# Introduction to machine translation

Marco Kuhlmann

Department of Computer and Information Science

LINKÖPING
UNIVERSITY

# Machine translation

# A timeline of machine translation

rule-based machine translation

statistical machine translation

1950　1980　1990　2015　2020

example-based machine translation

neural machine translation
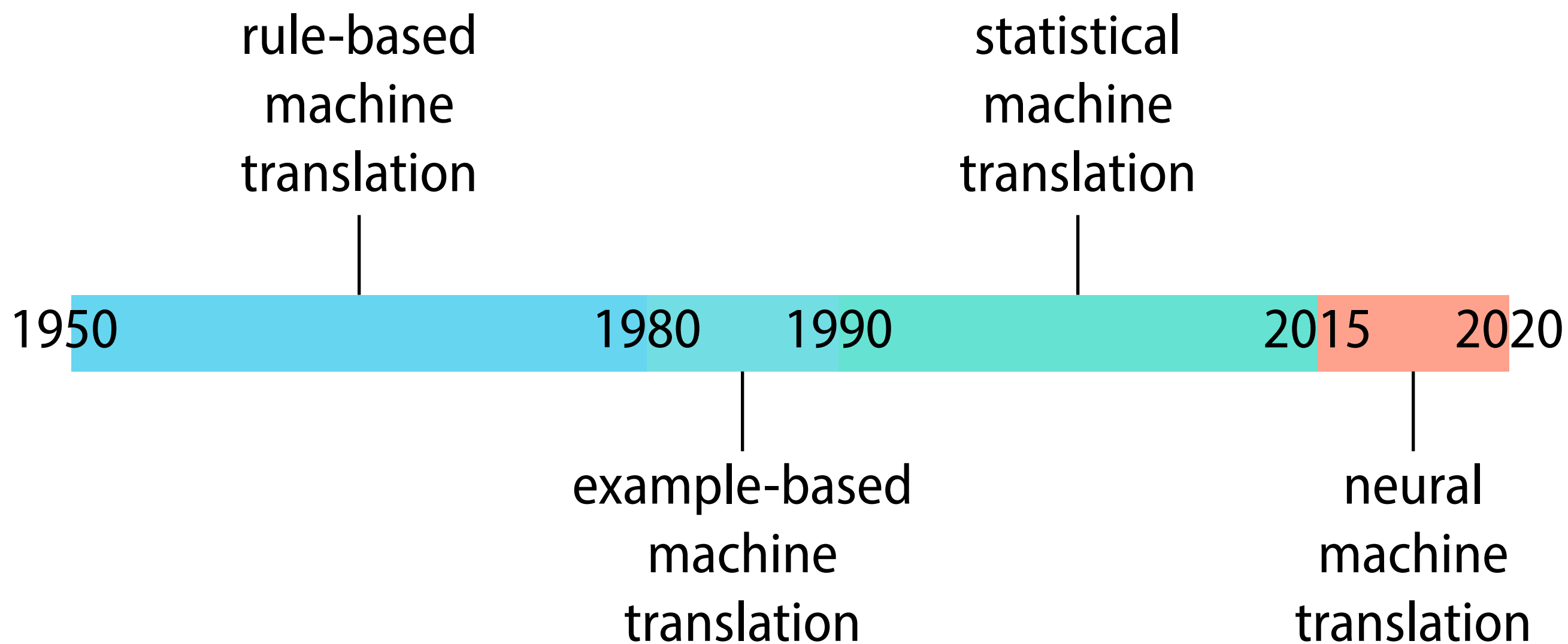
# Interlingual machine translation

abstract representation

**Intentionally_create**
Creator: dbr:Daphne_Koller
Created_entity: dbr:Coursera

linguistic structure

subject    object

Koller    co-founded    Coursera
PROPN      VERB          PROPN

object
subject

bhunaigh    Koller    Coursera
VERB        PROPN     PROPN

Koller co-founded Coursera

surface sentence

Bhunaigh Koller Coursera

# Noisy Channel Model

sender ——————— channel ————▶ receiver

(English)         (distorts the message)      (Russian)
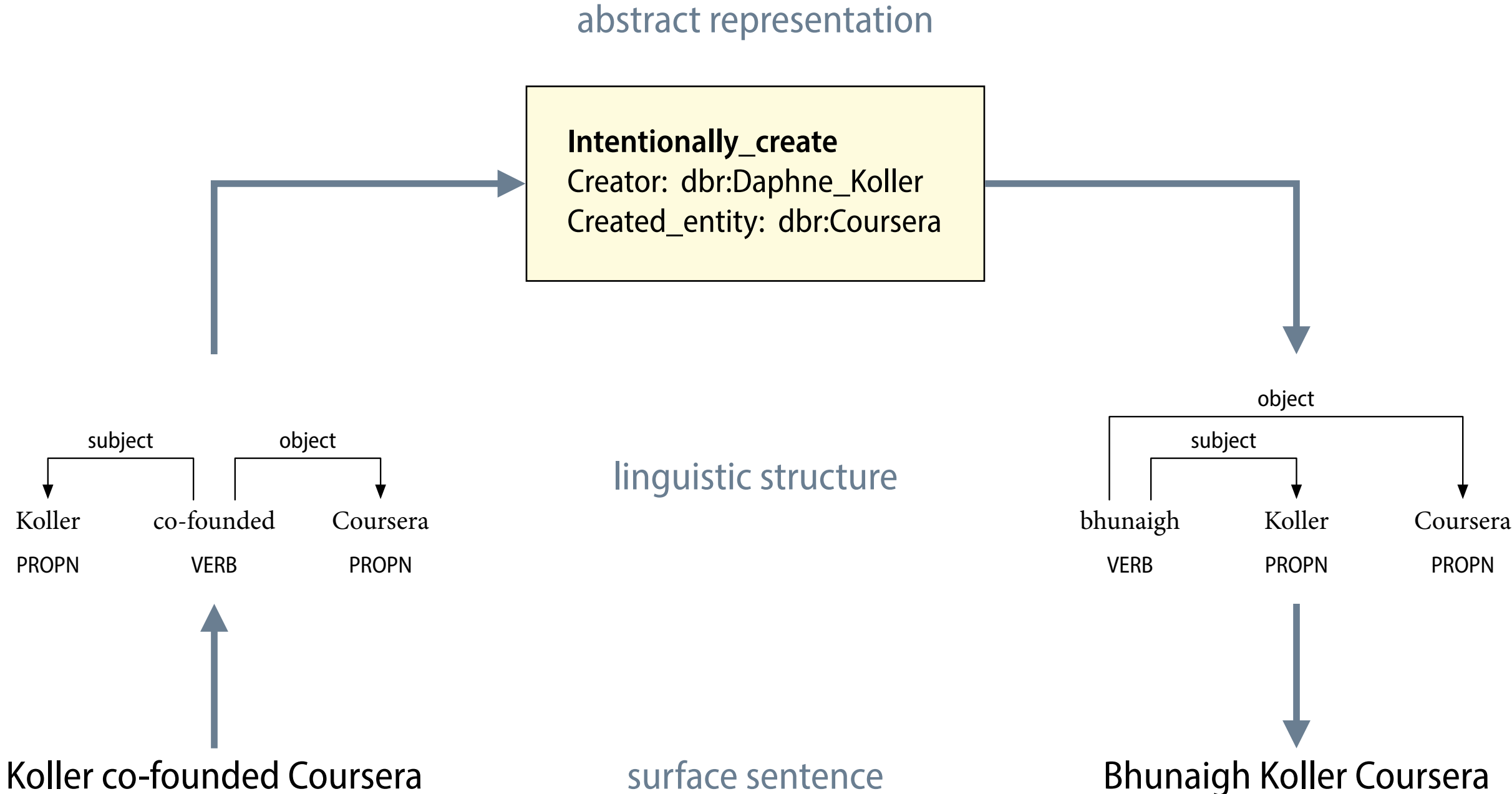
When I look at an article in Russian, I say:
'This is really written in English,
but it has been coded in some strange symbols.
I will now proceed to decode.'
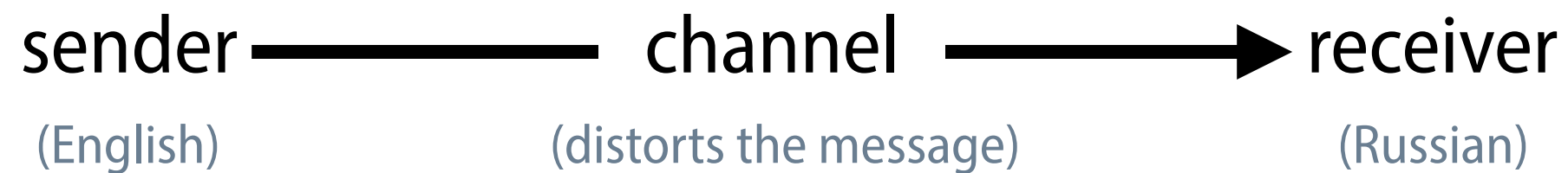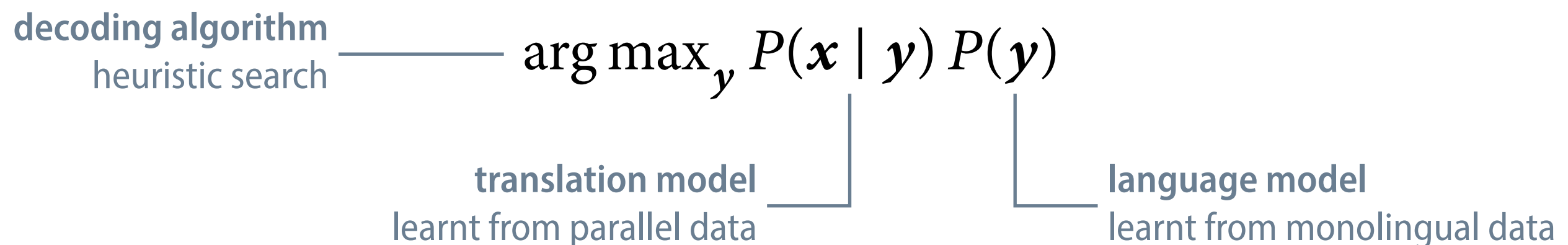
Warren Weaver (1894–1978)

# Statistical machine translation (SMT)

- Formulate machine translation as an optimisation task: Given a source sentence $\boldsymbol{x}$, find the most probable target sentence $\boldsymbol{y}$:

$$\arg\max_{\boldsymbol{y}} P(\boldsymbol{y} \mid \boldsymbol{x})$$

- Use Bayes' rule to decompose the probability model into two components that can be learned separately:

**decoding algorithm**
heuristic search ——— $\arg\max_{\boldsymbol{y}} P(\boldsymbol{x} \mid \boldsymbol{y}) \, P(\boldsymbol{y})$

**translation model**
learnt from parallel data

**language model**
learnt from monolingual data

# Parallel corpora

- Canadian Hansard (English–French); extracted from the proceedings of the Canadian Parliament.

- Europarl (21 languages); 30.32 M parallel sentences extracted from the proceedings of the European parliament.

  Link to the Europarl website

- OPUS (several languages); growing collection of translated texts, automatically preprocessed and aligned.

  Link to the OPUS website

# Word-to-word alignments

# IBM Model 1
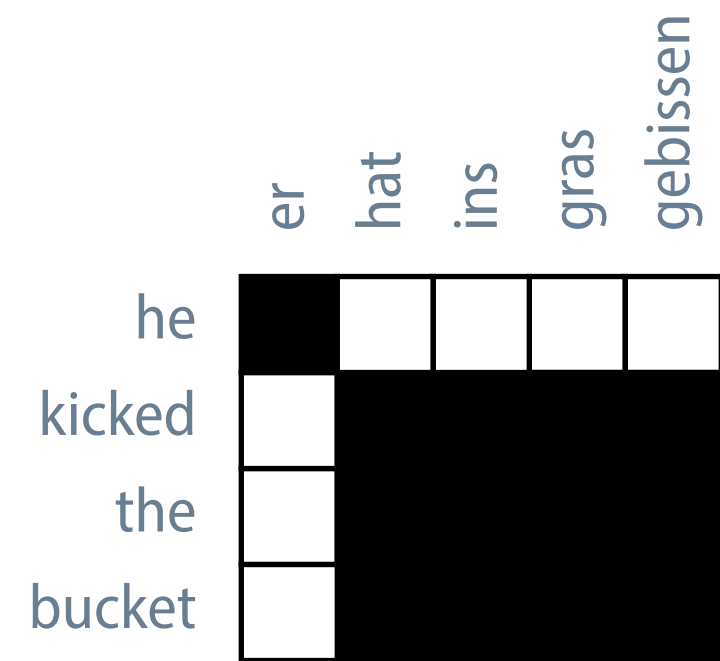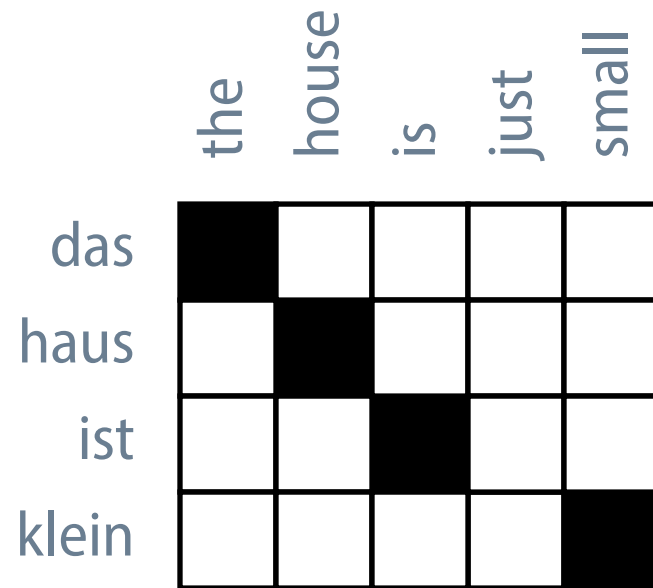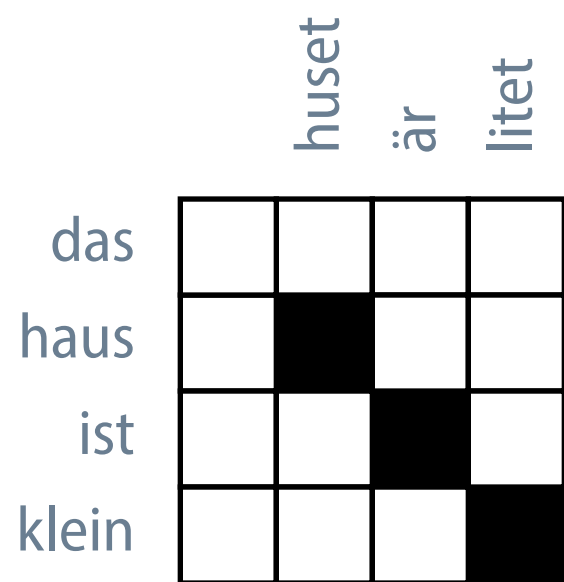
normalisation constant

$$P(\boldsymbol{x}, a \mid \boldsymbol{y}) = \frac{\varepsilon}{(|\boldsymbol{y}| + 1)^{|\boldsymbol{x}|}} \prod_{j=1}^{|\boldsymbol{x}|} t(x_j \mid y_{a(j)})$$

alignment function

- First in a series of increasingly complex statistical translation models; deals only with lexical (word-to-word) translation.

- Central component: The lexical translation probability $t$ of observing a source word $x$, given the aligned target word $y$.

Brown et al. (1993)

# Training statistical machine translation models

- We would like to estimate the lexical translation probabilities from a parallel corpus – but we do not have the alignments.

- We can bootstrap the translation probabilities and alignments in parallel using the Expectation Maximization (EM) algorithm:

  1. initialise the model parameters randomly
  2. calculate alignments based on the current model parameters
  3. estimate new model parameters from the new alignments
  4. repeat steps 2–3 until convergence

Brown et al. (1993)

# Statistical machine translation (SMT)

- Research on statistical machine translation led to significant improvements in the availability and quality of translation.

  The first version of Google Translate (2006–2016) was an SMT system.

- However, the best systems were extremely complex and required large amounts of external resources and feature engineering.

  Open-source example: Moses

# Evaluation: BLEU (Bilingual Evaluation Understudy)

- BLEU compares the automatic translation of a source sentence to one or several human-created translations.

- BLEU combines $n$-gram precision (for $n$ up to 4) with a brevity penalty for too-short translations.

- BLEU has been criticised for not correlating well with human judgement, and several other evaluation measures exist.

Papineni et al. (2002); Callison-Burch et al. (2006)