

# språkvetenskap, lärande språkmodeller och många sorters språk

jussi karlgren

språk i Norden och AI  
november 2024

# generativa språkmodeller

sannolikhetsfördelningar över strängar

generella inlärningsmekanismer för sekvenser

inte helt skraddarsydda för egenheter hos mänskliga språk

saknar kommunikativa färdigheter och intentioner

vi är just nu i ett Eliza-skede

Welcome to

```
EEEEEE LL      IIII  ZZZZZZ  AAAAA  
EE      LL      II     ZZ     AA   AA  
EEEEEE LL      II     ZZZ    AAAAAAA  
EE      LL      II     ZZ     AA   AA  
EEEEEE LLLLLL IIII  ZZZZZZ  AA   AA
```

Eliza is a mock Rogerian psychotherapist.

The original program was described by Joseph Weizenbaum in 1966.

This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?

YOU: Men are all alike.

ELIZA: What is the connection, do you suppose ?

YOU: They're always bugging us about something or other.

ELIZA: Can you think of a specific example ?

YOU: Well, my boyfriend made me come here.

ELIZA: Is it important to you that your boyfriend made you come here ?

YOU: He says I'm depressed much of the time.

ELIZA: I am sorry to hear that you are depressed.

YOU: It's true. I am unhappy.

ELIZA: Can you explain what made you unhappy ?

YOU:

# de funkar ganska bra!

dagens generation av språkmodeller hanterar strängar imponerande väl!

vissa språk med större kompetens än andra

- tillgång på språkiga data
- investeringsvilja och affärsmässiga betingelser
- språkliga egenskaper

vad är språkvetarens och språkvetenskapens uppgift?

i synnerhet i språkområden med färre resurser och färre talare?

hur kan vi hantera språkmodeller?

# är språkmodeller en svart låda?

det är hjärnan också

vi (lingvister och andra språkvetare) studerar språk

utan att bry oss nämnvärt om processorarkitektur och minnesmodeller

men vi bryr oss om perception, produktion och språkbruk

det ska vi fortsätta göra

vad är språkvetarens och språkvetenskapens uppgift?

vår uppgift (en av dem) är att bistå med kunskap om språk  
för att bygga bättre språkmodeller

# påståenden

språk är på det stora hela mycket lika

det finns vissa specifika parametriserbara skillnader

människor och kulturer är på det stora hela mycket lika

det finns vissa specifika delvis parametriserbara skillnader

mänskliga språk utnyttjar inte hela det möjliga parameterområdet för strängföljder

språkmodellens arkitektur tar inte hänsyn till mänskliga språks egenheter

språkmodeller lär sig mycket illa

språkmodeller är tränade att passa vissa kulturella preferenser

det finns mycket som går att göra för att språkmodeller ska bli bättre

det kommer nya språkmodeller hela tiden



# typologiska karakteristika

hur ser världens stora språk ut?

hur ser världens smärre språk ut?

# typologiska karakteristika

hur ser världens stora språk ut?

hur ser världens smärre språk ut?

kan vi anta att modellers inlärningsmekanismer möjligen är optimerade för någondera?

G. änen.	G. änein.
D. änelle.	D. äneille.
A. änen/ ändä.	A. änat/ änejä.
V. äni.	V. änat.
A. änelä/ eſtä/ eſä/ ellä.	A. äneildä/ eifſtä/ eifſä/ eillä.

*Paradigma Tertia Declinationis.*

Sing.	Plur.
N. äiti/ Mater.	N. äite.
G. äitin.	G. äitein.
D. äiteille.	D. äiteille.
A. äitin/ äitiä.	A. äitte/ äitejä.
V. äiti.	V. äite.
A. äiteildä/ eifſtä/ eifſä/ eillä.	A. äiteildä/ eifſtä/ eifſä/ eillä.

*Paradigma Quarta Declinationis.*

Sing.	Plur.
N. aallo/ Unda.	N. aallo/ aalloja.
G. aallon.	G. aalloin.
D. aallole.	D. aalloille.
A. aallon/ aalloa.	A. aallo/ aalloja.
V. aallo.	V. aallo.
A. aalloida/ loſta/ loſſa/ ſa/ ſolla.	A. aalloida/ loifſta/ loifſä/ oilla.

Eskil Petraeus. 1649. *Linguae Finnicæ Brevis Institutio.* Åbo.

lokativa kasus alla hopsamlade under etiketten "ablativ"; båda objektskasus under etiketten "ackusativ"

# hur varseblir lärande språkmodeller språk?

de matas med "raw data"

dvs teckensträngar

som styckas till semantiska enheter genom "tokenisering"

hur är tokeniseringen beskaffad? vad är den optimerad för?  
vilken sorts språk?

# hur analyserar språkmodeller syntax?

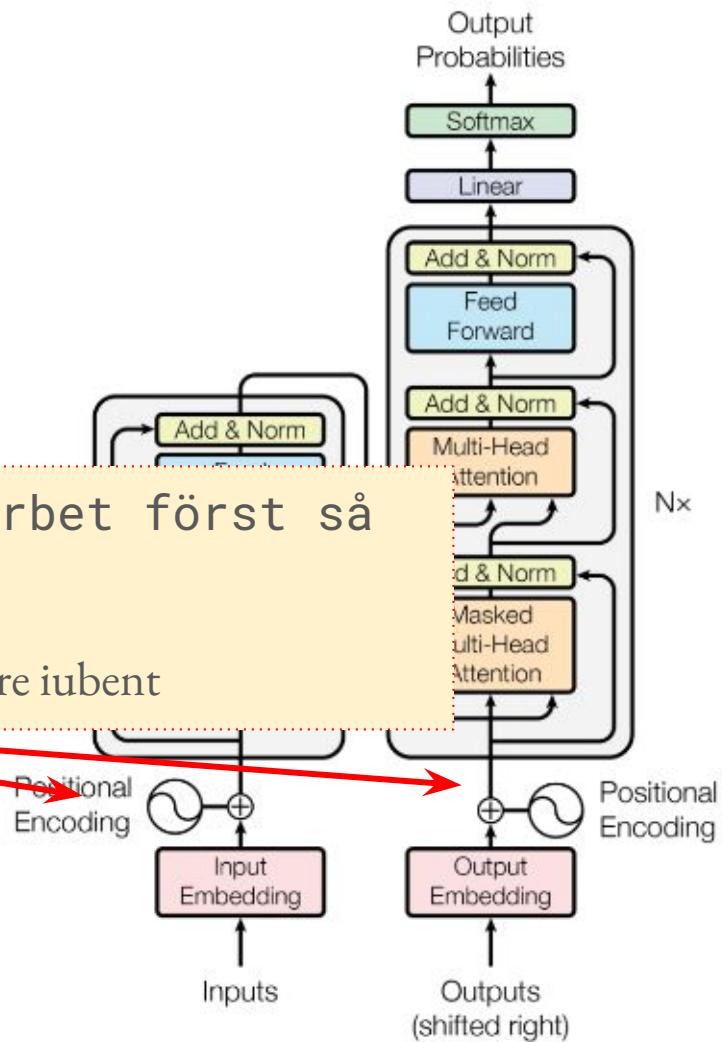
transformern och "attention is all  
you need" är en genial  
processorarkitektur!

förutsätt  
alla strä

kräver "positional encoding"

latinlärarens råd: hitta verbet först så  
löser sig resten

cibaria quemque domo effere iubent



hur är språkmodellens  
inlärningsprocess?

Wikipedia och 1T ord från  
Reddit och Flashback

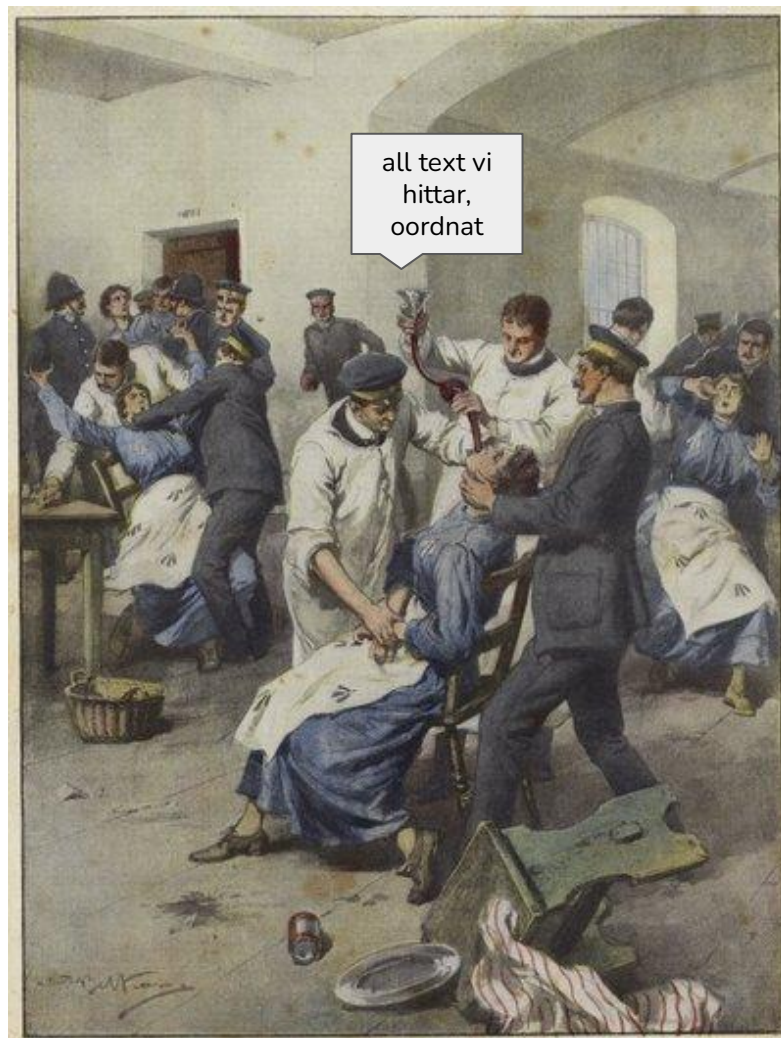
hur är människors  
inlärningsprocess?

börja med enkla byggstenar

bygg på situationer

med rik kontext

ta fasta på anomalier för  
opportunistisk inlärning



vilka byggstenar ser språkmodeller i den språkliga signalen?

vilka byggstenar ser människor i den språkliga signalen?

vilka ordklasser och satsdelar kan vi postulera?

(för att förklara lärande, inte för att modellera hela språket)

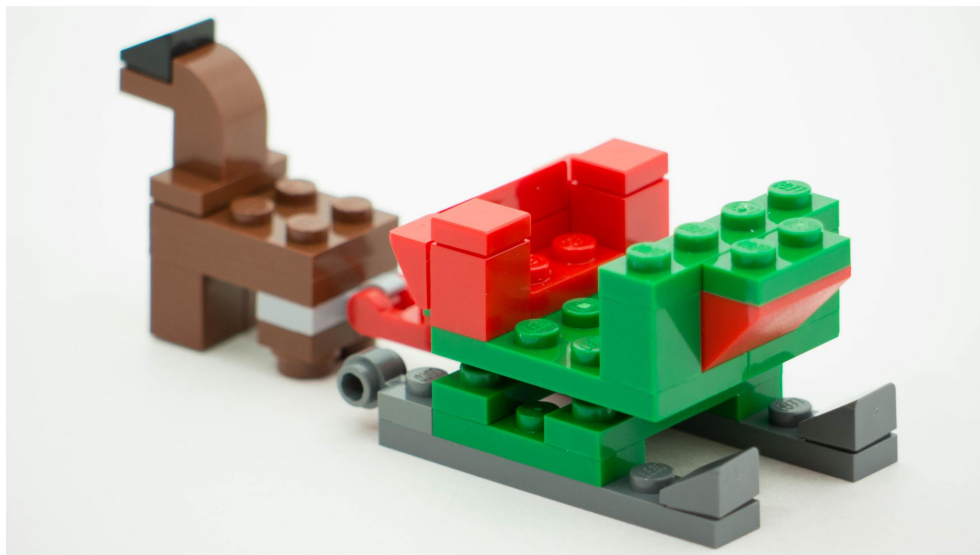


photo: billward@flickr

# hur lätt lär sig språkmodeller språk?

maskinell kurva

få parametrar: dålig modell

fler parametrar: bättre modell!

träningdatamängden växer linjärt med antal parametrar

dagens stora modeller: 10-100B parametrar, 1-10T ord

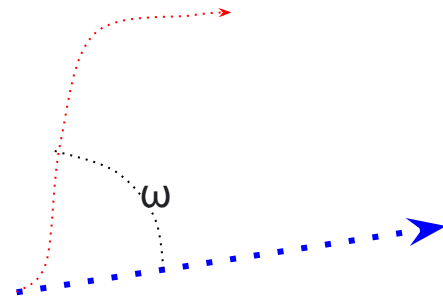
mänsklig kurva

100M-1B hjärnceller med 1-10B synapser i språkperception och språkproduktion

1-10M ord / år under 20 år

varför är skillnaden så stor?

detta ger språkteknologin en forskningsmetodologi! ( $\omega$ )





# hur instruerar vi en språkmodell?

en språkmodell är inte en  
språklig agent

instruktionsträning ger  
en språkmodell en röst

vad finns det för  
kulturella  
förutsättningar i  
instruktionsträningen?

- "Two people are seen sitting before a wave pool and one leads another out onto the water on a board. The person ..."
- "A cowboy rides a horse out of a corral and enters into a fenced off area. The man rides his horse out of the fenced are and throws a rope ..."
- "A small group of people are seen sitting around a casino table speaking to one another and playing a game of poker ..."

# språkteknologien och språkmodellen

språkmodeller är (på väg att bli) goda representationer av lexikon och syntax och språksignalintern semantik

språkmodeller saknar perception, kontextuell förståelse, kommunikativa intentioner och lär sig illa

språkmodeller är byggda synbart neutralt men i själva verket riggade för vissa typologiskt speciella språk och tränade för att passa vissa kulturområden

# språkteknologens uppgift

## **viktigt:**

vilka egenskaper hos språk gör dem lätta för oss att lära?

vilka intentioner och planer styr språkanvändning?

ω!

## **mindre viktigt:**

algoritmer för syntaktisk analys

algoritmer för att lösa strukturella eller semantiska flertydigheter

# språkvetarens uppgift

obs: tänk hållbart -  
det kommer nya  
modeller snart!

formulera en språklig representation som hanterar det enkla först och det sofistikerade sedan (ordklasser, satsdelar, komplexitet)

formulera kunskap om **särskilda språk** på ett sådant sätt att den kunskapen hjälper oss förstå inlärning av språket

beskriva situationella särdrag som driver språkinlärning (t ex anomalier och opportunist)

parametrisera språkliga skillnader; kontrastera mot språkliga generika

parametrisera kulturella språkliga skillnader

beskriva kommunikativa preferenser som styr och begränsar våra val i parameterutrymmet

beskriva språkliga val i ett system av många olika samverkande val och variabler (SFG? :-)