# Maskinoversættelse: Hvordan?

(http://cst.dk/anders/)

UNIVERSITY OF COPENHAGEN

# Hvad jeg vil tale om:

- Hvor kommer maskinoversættelsesfejl fra?
- En smule om funktioner og endelige samples
- En *bittelille* smule om Chomsky
- Frasebaseret maskinoversættelse
- Maskinoversættelse & deep learning
- Hvordan minimerer *I* maskinoversættelsesfejl?

| | |
|---|---|
| 1886 | Salige er de fattige i Ånden, thi Himmeriget er Deres. |
| 1907 | Salige er de fattige i Ånden, thi Himmeriges Rige er Deres. |
| 1926 | Lyksalige er de der føler sig fattige i Ånden, for Himlenes Kongerige er deres. |
| 1948 | Salige er de fattige i ånden, thi Himmeriget er deres. |
| 1959 | Velsignede de, som ikke tiltror sig at have noget at pukke paa overfor Gud - for Guds verden er netop for dem. |
| 1984 | Salige er de magtesløse, deres er Guds rige. |
| 1992 | Salige er de fattige i ånden, for Himmeriget er Deres. |
| 1997 | Salige de fattige i ånden - thi Himlenes Rige er deres. |
| 2007 | Velsignede er de, der erkender deres afhængighed af Gud - for de skal få del i Guds rige. |
| 2007 | I er heldige hvis I er fattige og har fået Helligånden, Guds rige tilhører jer. |

# Maskinoversættelsesfejl

## Nye ord og tvetydigheder


German
French
English
Detect language
Danish
English
Hausa
Translate
Villante looks like a dench Adam Johnson.
Villante ligner en Dench Adam Johnson.
Suggest an edit

*Nye ord*


German
French
English
Detect language
Danish
English
Hausa
Translate
Villante looks like a dench Adam Johnson.
Villante ligner en Dench Adam Johnson.
Suggest an edit

*Nye ord*


JJH @JaackJH · 20h
Villante looks like **a dench** Adam Johnson


*Tvetydighed*

German | French | English | Detect language

I love no case on the iPhone

Danish | English | Hausa | Translate

Jeg elsker intet tilfælde på iPhone

Suggest an edit

*Tvetydighed*

**Łêêk** @leekyleek__ · 2h

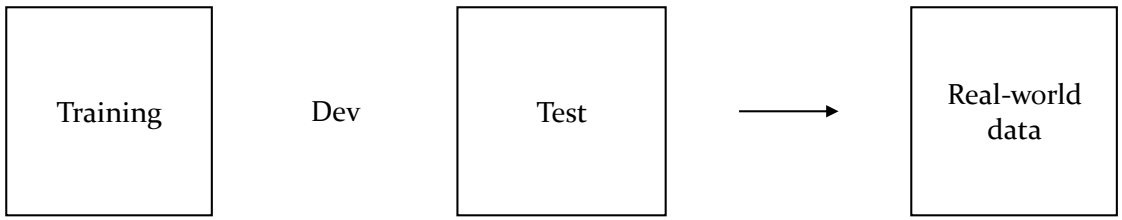
**I love** no case on the **iPhone**

1 1

# Funktioner

| | **x** | *y* |
|---|---|---|
| **POS tagging** | sentence | sequence of labels |
| **NER** | sentence | sequence of labels |
| **Parsing** | sentence | grammatical analysis |
| **Document classification** | document | class |
| **Information retrieval** | query + documents | snippets |
| **Summarization** | text | short text |
| **Machine translation** | sentence | sentence |

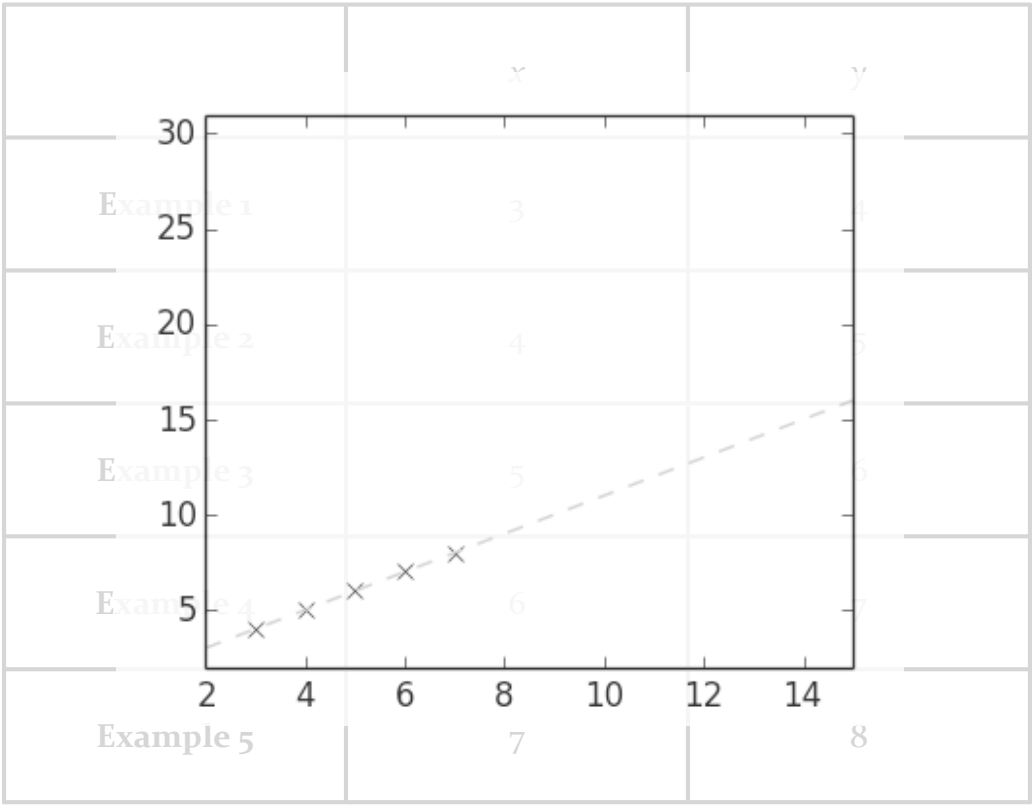
| | **x** | *y* |
|---|---|---|
| **POS tagging** | sentence | sequence of labels |
| **NER** | sentence | sequence of labels |
| **Parsing** | sentence | grammatical analysis |
| **Document classification** | document | class |
| **Information retrieval** | query + documents | snippets |
| **Summarization** | text | short text |
| **Machine translation** | sentence | sentence |



| | **x** | *y* |
|---|---|---|
| **Example 1** | 3 | 4 |
| **Example 2** | 4 | 5 |
| **Example 3** | 5 | 6 |
| **Example 4** | 6 | 7 |
| **Example 5** | 7 | 8 |

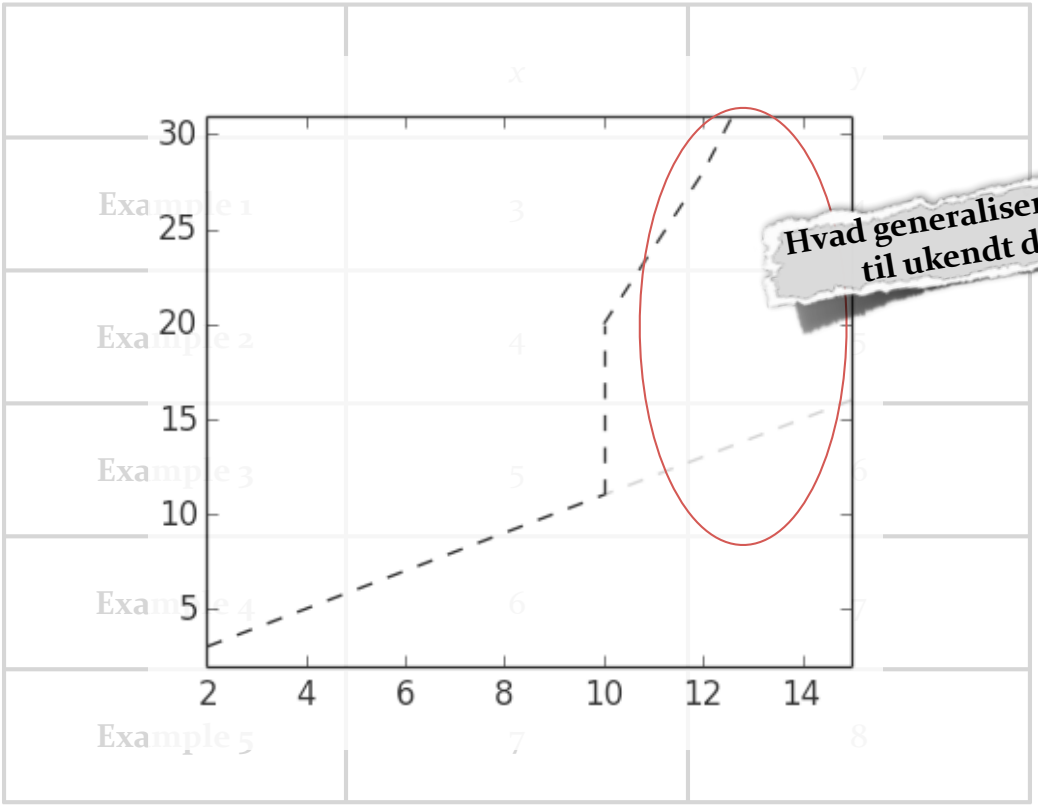
a) $y=x+1$

b) if $x<10$, $y=x+1$, else $y=x^2/20$




a) $y=x+1$

b) if $x<10$, $y=x+1$, else $y=x^2/20$




Hvad generaliserer bedst til ukendt data?

a) $y=x+1$

b) if $x<18$, $y=x+1$, else $y=x^2/20$

Ockhams rageknив...

| | *x* | *y* |
|---|---|---|
| **Example 1** | Our house is red. | Vores hus er rødt |
| **Example 2** | Our house is old, too. | Vores hus er også gammelt. |
| **Example 3** | My bike is behind the house. | Min cykel står bagom huset. |
| **Example 4** | My bike is red. | Min cykel er rød. |
| **Example 5** | My bike is rather than yours. | Min cykel er hurtigere end din. |

## Om Kunsthal Charlottenborg

Foto: Anders Sune Berg

Kunsthal Charlottenborg er et af de største og smukkeste udstillingssteder for samtidskunst i Europa, hvor der præsenteres et ambitiøst udstillingsprogram suppleret med aktiviteter som artist talks, performances, koncerter og videovisninger. De mange forskellige aktiviteter retter sig mod et bredt publikum både i og omkring København, og gør Charlottenborg til dét centrale samlingspunkt for samtidskunst i byen.

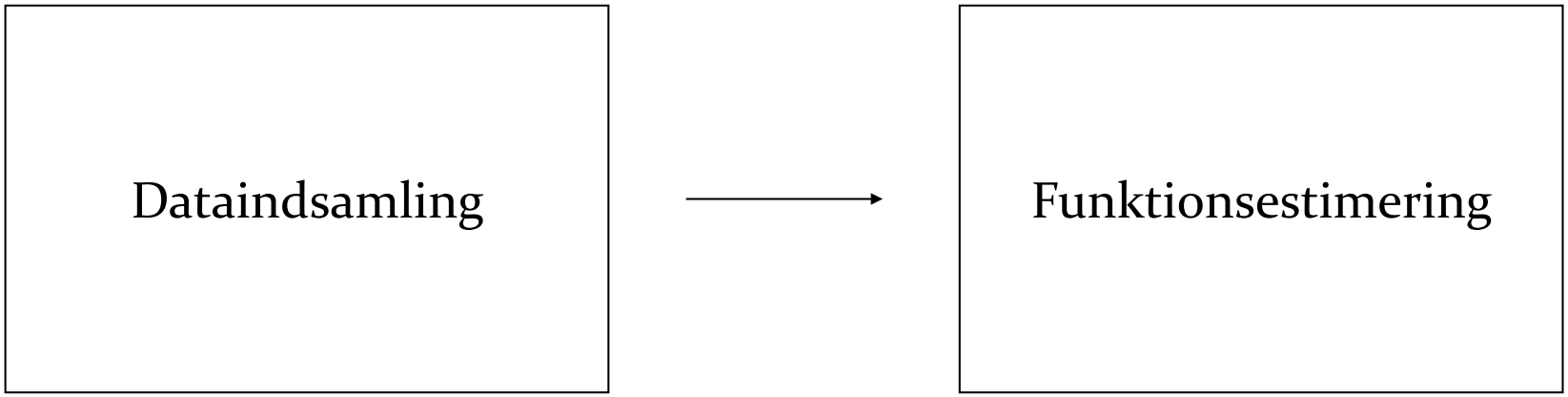
Kunsthal Charlottenborg og Det Kongelige Danske Kunstakademis Billedkunstskoler er med virkning fra 1. september 2012 fusioneret til en samlet organisation på initiativ af Kulturministeriet.
Kunsthal Charlottenborg er derfor nu en del af Det Kongelige Danske Kunstakademis Billedkunstskoler.

## About Kunsthal Charlottenborg

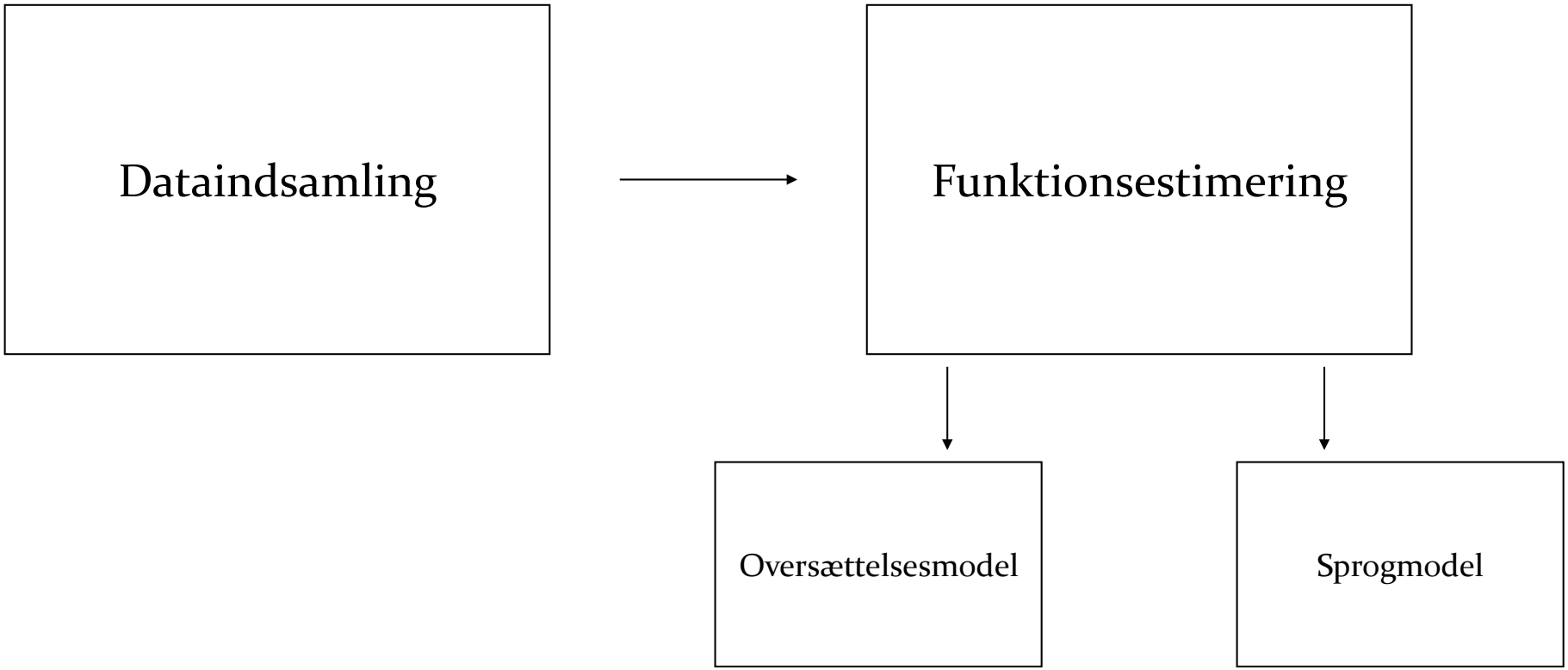
Photo: Anders Sune Berg

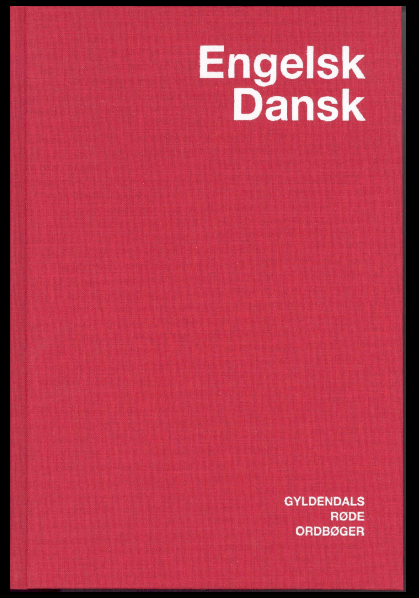
Kunsthal Charlottenborg is one of the largest and most beautiful spaces for contemporary art in Europe, presenting an ambitious program of exhibitions and events including talks, performances, concerts and screenings. This spread of activities is designed to speak to a wide range of audiences in Copenhagen and beyond, making Charlottenborg the main cross-roads for contemporary art in the city.

Kunsthal Charlottenborg and The Royal Danish Academy of Fine Arts' Schools of Visual Arts have with effect from 1 September 2012 merged into a single organization on the initiative of The Ministry of Culture. Kunsthal Charlottenborg has thereby become a part of The Royal Danish Academy of Fine Arts' Schools of Visual Arts.


Dataindsamling
Funktionsestimering

# Frasebaseret maskinoversættelse

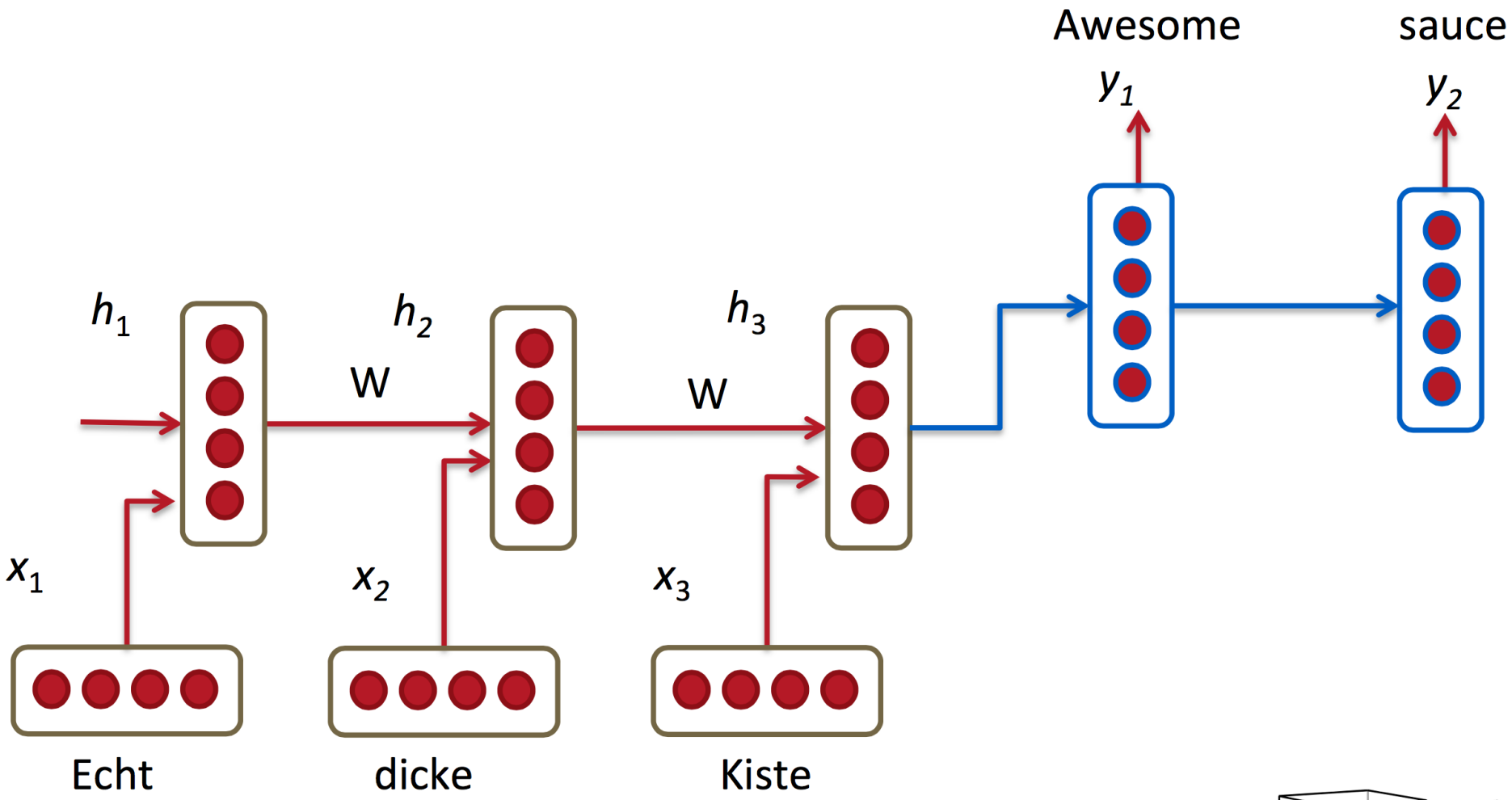

Dataindsamling
Funktionsestimering
Oversættelsesmodel
Sprogmodel

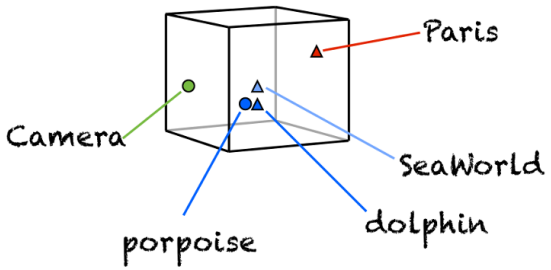



- **Oversættelsesmodel**
- **IBM-1:** P(the house|huset) x P(is|er) x P(red|rødt)
- **IBM-2:** P(of course|naturligvis) x P(the house|huset) x P(is|er) x P(red|rødt) x P(1|1,6,4) x P(2|1,6,4) x P(3|3,6,4) x P(4|3,6,4) x **P(5|2,6,4)** x P(6,4,6,4)


- **Sprogmodel**
- P(Naturligvis|NULL) x P(er|Naturligvis) x P(huset|er) x P(rødt|huset)

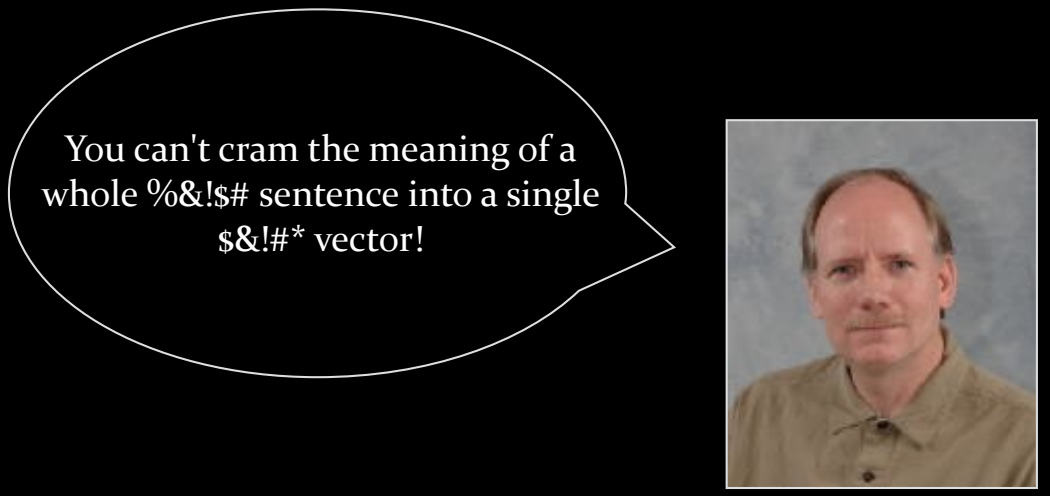
# Maskinoversættelse & deep learning


Awesome
sauce
$y_1$
$y_2$
$h_1$
$h_2$
$h_3$
W
W
$x_1$
$x_2$
$x_3$
Echt
dicke
Kiste

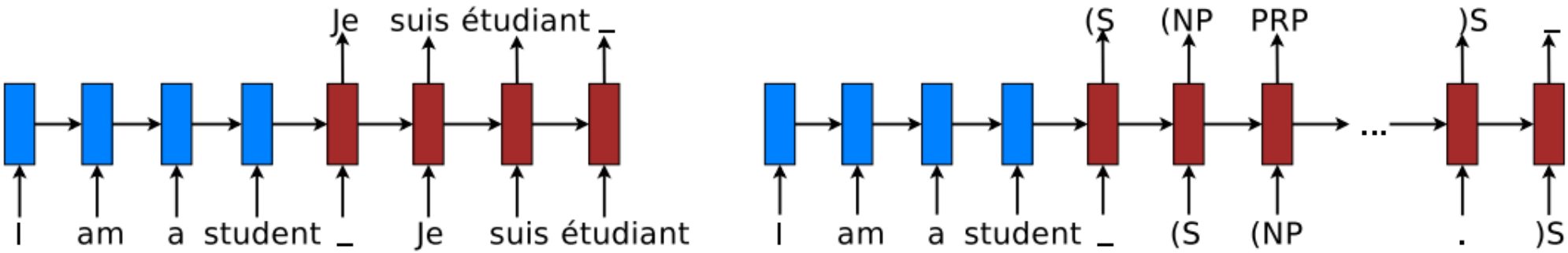

Paris
Camera
SeaWorld
porpoise
dolphin

# Fordelene

- Ingen problemer med nye ord
- Samme arkitektur kan genbruges på tværs af sprog
- *Multi-task learning…*

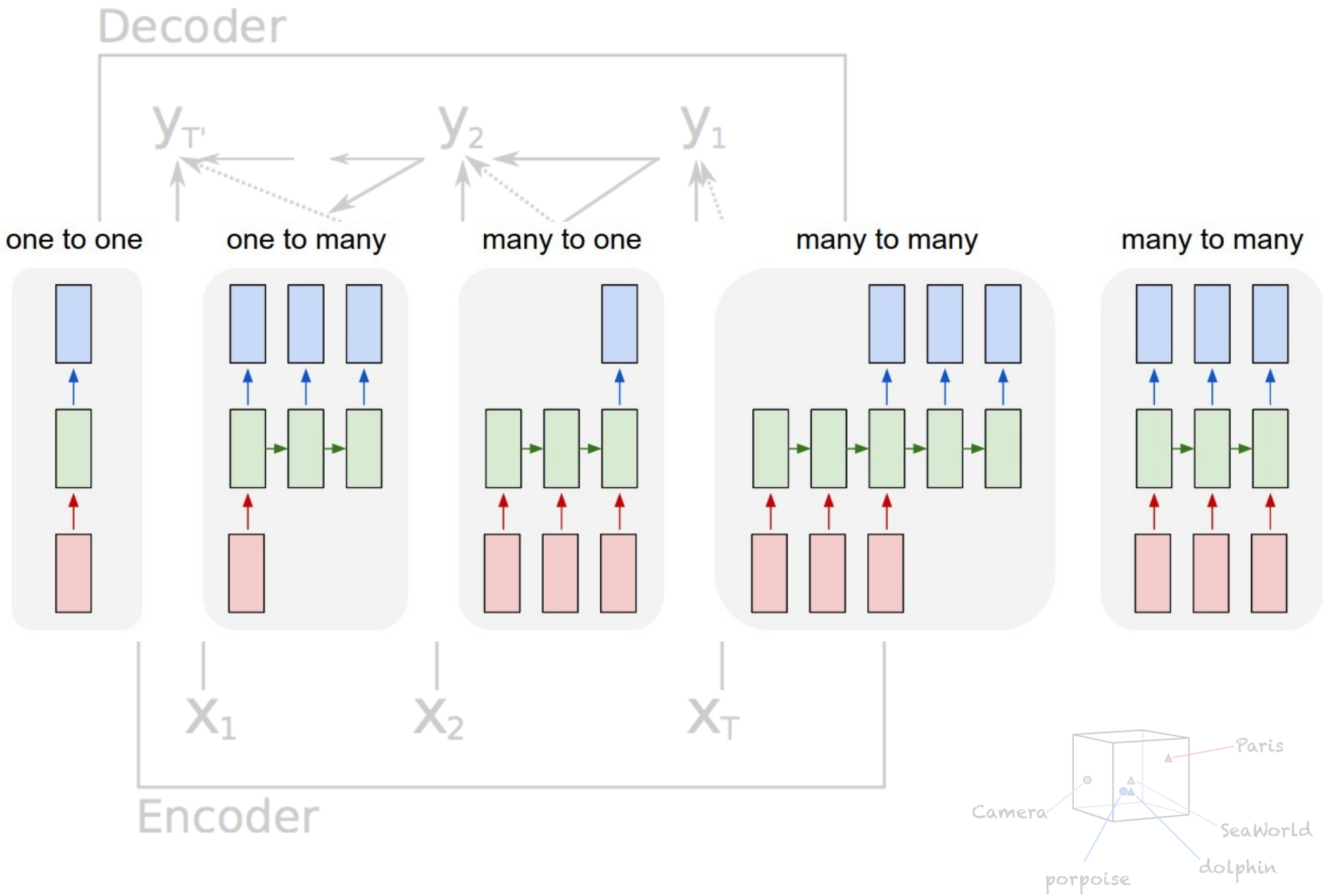


# Luong et al. (2016)




- Multi-task RNN encoder-decoders
- Samtidig læring af oversættelse, syntaktisk analyse og billedtekst-generering

http://arxiv.org/pdf/1511.06114.pdf


Decoder
$y_{T'}$
$y_2$
$y_1$
one to one
one to many
many to one
many to many
many to many
$x_1$
$x_2$
$x_T$
Encoder
Paris
Camera
SeaWorld
dolphin
porpoise

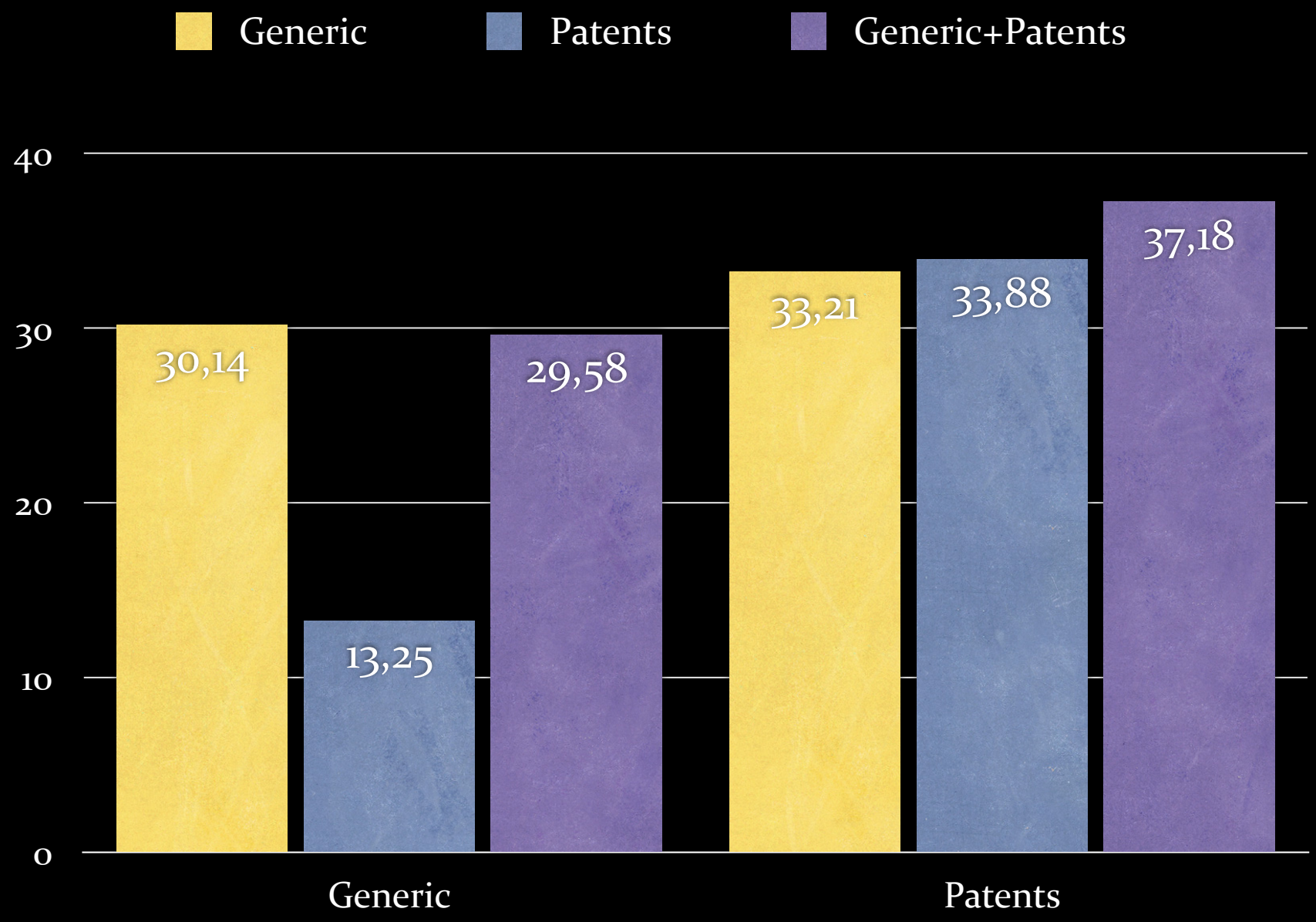
Send mere data!


# Send mere data!

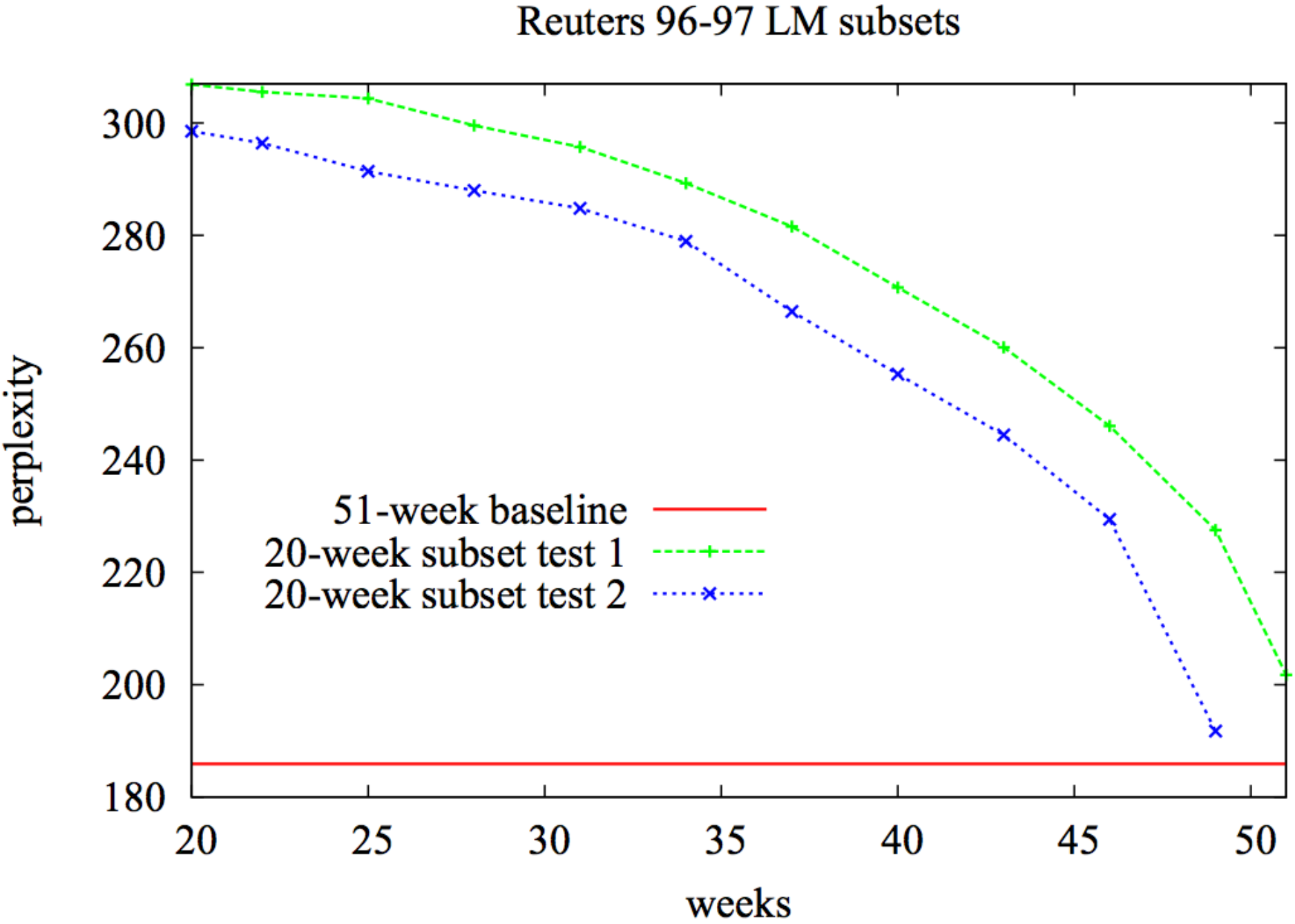

# Send mere data!



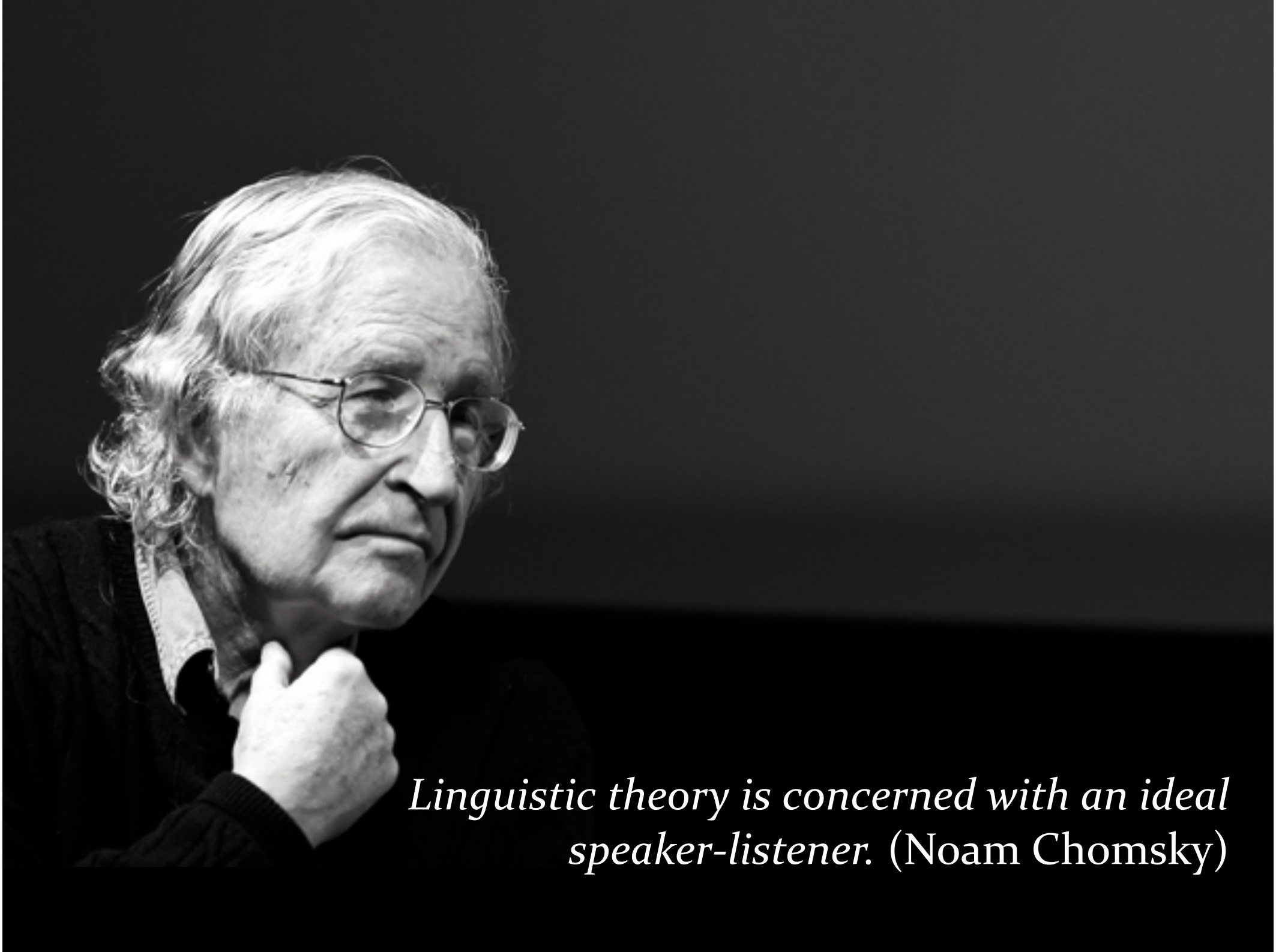
# Send mere data!




Wang et al. (2012)




Levenberg (2011)


*Linguistic theory is concerned with an ideal speaker-listener.* (Noam Chomsky)

*Everyone speaks transitional dialects.* (Max Weinreich)

# Andre udfordringer for MT:

- Det er svært at evaluere kvaliteten af oversættelser.
- Vi har MT for omkring 80 sprog, men der er tusinder af sprog derude.
- Tekstkohærens.

?