



UPPSALA  
UNIVERSITET

# Matematik för språkteknologer

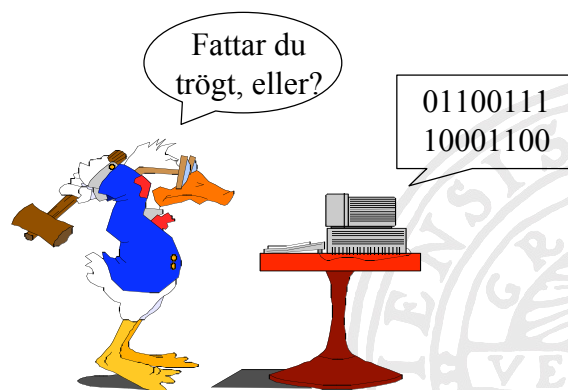
Joakim Nivre

Uppsala universitet  
Institutionen för lingvistik och filologi



UPPSALA  
UNIVERSITET

## Varför matematik?



2



## Matematisk modellering

### Matematiska modeller:

- Formella (exakt definierade)
- Abstrakta (förenklade)
- Användbara (praktisk tillämpning, förståelse)

### Exempel: Informationssökning

- Informationsbehov  $I$ :
  - Teckensträng, t.ex. "matematisk modellering"
- Dokument  $D$ :
  - Teckensträng, t.ex. "I den här artikeln ska vi inte diskutera matematisk modellering ..."
- Relevans:
  - $D$  är relevant för  $I$  om  $I$  förekommer som en delsträng i  $D$ .

3



## Matematik är inte (bara) siffror

### Definition:

- Matematik (av grekiska *máthema*, "vetenskap") är läran om abstrakta kvantiteter, strukturer och mönster. (Wikipedia)

### Matematik för språkteknologer:

- Diskret matematik:
  - Diskreta strukturer och mönster
  - Exempel: Vilka ord **kan** förekomma i kontexten "I den här artikeln ska vi ..."?
- Sannolikhet och statistik:
  - Frekvens och variation
  - Exempel: Vilka ord **brukar** förekomma i kontexten "I den här artikeln ska vi ..."?

4



## Diskret matematik

### Definition:

- Diskret matematik är studiet av matematiska strukturer som är fundamentalt diskreta [dvs.] inte kräver begreppet kontinuitet. (Wikipedia)

### Exempel:

- Elspisar (i motsats till gasspisar).
- Digital termometer (i motsats till analog).
- Heltal i motsats till decimaltal.

### Är språket diskret eller kontinuerligt?

- Ordförråd?
- Språkljud?
- Dialekter?



## Sannolikhet och statistik

### Teorier om osäkerhet:

- Slantsingling: **krona** eller **klave**?
- Språkligt yttrande: **nästa ord**?

### Sannolikhet:

- Slantsingling:  $P(\text{krona}) = P(\text{klave}) = 0.5$
- Språkligt yttrande:  $P(\text{"och"}) = ?$

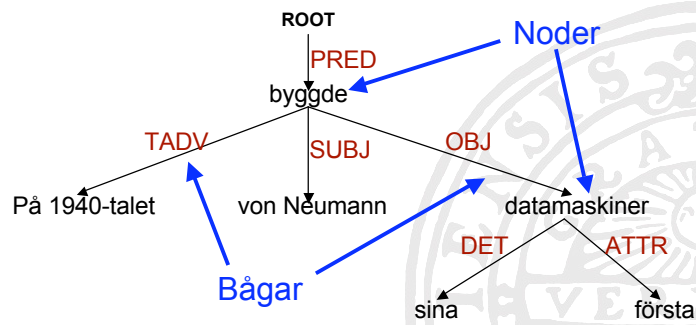
### Statistik:

- Stickprov som grund för bedömningar av sannolikhet.
- Exempel:
  - Ordet **"och"** observerades 30852 gånger i ett stickprov på 1 miljon ord.
  - $P(\text{"och"}) \approx 0,03$



## Exempel 1: Syntaktisk struktur

På 1940-talet byggde von Neumann sina första datamaskiner.

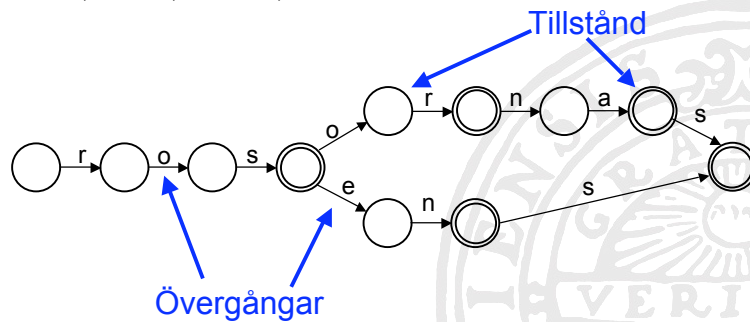


7



## Exempel 2: Morfologisk analys

ros, ros, rosen, rosens  
rosor, rosors, rosorna, rosornas



8

## Exempel 3: Statistisk taggning

- Märka upp alla ord i en text med rätt ordklass

Jag bad om en kort bit.  
 PN VB PP DT JJ NN  
 NN NN PL RG AB VB  
 AB PN NN  
 SN NN

384 möjliga analyser!

9

## Korpusstatistik

- Systematiska observationer av samförekomst av ord och ordklasser.

	AB	DT	JJ	NN	PL	PN	PP	RG	SN	VB
jag				25		4614				
bad				11						43
om	149				377		5010		2373	
en		16340		0		407		397		
kort	37		125	18						
bit				94						0

	DT	JJ	NN	PP	VB
DT	188	21469	23939	277	27
JJ	274	2763	46651	3707	1915
NN	1648	3537	11221	52652	42393
PP	19866	10613	52895	543	312
VB	23442	8935	19589	22581	19810

10



## Tillämpning av statistik

- Beräkning av den mest sannolika sekvensen av ordklasser.

Jag	bad	om	en	kort	bit
NN	NN	AB	DT	AB	NN
PN	VB	PL	NN	JJ	VB
		PP	PN	NN	
		VB	RG		

11



## Generalisering

- Analys av okända ord.

Hon tog fram tre nya **boskar** ur skåpet.

PN	VB	AB	RG	JJ	NN	PP	NN
----	----	----	----	----	----	----	----

Alla bokar **boskar**.

DT	NN	NN
PN	VB	

12



## Kursmoment

### Mängder:

- Mängder av objekt
- Grunden för all (diskret) matematik

### Sannolikhet och statistik:

- Beskrivning och kvantifiering av osäkerhet
- Metoder för datainsamling och generalisering

### Grafer, relationer och funktioner:

- Relationer mellan objekt
- Abstrakta strukturer som mängder + relationer

### Automater:

- Formella beskrivningar av mönster