- ▶ vi vill ha “äkta” texter (korpuslingvistik är ju empirisk), men
- ▶ äkta texter är fulla av oväsentligheter som vi vill ta bort (men bara oväsentligheterna!), så
- ▶ korpusinsamling och -beredning innebär abstraktion
 - ▶ hur mycket?

kodning av tecken

- ▶ datorn lagrar text som en följd av symboler (siffror)
- ▶ symbolerna är specificerade i en fix teckenuppsättning
- ▶ teckenuppsättningen är numrerad och positionerna kallas *kodpunkter* (*codepoints*)
 - ▶ för länge sedan: 128 symboler (ASCII)
 - ▶ därefter: 256 symboler (Latin-1, ...)
 - ▶ numera: obegränsat antal symboler (Unicode)
- ▶ t.ex. texten *lök* motsvaras av Unicode-symbolerna med kodpunkterna 108, 246, 107.

vad är bokstäver egentligen?

- ▶ är ligaturen *fi* en bokstav?
- ▶ bokstavsbegreppet ännu mer komplicerat för en del icke-latinska skriftsystem
 - ▶ arabiska
 - ▶ Devanagari och dess släktingar
 - ▶ Hangul
- ▶ Unicode-standarden definierar också regler för rendering (uppritning) av bokstavskombinationer

exempel: arabiska

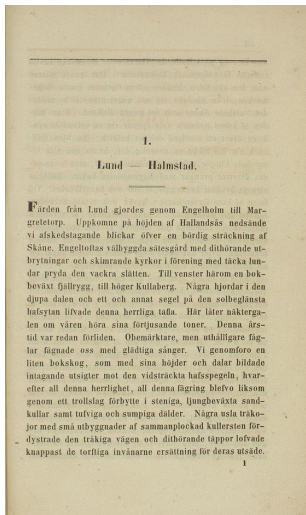
- ▶ arabiska bokstäver renderas (ritas) olika beroende på position i ordet
- ▶ från höger till vänster!
- ▶ men det är ändå samma symboler, och i datorns minne finns inget höger/vänster

kaf, teh, 'alef, beh → كتاب

digitalisering...



från bild till text: OCR



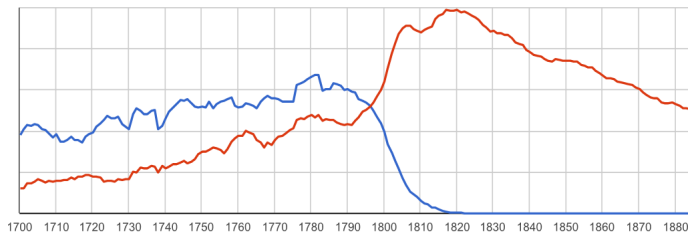
1. Lund – Halmstad.

Färden från Lund gjordes genom Engelholm till Margretetorp. Uppkomne på höjden af Hallandsås nedsände vi afskedstagande blickar öfver en bördig sträckning af Skåne. Engeltoftas välbyggda sätesgård med dithörande utbrytningar och skimrande kyrkor i förening med täcka lundar pryda den vackra slätten. ...

när OCR ställer till det

- ▶ Google ngram viewer: <http://books.google.com/ngrams>

■ funk ■ sunk



Stone House Day.

Showing several of America's oldest private homes

<http://languagelog.ldc.upenn.edu/n11/?p=2848>

- “boilerplate removal” <http://code.google.com/p/boilerpipe>

TORSDAG 5 SEPTEMBER 2013 Arets Dagstidning digitalt. Arets Nyhetssajt.

AFTONBLADET

Min Ekonomi



Skapa konto

Nyheter Sport Nöje TV **Ekonomi** Väder Ledare Kultur Plus A-Ö

Min Ekonomi Shopping **Bredband/telefoni** Boende Spara/Pension El Försäkringar Lån Familj

KRÖNIKÖRER



Annika
Creutzer

• Tipsen - så samlar ni in
pengar till skolresan

TIPSA MIN EKONOMI!!

SENASTE NYTT

BREDBAND/TELEF

Telefonsäljarnas nya
fuskmeter 09:26

Startsidan / Min Ekonomi / Bredband/telefoni

2013-09-05

Telefonsäljarnas nya fuskmeter

”Slamming” ger jätteökning i antalet klagomål - Så
skyddar du dig

Felaktig anslutning (slamming)

160

PERSONER



SKYDDA SEMEST
ERBJUDANDE PÅ MAXA
VÄGASSISTANS HÄR >>

5. Information om texten och dess egenskaper



metadata och annotering

- ▶ för att göra intressantare analyser behöver vi mer information än bara texten i sig

metadata: information om texten

- ▶ information om dokumentet i stort:
 - ▶ språk
 - ▶ tillkomst- eller publiceringstid
 - ▶ författaren:
 - ▶ modersmål och hemort
 - ▶ inlärnivå
 - ▶ kön
 - ▶ ålder
 - ▶ genre
 - ▶ originalets modalitet (skrivet? talat?)
 - ▶ klassificering, t.ex. enligt SAB eller enligt en svårighetsskala
- ▶ information om delar av dokumentet:
 - ▶ stil- och typsnittsinformation
 - ▶ skadade / oläsliga partier
 - ▶ språk i delar av texten
 - ▶ rättning av stavfel

indelning av texten

- ▶ texten delas in i lingvistiskt mer intressanta delar
 - ▶ meningar
 - ▶ ord
- ▶ detta görs vanligen automatiskt eftersom det är för tråkigt att göra det för hand

indelning av texten: svåra fall

- ▶ punkter kan orsaka problem

“... an account with the **U.S. Treasury** to buy Savings Bonds online ...”

“... then I went back to the **U.S. My** dad and I moved ...”

- ▶ avstavade ord: ska vi ta bort bindestrecket?
- ▶ ordindelning i t.ex. kinesiska är inte enkelt att göra automatiskt

民主
min-zhu
people-dominate
“democracy”



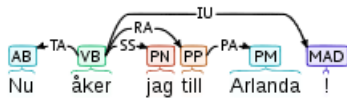
江泽民 主席
jiang-ze-min zhu-xi
... - ... - people dominate-podium
“President Jiang Zemin”



exemplet lånat av Liang Huang

lingvistisk annotering

- ▶ vi kan tillföra lingvistisk information, t.ex. morfologisk eller syntaktisk analys

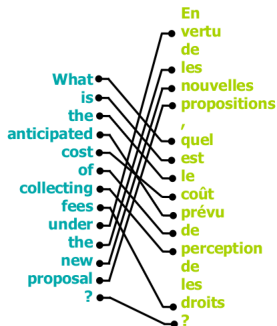


- ▶ manuellt för kvalitet, automatiskt för kvantitet
 - ▶ mer om detta på föreläsning 3 och 5

länkning av ord i parallelltext

What is the anticipated
cost of collecting fees
under the new proposal?

En vertu des nouvelles
propositions, quel est le
coût prévu de perception
des droits?



6. Sökning och statistik



konkordanser

- ▶ t.ex. Korp (föreläsning 2, uppgift 1), AntConc, Wordsmith
- ▶ lista över de sammanhang där ett ord förekommer

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 stai
s ganska experimentell och inte längre så sk (r) amlig	rock	.
	Rock	mot våld i Högsättrahallen 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

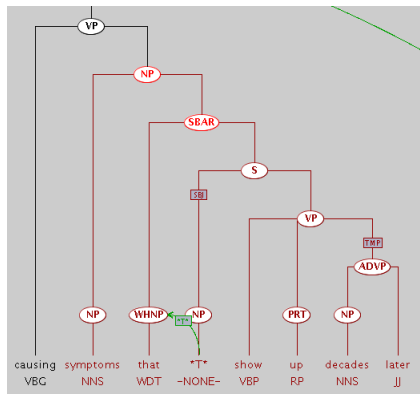
- ▶ pionjären: Hugues de St. Cher (d. 1263) och 500 munkar

syfte med konkordanser

- ▶ översikt av ett ords betydelser
- ▶ idiom
- ▶ vanliga samförekomster
- ▶ “typisk” användning, t.ex. för ordböcker

syntaktiska sökverktyg

- T.ex TIGERSearch (föreläsning 5, uppgift 2)



- ▶ frekvenser
 - ▶ jämför bruklighet av olika ord eller uttryck
 - ▶ utveckling över tid (neologismer, nya betydelser)
 - ▶ ord, fraser, syntaktiska konstruktioner, ...
- ▶ samförekomster
 - ▶ vilka ord brukar förekomma nära ordet *jazz*?
 - ▶ vilka verb tar substantivet *tårta* som objekt?
 - ▶ hur brukar ordet *case* översättas från engelska till svenska?

nästa föreläsning (20 september)

- ▶ Språkbankens korpusar
- ▶ sökverktyget Korp
- ▶ tag gärna med egen laptop!
- ▶ OBS! hålls i **K235**!

