

729G49 Språk och datorer (VT2024)

# Kursintroduktion

Marcel Bollmann & Jenny Kunz

Institutionen för datavetenskap (IDA)



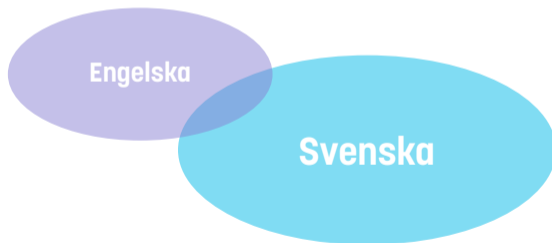
Baserat på material från Marco Kuhlmann,  
Arne Jönsson, och Evelina Rennes.

# Denna föreläsning

1. Lärandemål och organisation
2. Användningsområde korpuslingvistik
3. Användningsområde språkteknologi
4. Programmeringstekniker för textdata

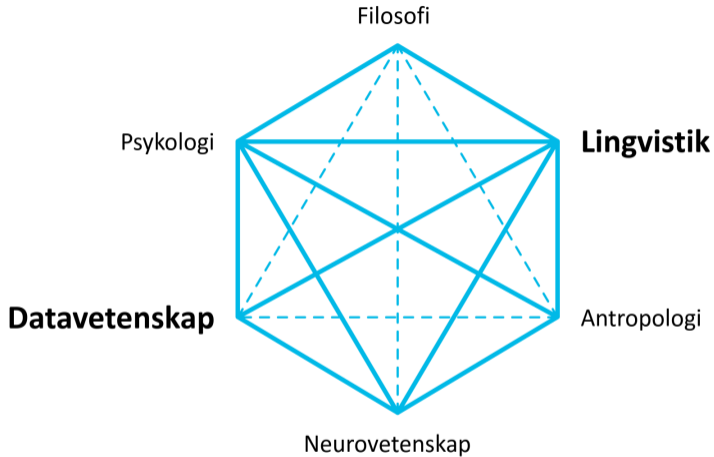
## Språk i kursen

- Föreläsningar om lingvistik kommer att vara på **engelska!**
  - Material från 2023 (på svenska) finns på webbsidan och är fortfarande aktuellt.
- Allt annat sker på **svenska.**
  - seminarier, duggan, föreläsningar om språkteknologi, rapportskrivning, ...



# Lärandemål och organisation





Källa: [Cognitive Science 1978](#)

## Lärandemål

- Redogöra för **grundläggande begrepp inom språkvetenskap**, särskilt morfologi, syntax och semantik
- Utföra **grundläggande språkvetenskapliga analyser** såsom ordklassbestämning och dependensanalys
- Använda **datorbaserade verktyg och programmering** för att samla in, analysera och validera språkliga datamängder (korpora)
- Planera och utföra **enklare korpusundersökningar**, samt redovisa och värdera resultaten

## Lärare på kursen

**Marcel Bollmann** — ✉ [marcel.bollmann@liu.se](mailto:marcel.bollmann@liu.se)

- examinator, kursansvarig, lingvistiklärare

**Jenny Kunz** — ✉ [jenny.kunz@liu.se](mailto:jenny.kunz@liu.se)

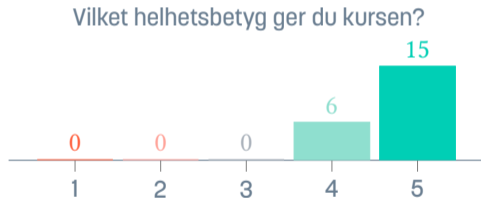
- kursledare, språkteknologilärare, laborationer, fördjupningsuppgift

**Daniel Holmer** — ✉ [daniel.holmer@liu.se](mailto:daniel.holmer@liu.se)

- laborationer, fördjupningsuppgift

## Kursutvärdering 2023: Vad var bra?

- **54** studenter läste kursen 2023, **21** svarade på kursutvärderingen (→ 39%).
- Många studenter uppskattade **seminarierna!**
- Mycket beröm också för Evelinas engagemang (*som tidigare kursansvarig*).





## Kursutvärdering 2023: Van kan vi göra bättre?

- Tydligare instruktioner till labbarna
  - **Åtgärd:** Jenny kommer att ge en översikt och förklaring av labbar under labbföreläsningarna
- Svårighetsgrad på labbar, särskilt VG-delen av L1
  - **Åtgärd:** VG-delen av L1 ändrades lite och är nu konceptuellt enklare
- Två labbar skulle inte schemaläggas på samma dag
  - **Åtgärd:** Inga två labb på samma dag den här gången
- Inte tillräckligt med labbassistenter
  - **Åtgärd:** Vi försöker att ha extra assistenter närvarande under labbarna





All information finns på kurswebbsidan!







[www.ida.liu.se/~729G49](http://www.ida.liu.se/~729G49)

# Kursöversikt





## Lingvistik

-  80 timmar
-  6 föreläsningar,  
5 seminarier
-  Dugga
-  DUG1: U/G

## Språkteknologi

-  80 timmar
-  5 föreläsningar,  
5 laborationer
-  Labbuppgifter
-  LAB1: U/G/VG

## Fördjupningsuppgift

-  80 timmar
-  Introföreläsning,  
2 seminarier
-  Rapport
-  UPG1: U/G/VG

# Laborationer

- Inlämningar via **Lisam**.
- Återkoppling via **Webreg**.
- Anmäl er i Webreg!
  - [www.ida.liu.se/webreg3/729G49-2024-1/LAB1](http://www.ida.liu.se/webreg3/729G49-2024-1/LAB1)  
eller använd QR-koden



# Frågor?

Dugga, lingvistikdel,  
administrativa frågor



Marcel Bollmann

✉ [marcel.bollmann@liu.se](mailto:marcel.bollmann@liu.se)

Labbar, fördjupningsuppgift,  
organisatoriska frågor



Jenny Kunz

✉ [jenny.kunz@liu.se](mailto:jenny.kunz@liu.se)

# Användningsområde korpuslingvistik



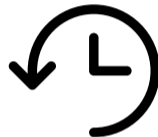
## Vad är korpuslingvistik?

- Korpuslingvistik handlar om att undersöka **språkvetenskapliga frågor** utifrån språkliga **data**
  - s.k. **korpusar**
- Ett grundantagande är att språkvetenskapliga undersökningar bör baseras på **texter från fältet**, i deras naturliga kontext



## Historisk bakgrund

- Sedan mitten av 1900-talet är språkvetenskap i huvudsak en **deskriptiv** vetenskap
  - tidigare ofta **preskriptiv**
- Precis som i andra humaniora kan språkvetenskapliga analyser **grundas i data** och **kvantitativa metoder**





## Historisk bakgrund

- Utvecklingen av kvantitativa och statistiska metoder inom språkvetenskapen **avstannade under 1950-talet**
  - Det fanns **inte tillräckligt mycket språkdata** för att stödja intressanta statistiska generaliseringar
- Humaniora rörde sig bort från data-orienterade beskrivningar mot introspektiva modeller av kognition



“*When measurement departs from theory, it is likely to yield mere numbers, and their very neutrality makes them particularly sterile as a source of remedial suggestions. But numbers register the departure from theory with an authority and finesse that no qualitative technique can duplicate, and that departure is often enough to start a search.*”

— Thomas Kuhn (1961)

# Vad är en korpus?

**SAOL**

publicerad: 2015

**corp-us** [kå̀r`p- el. kår`p-] substantiv ~en  
~ar

1 (språkv.) stor mängd av databehandlade  
texter

2 huvuddel t.ex. av en kanna; mittparti t.ex. i  
altarskåp

3 en medelstor stilgrad

## corpus noun

 Save Word

cor-pus | \ `kôr-pəs \

plural **corpora** \ `kôr-p(ə-)rə \

### Definition of *corpus*

- : the body of a human or animal especially when dead
- a** : the main part or body of a bodily structure or organ  
*// the corpus of the uterus*  
**b** : the main body or corporeal substance of a thing  
*specifically* : the principal of a fund or estate as distinct from income or interest
- a** : all the writings or works of a particular kind or on a particular subject  
*especially* : the complete works of an author  
**b** : a collection or body of knowledge or evidence  
*especially* : a collection of recorded utterances used as a basis for the descriptive analysis of a language

Källor: SAOL & Merriam-Webster

## Korpusar av olika slag

- Generella korpusar
  - problem: representativitet
- Specifika korpusar
  - inriktning mot t.ex. en viss genre, författare, historisk period
- Parallela korpusar
  - samma text i flera språk eller versioner

## Några viktiga korpusar

Namn	År	Storlek	Språk och sort
British National Corpus (BNC)	1994	100 miljoner ord	brittisk engelska, talad och skriven
American National Corpus (ANC)	2003	22 miljoner ord	amerikansk engelska, talad och skriven
Corpus of Contemporary American English (COCA)	2008	1 miljard ord	amerikansk engelska, talad och skriven
Stockholm-Umeå Corpus (SUC)	2006	1 miljon ord	svenska, skriven

## Länkar till korpusar och korpusverktyg

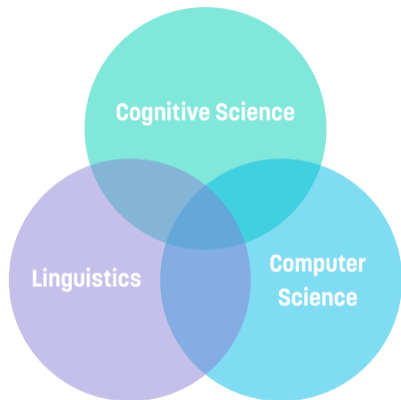
- [🔗 Språkbankens resurser](#) (*svenska*)
- [🔗 Korp, Språkbankens ordforskningsplattform](#) (*sökverktyg*)
- [🔗 Google Books Ngram Viewer](#) (*flera språk, men inte svenska*)
- [🔗 Universal Dependencies Project](#) (*fler än 100 olika språk*)

# Användningsområde språkteknologi



## Vad är språkteknologi?

- Teknologi för **analysera och tolka** naturligt (mänskligt) språk
- **Tvärvetenskapligt** forskningsområde
- **Relaterade områden:**
  - Natural language processing (NLP)
  - Computational linguistics (CL)
  - Speech processing





# Kommerciellt intresse

SONY



Colossal-AI

GTCOM

LIVEPERSON



Google Research

Alibaba Cloud

Megagon Labs

Baidu 百度

ByteDance

amazon | science

Diamond & Platinum sponsors from EMNLP 2023

# Exempel: ChatGPT

ChatGPT 3.5 ▾



**You**

Vad är språkteknologi?



**ChatGPT**

Språkteknologi är ett tvärvetenskapligt område som fokuserar på utvecklingen av datorbaserade system som kan förstå, tolka och generera mänskligt språk. Det omfattar olika teknologier och tillämpningar som syftar till att behandla och hantera språkdata på ett effektivt sätt. Språkteknologi involverar ofta områden som datorlingvistik, maskininlärning, datavetenskap, lingvistik och psykolingvistik.

Exempel på språkteknologi inkluderar:

1. Maskinöversättning: Utveckling av system som kan översätta text eller tal från ett språk till ett annat automatiskt.
2. Taligenkänning: Tekniker som gör det möjligt för datorer att omvandla talad språk till text.
3. Texttolkning: System som kan extrahera meningsfull information från text, till exempel namn, platser, datum och händelser.
4. Automatiskt sammanfattning: Algoritmer som sammanfattar längre texter till kortare versioner med bibehållen relevant information.

Via ChatGPT

## Varför språkteknologi?

“*We are drowning in information but starved for knowledge.*”

— John Naisbitt (1982)

- Vi **förmedlar information** främst genom språk.
- Språk **produceras av & är avsett för människor**, snarare än datorer.
  - s.k. ostrukturerad data
- Språkteknologi kan hjälpa till att extrahera **strukturerad data** från språk.

## En stor utmaning: Ambiguitet

<i>Jag</i>	<i>bad</i>	<i>om</i>	<i>en</i>	<i>kort</i>	<i>bit</i>
<b>PN</b>	<b>VB</b>	<b>PP</b>	<b>DT</b>	<b>JJ</b>	<b>NN</b>
<b>NN</b>	<b>NN</b>	<b>SN</b>	<b>PN</b>	<b>AB</b>	<b>VB</b>
		<b>PL</b>	<b>RG</b>	<b>NN</b>	
		<b>AN</b>	<b>NN</b>		

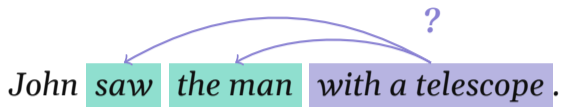
Exempel: Joakim Nivre

## En stor utmaning: Ambiguitet

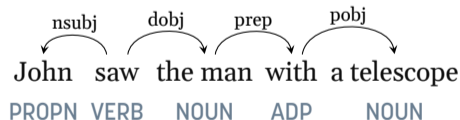
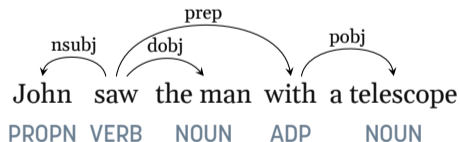
<i>Jag</i>	<i>bad</i>	<i>om</i>	<i>en</i>	<i>kort</i>	<i>bit</i>
<b>PN</b>	<b>VB</b>	<b>PP</b>	<b>DT</b>	<b>JJ</b>	<b>NN</b>
99,45%	80,93%	63,14%	95,26%	70,22%	100%
<b>NN</b>	<b>NN</b>	<b>SN</b>	<b>PN</b>	<b>AB</b>	<b>VB</b>
0,55%	19,61%	29,79%	2,37%	19,66%	0%
		<b>PL</b>	<b>RG</b>	<b>NN</b>	
		5,13%	2,28%	10,11%	
		<b>AN</b>	<b>NN</b>		
		1,92%	0,01%		

Data: Stockholm-Umeå Corpus

# Strukturell ambiguitet



- **Lingvistik analys** kan beskriva de underliggande strukturerna:



# Programmeringstekniker för textdata



# Programmeringstekniker för att hantera text

## **Laboration L0**

(ingen inlämning)



