# First Workshop *Data Science: Theory and Application*
# RWTH Aachen University, Oct. 26, 2015

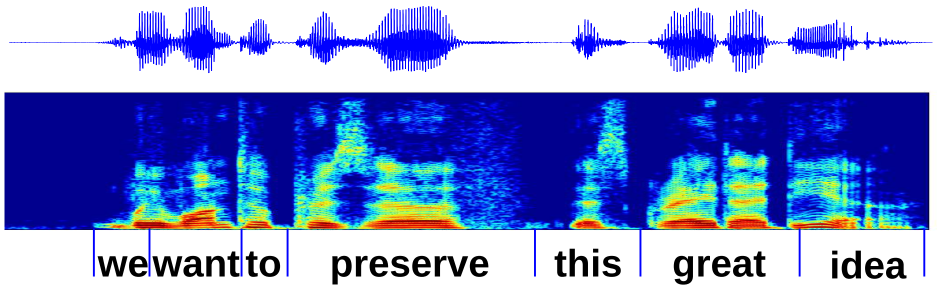
# The Statistical Approach to Speech Recognition and Natural Language Processing

**Hermann Ney**
**Human Language Technology and Pattern Recognition**

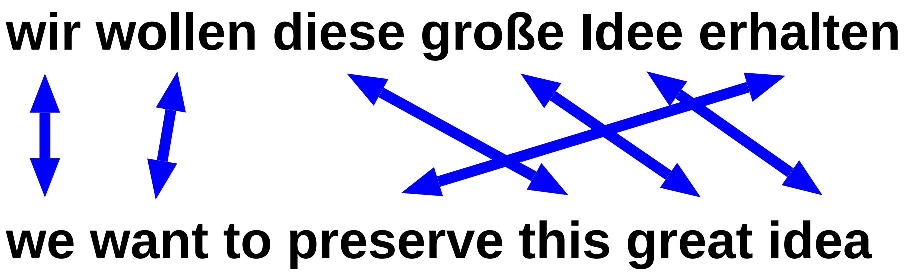
**RWTH Aachen University, Aachen**
**DIGITEO Chair, LIMSI-CNRS, Paris**

# Human Language Technology (HLT)

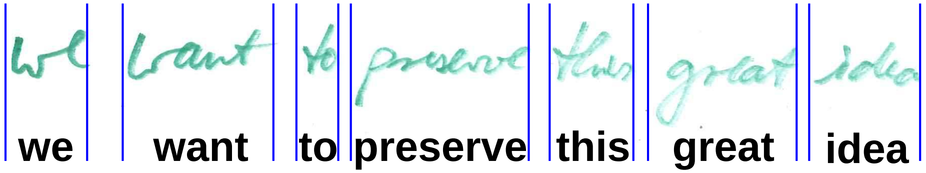
## Speech Recognition



## Machine Translation


wir wollen diese große Idee erhalten
we want to preserve this great idea

## Text Image Recognition



tasks:
- speech recognition
- text image recognition
- machine translation
  (+ sign language,...)

# Human Language Technology

characteristic properties:

- well-defined ’classification’ tasks:
  - due to 5000-year history of (written!) language
  - well-defined classes: letters or words of the language
- easy task for humans (but: native vs. foreign language ?)
- hard task for computers
  (as the last 50 years have shown!)

unifying view:

- formal task: input string → output string
- output string: string of words/letters in a natural language
- models of context and dependencies: strings in input and output
  - within input and output string
  - across input and output string

# Projects

**activities of RWTH team in large-scale joint projects:**

- **TC-STAR 2004-2007: funded by EU**
  **– first research system for speech-to-speech translation on real-life data (EU parliament)**
  **– partners: KIT Karlsruhe, FBK Trento, LIMSI Paris, UPC Barcelona, IBM-US Research, ...**

- **GALE 2005-2011: funded by US DARPA**
  **emphasis on Chinese and Arabic speech and text**
- **BOLT 2011-2015: funded by US DARPA**
  **emphasis on colloquial text for Arabic and Chinese**
- **QUAERO 2008-2013: funded by OSEO France**
  **European languages, more colloquial speech, handwriting**
- **BABEL 2012-2017: funded by US IARPA**
  **spoken term detection with noisy and limited training data**
- **EU projects 2012-2014: EU-Bridge, TransLectures**
  **emphasis on recognition and translation of lectures (academic, TED, ...)**

# Statistical Approach: 1975-2015

- two strings: input $x_1^M := x_1 ... x_m ... x_M$ and output $c_1^N := c_1 ... c_n ... c_N$
  with a probabilistic dependence: $p(N, c_1^N | x_1^M)$
- performance measure or loss function: $L[\tilde{c}_1^{\tilde{N}}, c_1^N]$
  between true output $\tilde{c}_1^{\tilde{N}}$ and hypothesized output $c_1^N$
- Bayes decision rule minimizes expected loss:

$$x_1^M \to \hat{c}_1^{\hat{N}}(x_1^M) \; := \; \arg\min_{N, c_1^N} \left\{ \sum_{\tilde{N}, \tilde{c}_1^{\tilde{N}}} p(\tilde{N}, \tilde{c}_1^{\tilde{N}} | x_1^M) \cdot L[\tilde{c}_1^{\tilde{N}}, c_1^N] \right\}$$

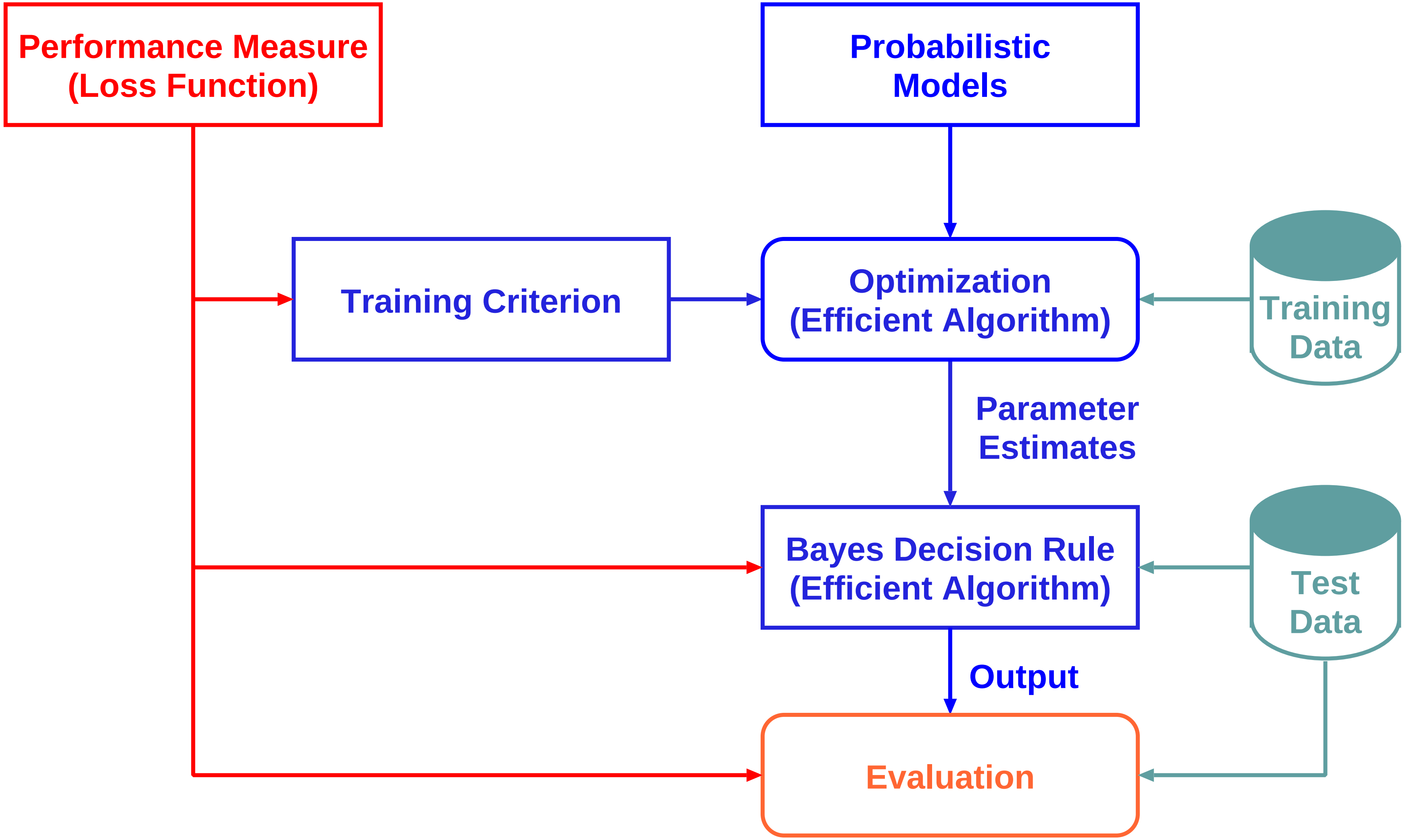
  rule for minimum string error: $x_1^M \to \hat{c}_1^{\hat{N}}(x_1^M) := \arg\max_{N, c_1^N} \left\{ p(N, c_1^N | x_1^M) \right\}$

- from true to model distribution: separation of language model $p(N, c_1^N)$

$$p(N, c_1^N | x_1^M) = p(N, c_1^N) \cdot p(x_1^M | c_1^N) \Big/ p(x_1^M)$$

  – advantage: huge amounts of monolingual training data
  – extension: log-linear modelling

# Statistical Approach to String Classification for HLT Tasks



# Statistical Approach: Interpretation

four ingredients:

- performance measure: often edit distance
  we have to decide how to judge the quality of the system output
- probabilistic models (with a suitable structure):
  to capture the dependencies within and between input and output strings
  – elementary observations: Gaussian mixtures, log-linear models, support vector machines (SVM), artificial neural nets (ANN), ...
  – strings: $n$-gram Markov chains, Hidden Markov models (HMM), recurrent neural nets (RNN), LSTM RNN, ...
- training criterion:
  to learn the free parameters of the models
  – ideally should be linked to performance criterion
  – might result in complex mathematical optimization (efficient algorithms!)
  – extreme situation: number of free parameters vs. observations
- Bayes decision rule:
  to generate the output word sequence
  – combinatorial problem (efficient algorithms)
  – should exploit structure of models
  examples: dynamic programming and beam search, A$^*$ and heuristic search, ...

# From Speech Recognition to Machine Translation (MT)

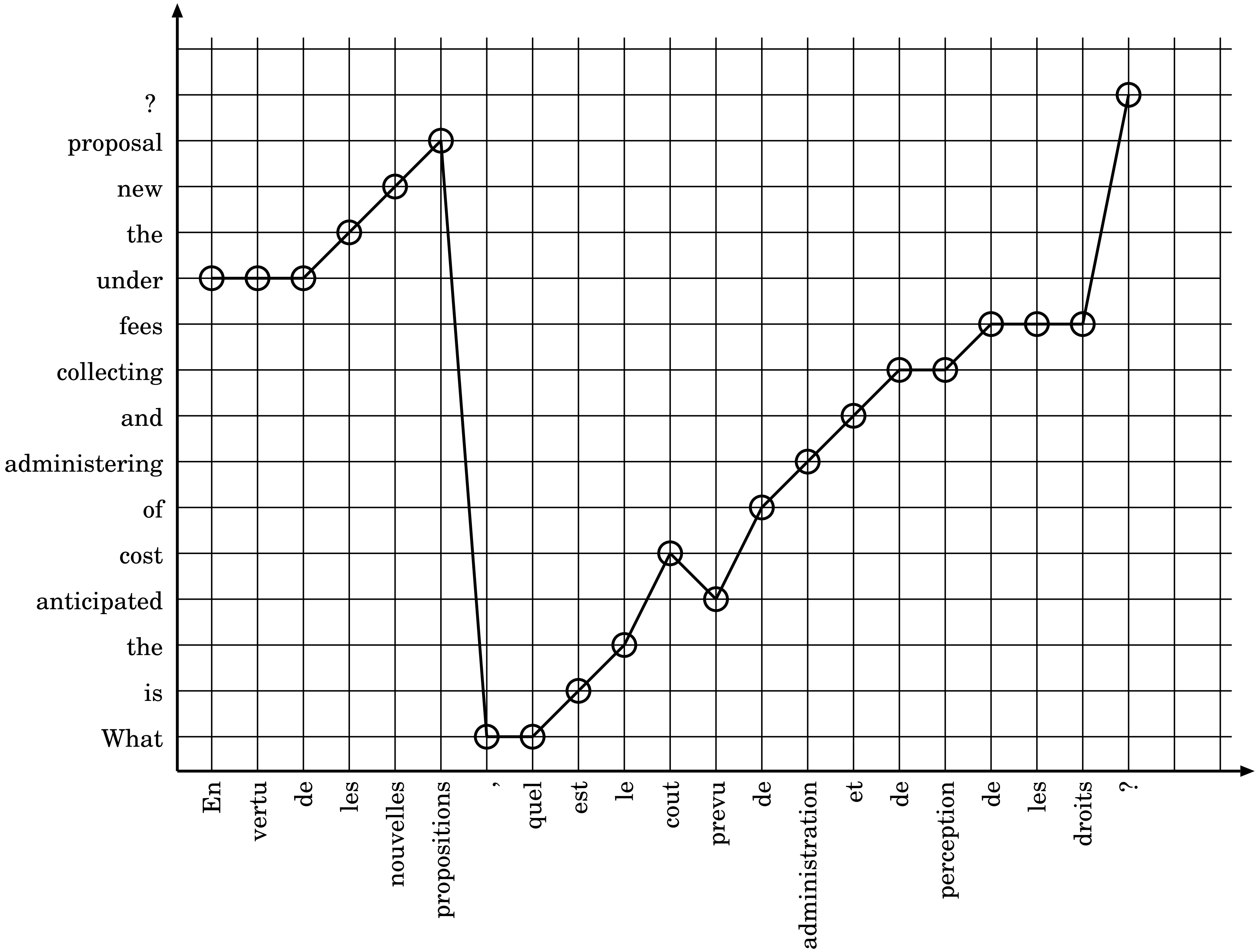
use of statistics has been controversial in symbolic processing and computational linguistics:

- Chomsky 1969:
  ... the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term.
- was considered to be true by most experts in (rule-based) natural language processing and artificial intelligence

history of statistical approach to MT:

- 1989-94: IBM's pioneering work
- since 1996: only a few teams advocated statistical MT:
  RWTH Aachen, UP Valencia, HKUST Hong Kong, CMU Pittsburgh
- since 2004: from singularity to mainstream in MT
- 2008 Google Translate

# Example of Alignment (Canadian Hansards)



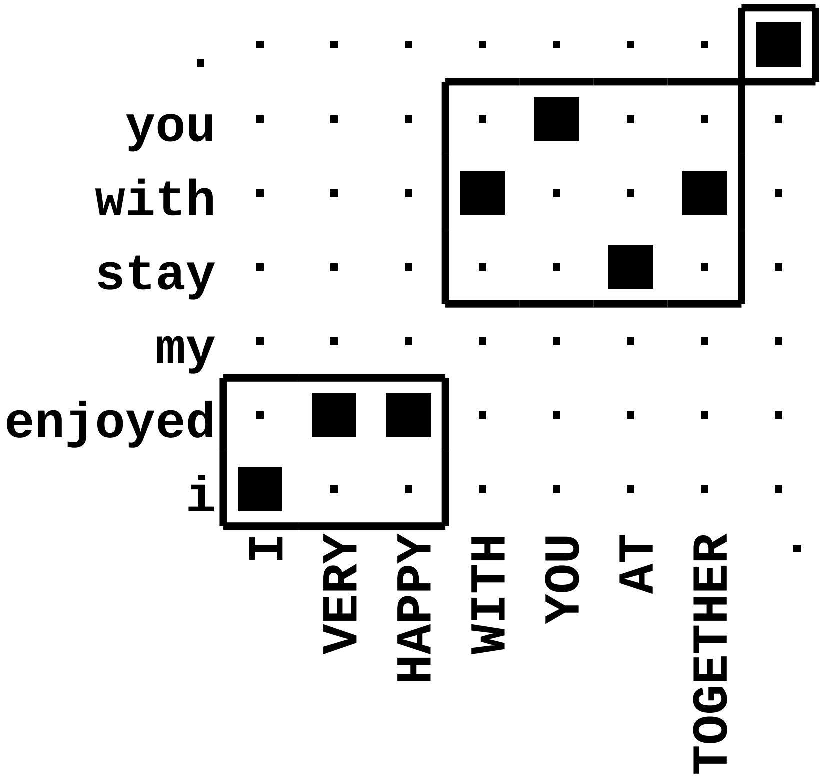
# From Words to Phrases

**source sentence** 我 很 高兴 和 你 在 一起 .

**gloss notation** **I VERY HAPPY WITH YOU AT TOGETHER .**

**target sentence** **I enjoyed my stay with you .**

**Viterbi alignment for $F \rightarrow E$:**

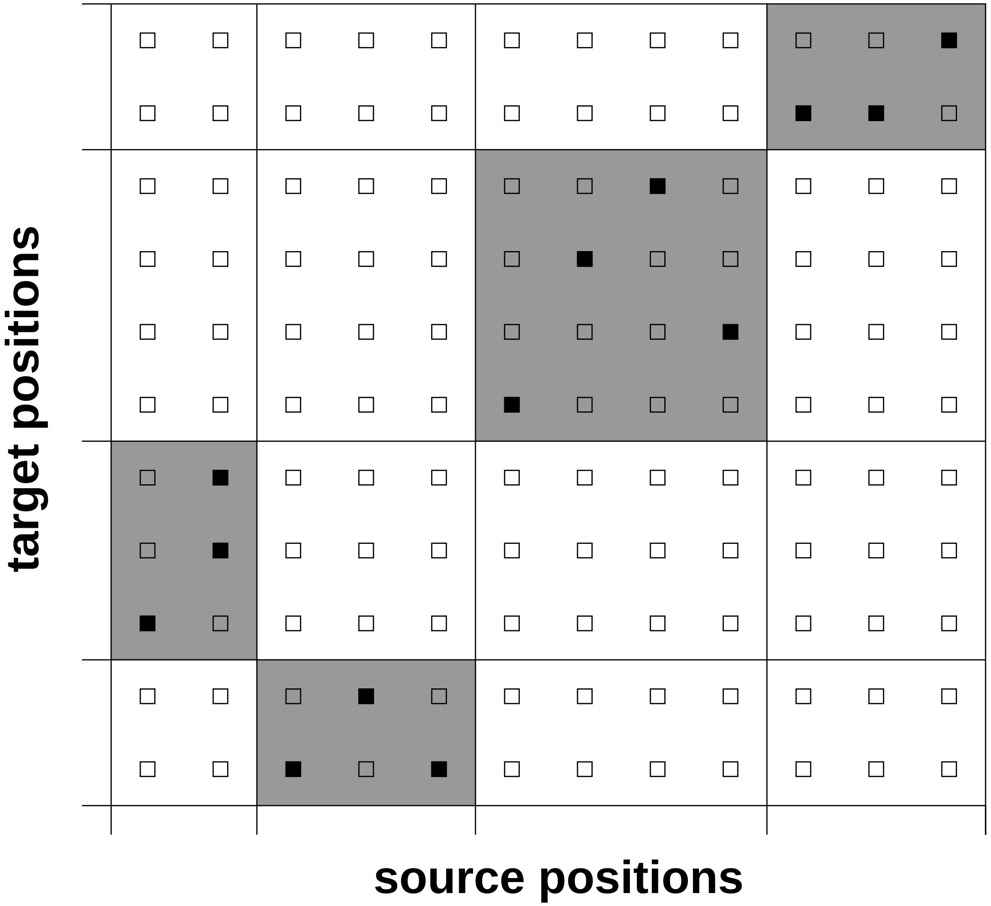


|  | I | VERY | HAPPY | WITH | YOU | AT | TOGETHER | . |
|---|---|---|---|---|---|---|---|---|
| . | · | · | · | · | · | · | · | ■ |
| you | · | · | · | · | ■ | · | · | · |
| with | · | · | · | ■ | · | · | ■ | · |
| stay | · | · | · | · | · | ■ | · | · |
| my | · | · | · | · | · | · | · | · |
| enjoyed | · | ■ | ■ | · | · | · | · | · |
| i | ■ | · | · | · | · | · | · | · |

# From Words to Phrases (Segments)

**phrase-based approach:**

- **training: extraction of phrase pairs (= two-dim. 'blocks') after alignment/lexicon training**
- **translation process: phrases are the smallest units**



# Conclusions
## HLT tasks: mapping from input string to output string

- statistical approach (inc. ANNs): four key ingredients
  - choice of performance measure: errors at string, word, phoneme, frame level
  - probabilistic models at these levels and the interaction between these levels
  - training criterion along with an optimization algorithm
  - Bayes decision rule along with an efficient implementation
- about recent work on artificial neural nets:
  - they result in significant improvements
  - they provide one more type of probabilistic models
  - they are PART of the statistical approach
- specific future challenges for statistical approach (incl. ANNs) in general:
  - complex mathematical model that is difficult to analyze
  - questions: can we find suitable mathematical approximations with more explicit descriptions of the dependencies and level interactions and of the performance criterion (error rate)?
- specific challenges for ANNs:
  - can the HMM-based alignment mechanism be replaced?
  - can we find ANNs with more explicit probabilistic structures?

# THE END