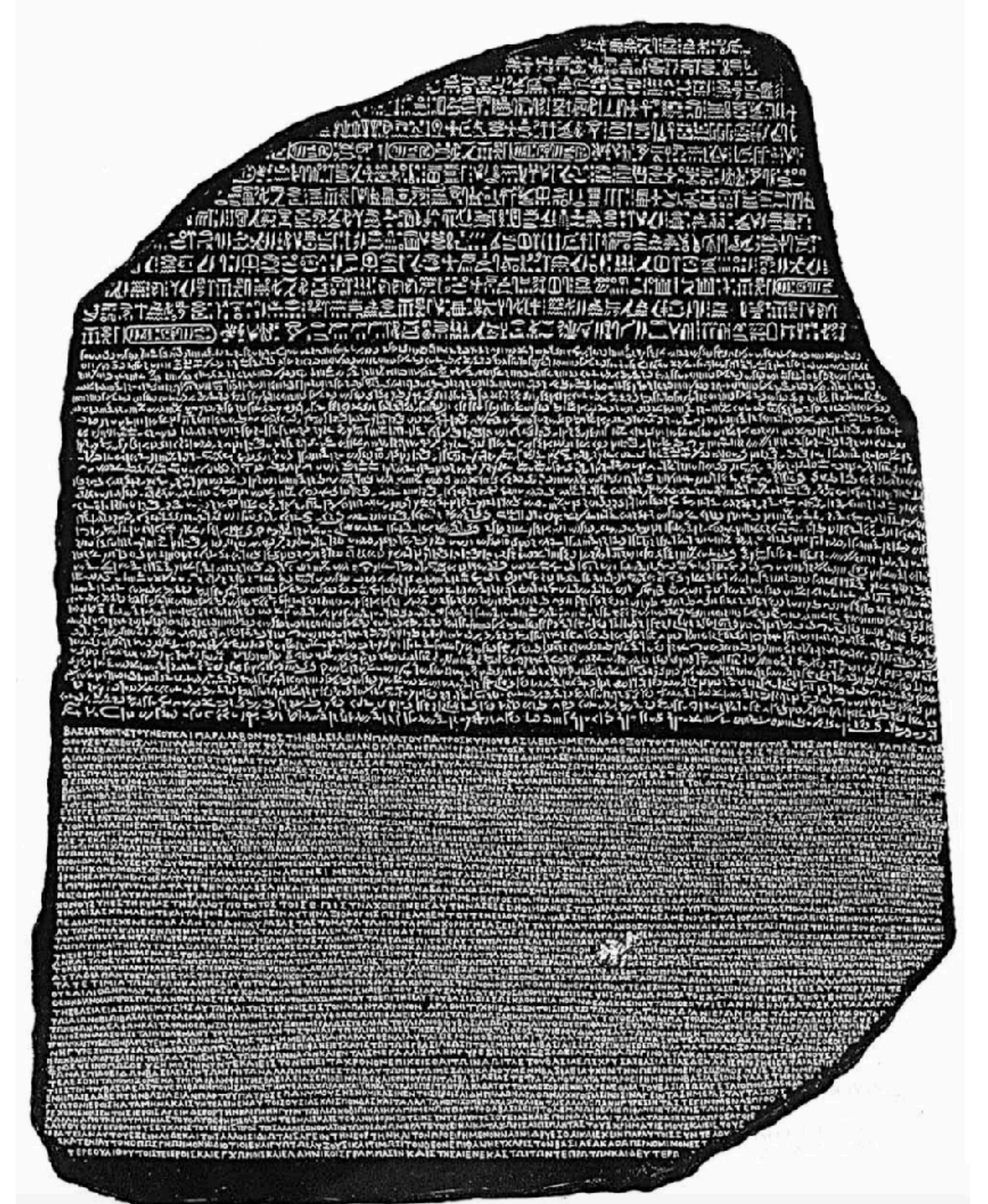
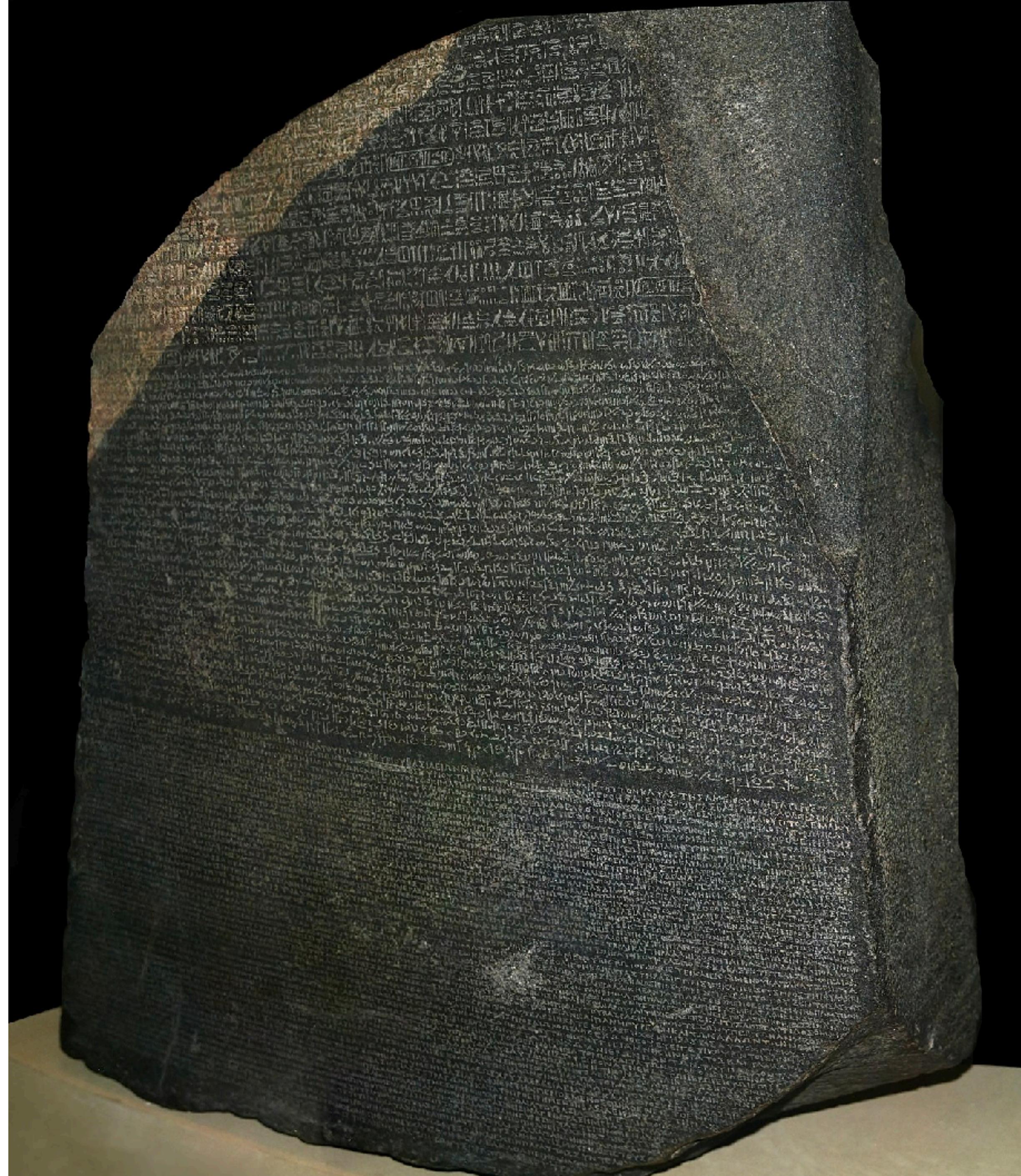


Lecture 19: Machine Translation

Zhizheng Wu

Agenda

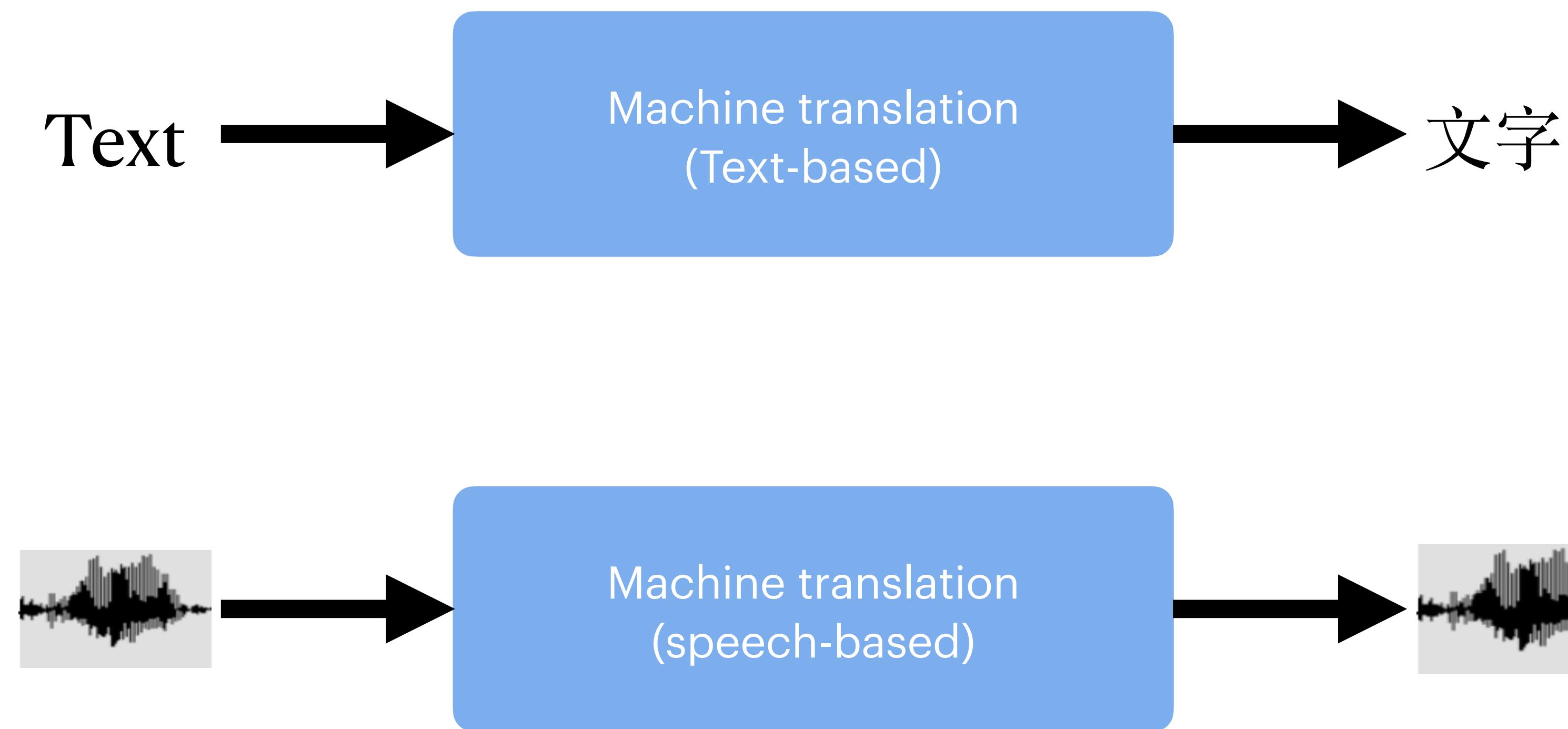
- ▶ Introduction
- ▶ Why machine translation is difficult
- ▶ Statistical MT
- ▶ Neural MT



The Rosetta Stone

- ▶ The same text in three languages
- ▶ This is an instance of parallel text
 - The Greek inscription allowed scholars to decipher the hieroglyphs

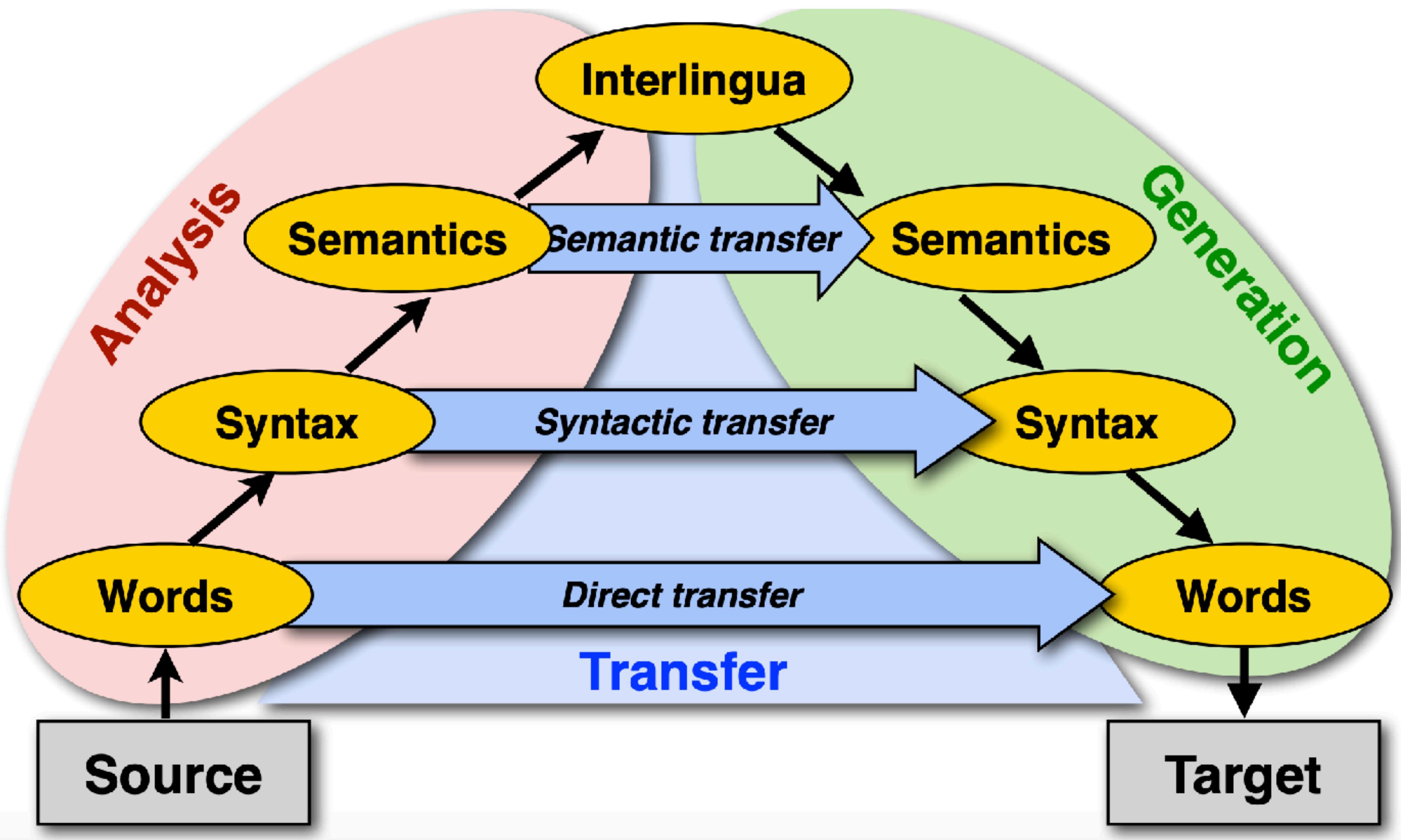
Machine Translation



Applications



The Vauquois triangle



Why machine translation is difficult?

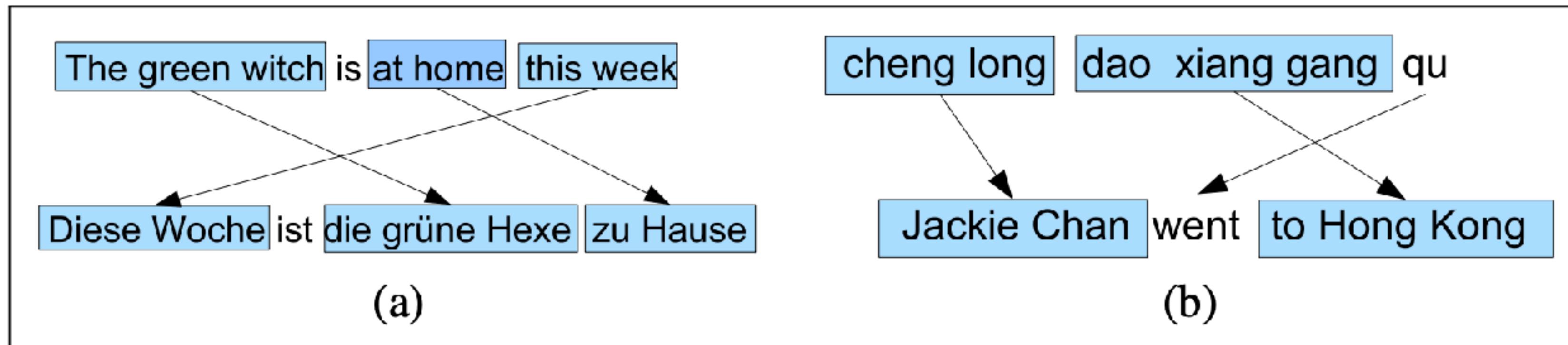
- ▶ Lexical divergences
- ▶ Syntactic divergences
- ▶ Semantic divergences
- ▶ Word-to-Word correspondence divergences

Lexical divergences

- ▶ English: brother vs Chinese: gege, didi (哥哥、弟弟)
- ▶ English: sister vs Chinese: jie jie, meimei (姐姐、妹妹)
- ▶ English: Play vs Chinese: 打、弹、...
- ▶ The different senses of homonymous words generally have different translations:
 - English-German:
 - (river) bank - Ufer
 - (financial) bank - Bank

Syntactic divergences

- Word order



- Head-marking vs. dependent-marking
 - Dependent-marking (English): **the man's house**
 - Head-marking (Hungarian): **the man house-his**

Semantic divergences

- ▶ English has a progressive aspect
 - ‘Peter swims’ vs. ‘Peter is swimming’
- ▶ German can only express this with an adverb
 - ‘Peter schwimmt’ vs. ‘Peter schwimmt gerade’ (‘swims currently’)

Word-to-word correspondences

One to-one:

John loves Mary.

| | |
Jean aime Marie.

**One-to-many:
(and reordering)**

John told Mary a story.
| | |
Jean [a raconté] une histoire [à Marie].

**Many-to-one:
(and elision)**

John is a [computer scientist].
| |
Jean est *informaticien*.

Many-to-many:

John [swam across] the lake.
|
Jean [a traversé] le lac [à la nage].

Statistical MT

Given input in the source language, S ,...

e.g. a Chinese sentence...

主席：各位議員，早晨。

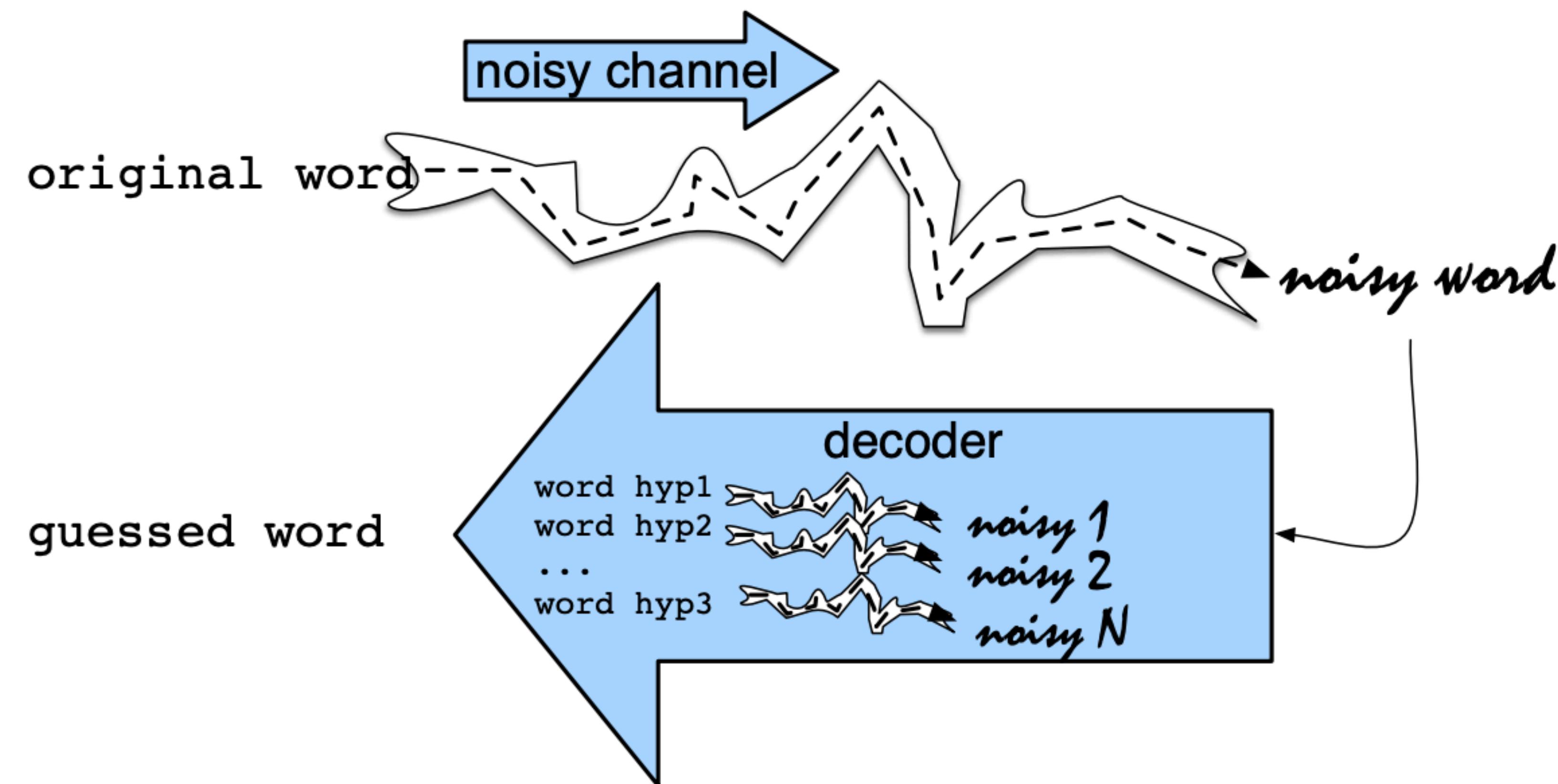
... return the best translation in the target language, T^*

e.g in English:

President: Good morning, Honourable Members.

We can formalize this as $T^* = \operatorname{argmax}_T P(T | S)$

Noise channel model



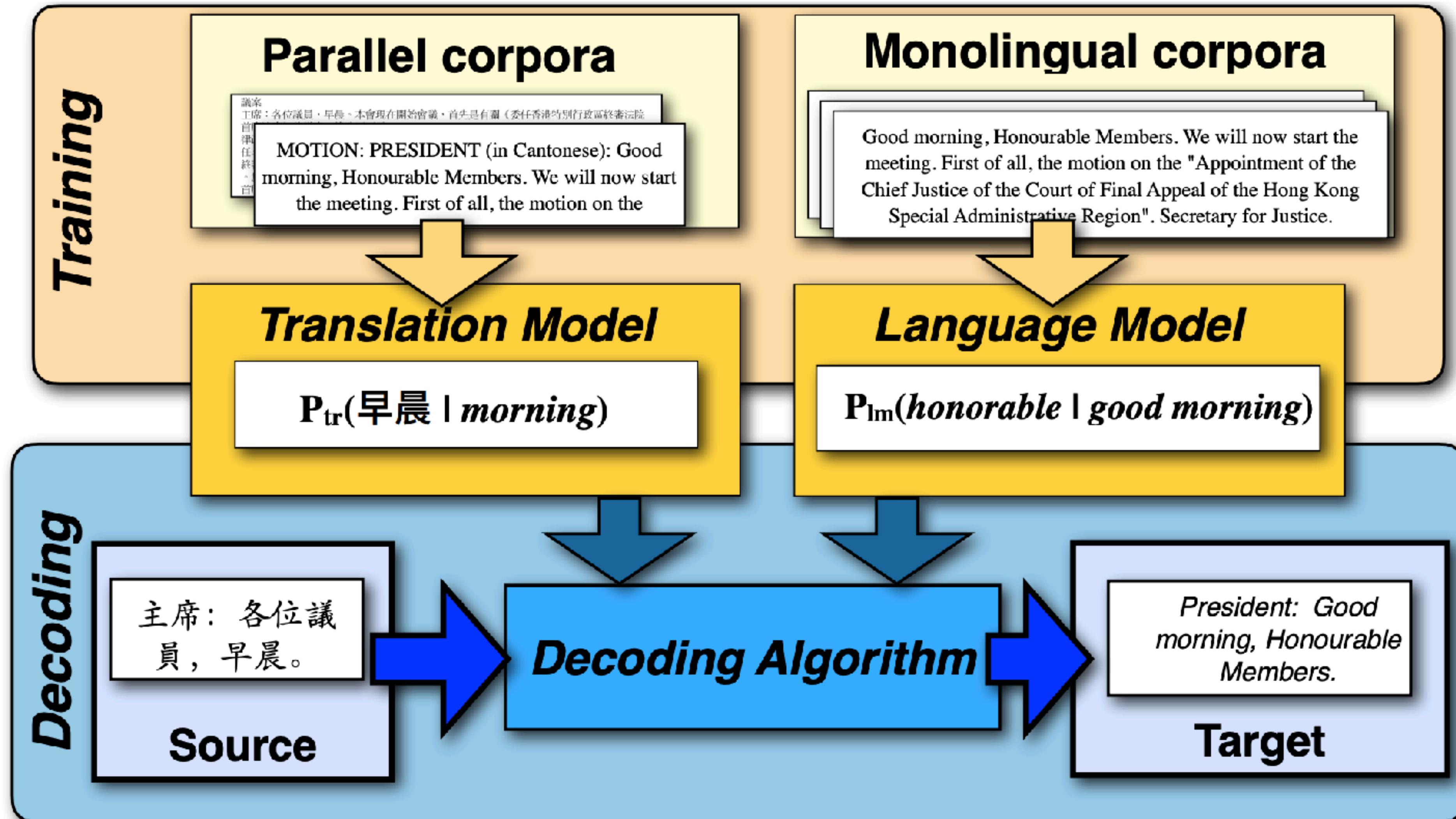
Noise Channel model

- ▶ Suppose we want to translate a foreign language to English, we can model $P(E|F)$
- ▶ by Bayes law, we have the equivalent equation

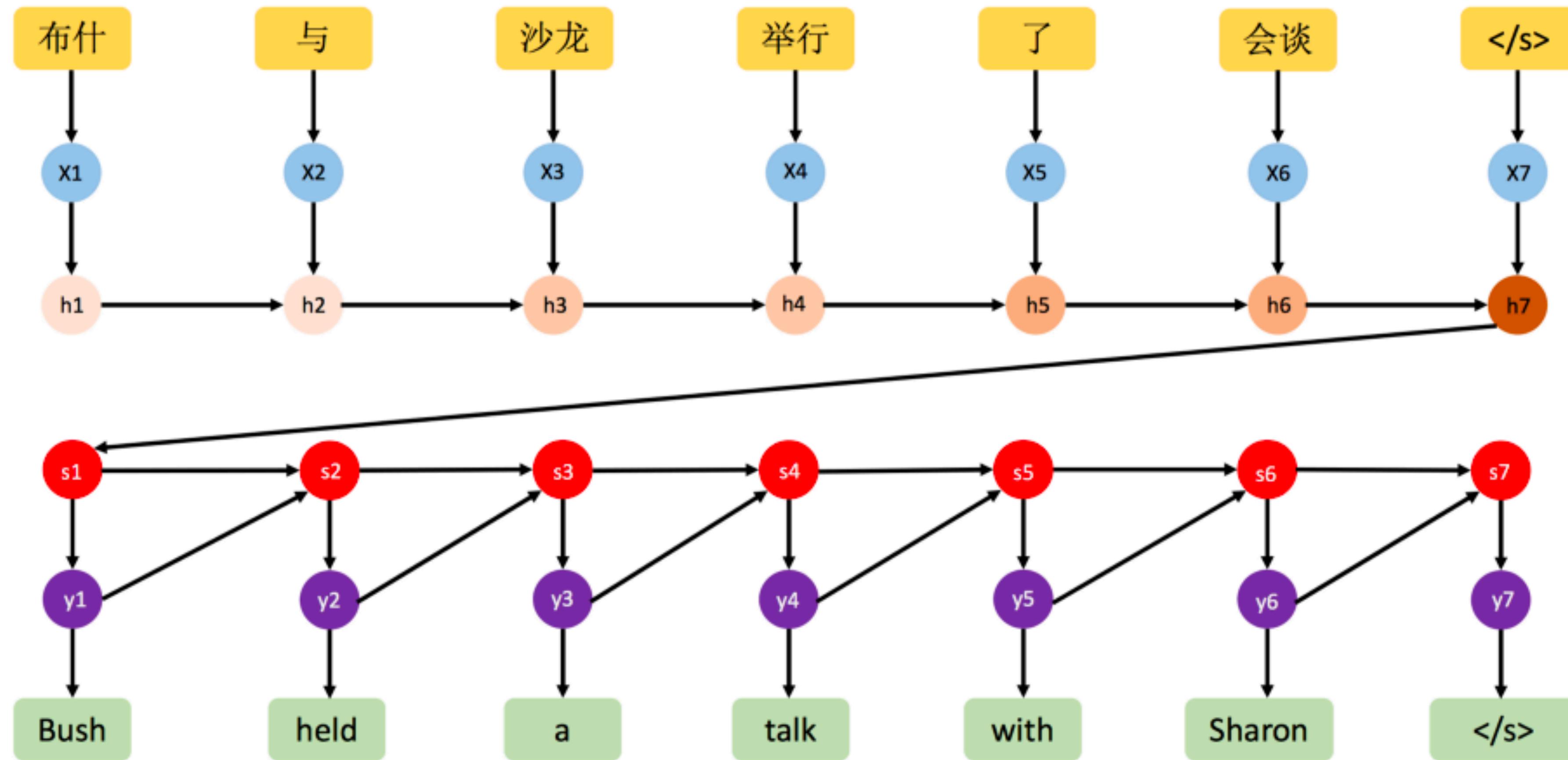
$$\hat{E} = \operatorname{argmax}_{E \in \text{English}} \underbrace{P(F | E)}_{\text{translation model}} \underbrace{P(E)}_{\text{language model}}$$

- Intuitively, a good model of English, and a good English-to-foreign translator

Training and decoding

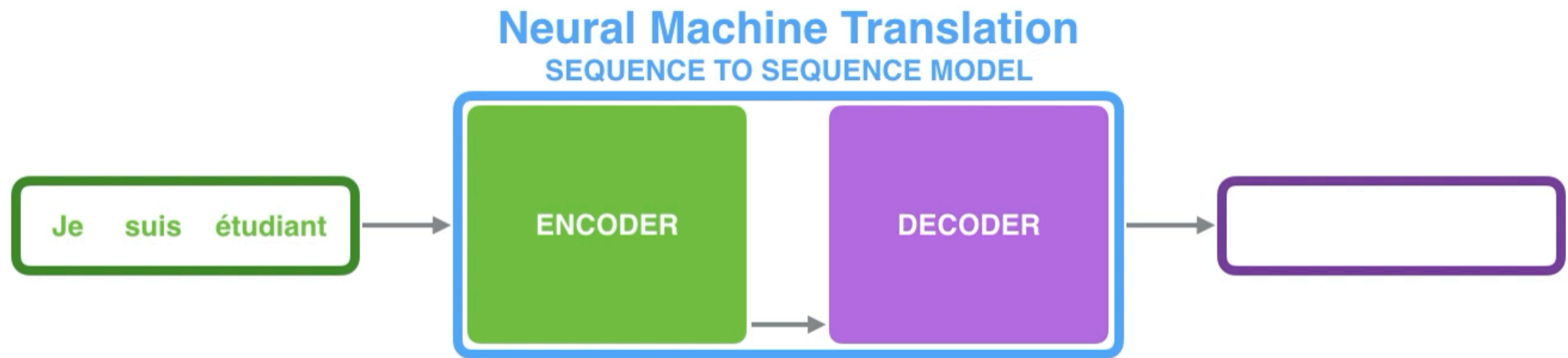


Neural MT



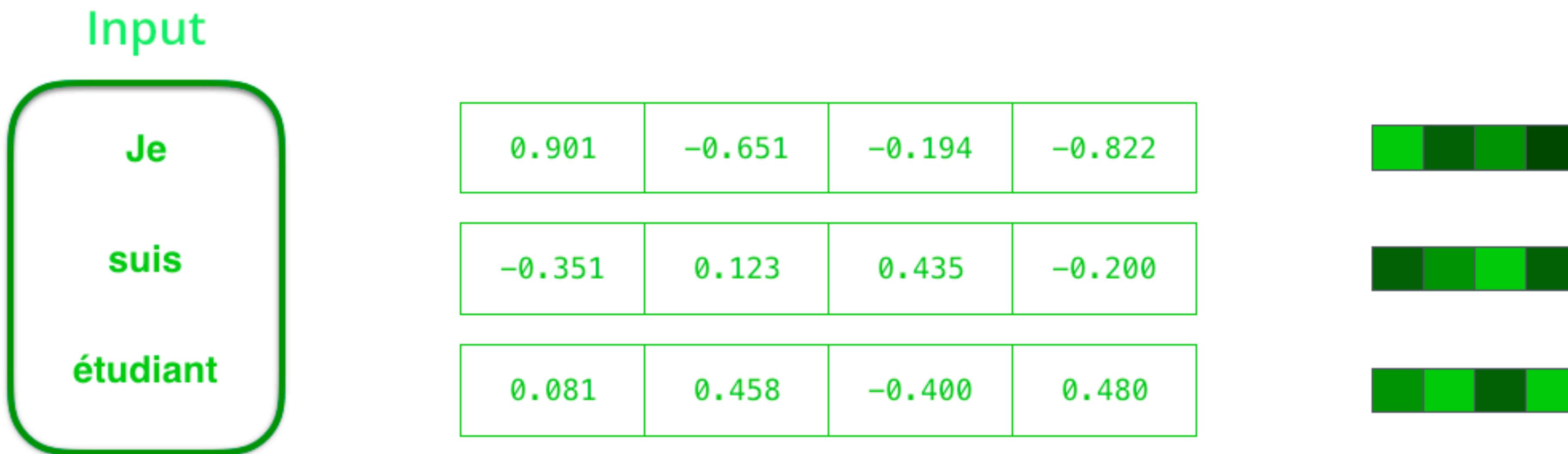
(Sutskever et al., 2014)

Encoder-decoder model



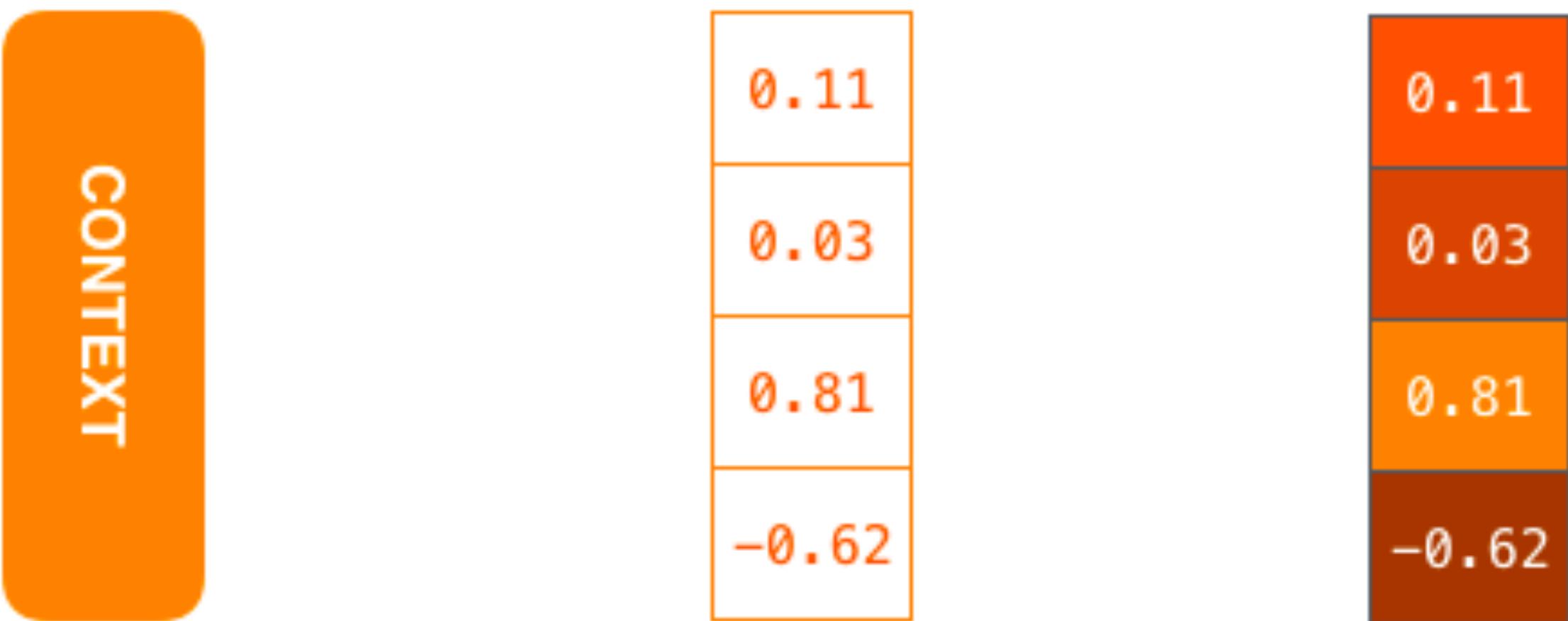
Input to the encoder

- ▶ Embedding



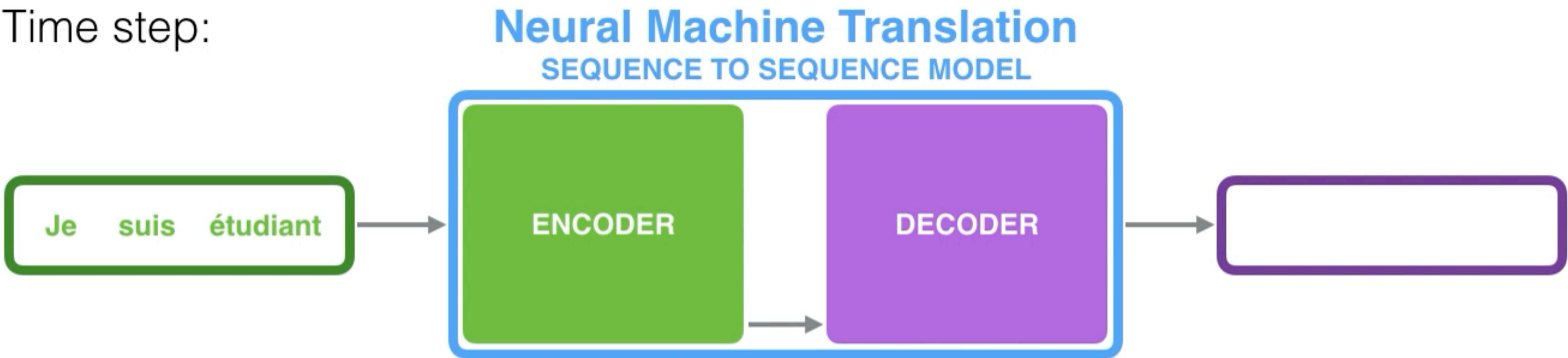
Context vector

- A continuous vector from the encoder



Seq2seq

Time step:



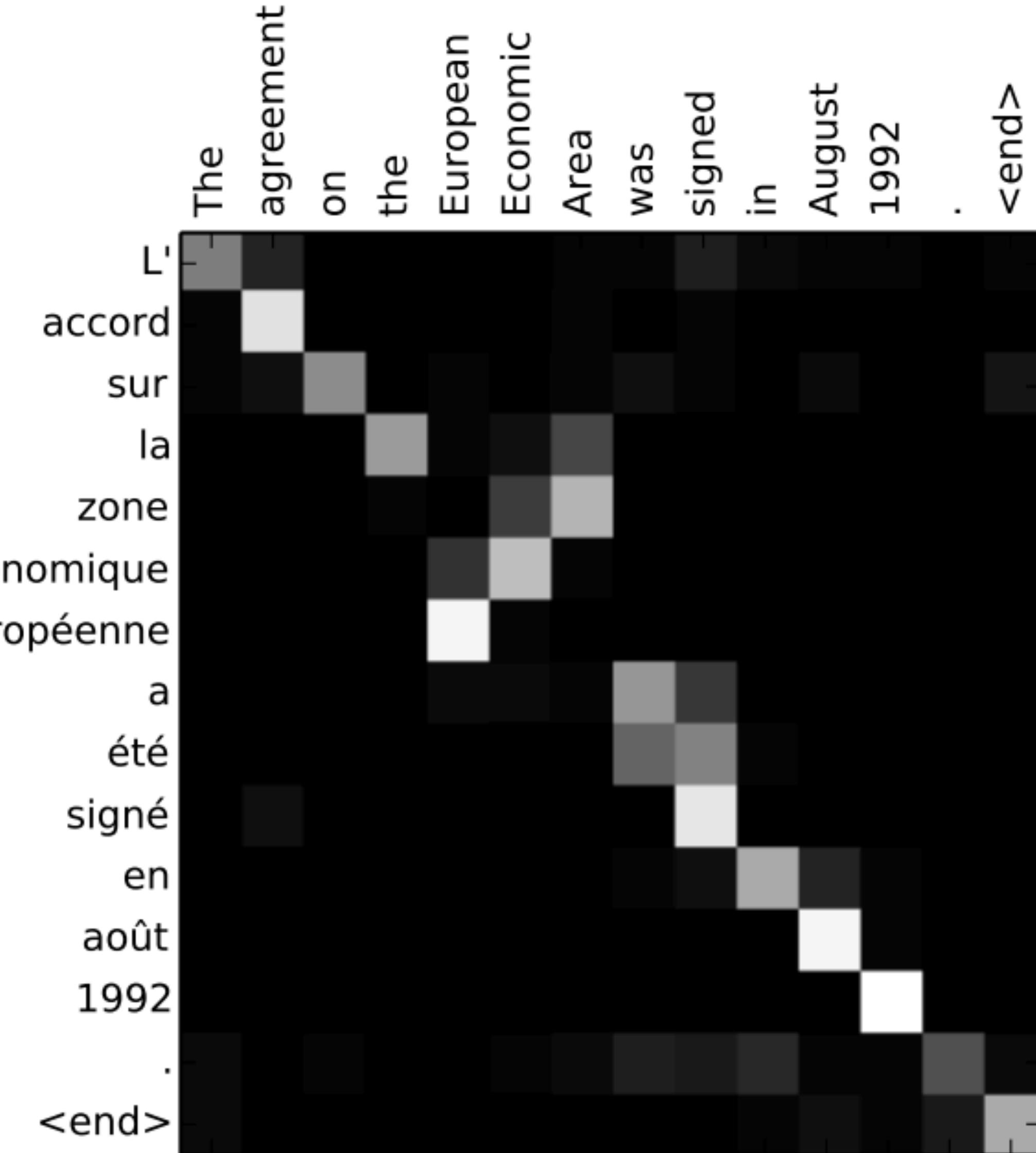
Let's pay attention

Time step: 7

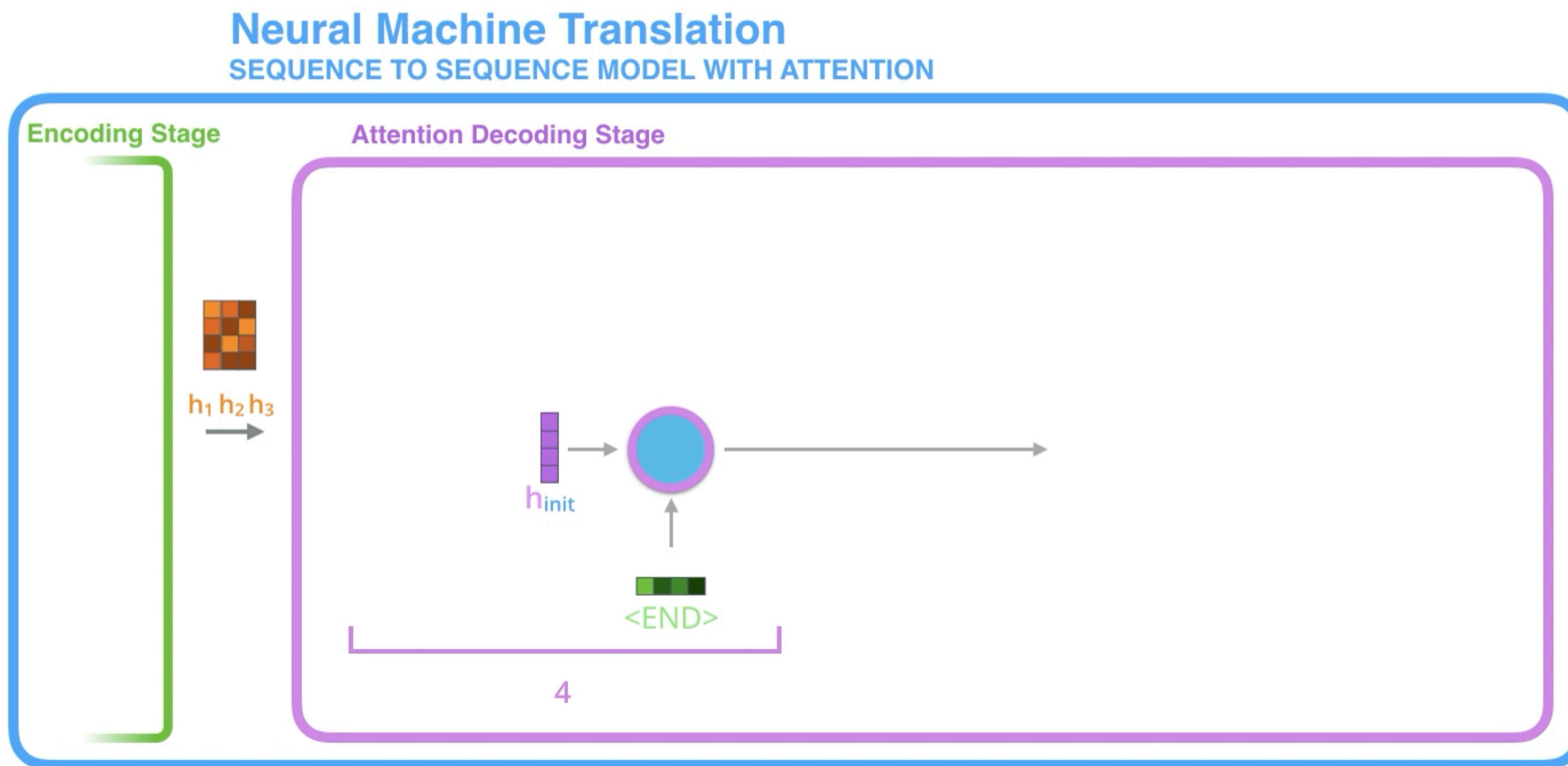
Neural Machine Translation SEQUENCE TO SEQUENCE MODEL WITH ATTENTION



Attention visualization

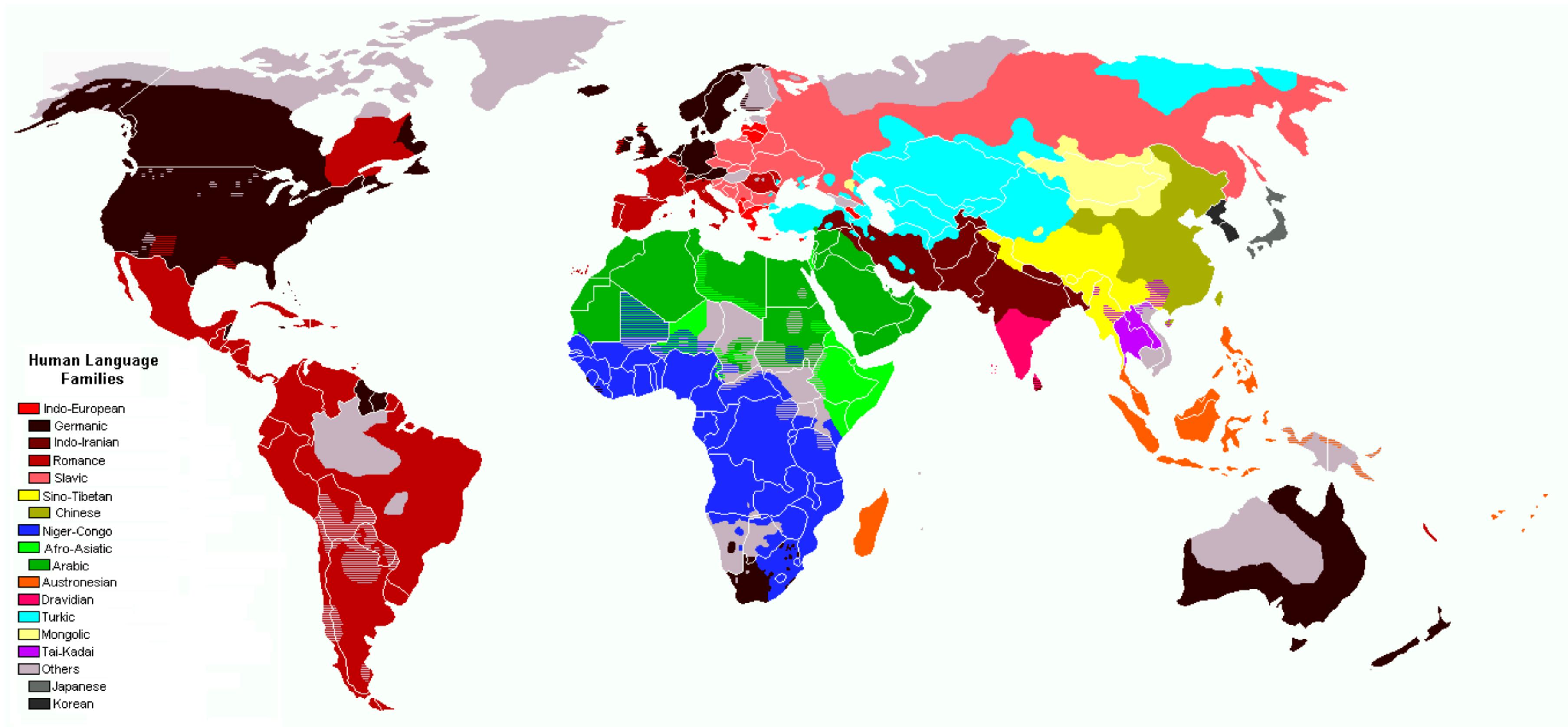


Seq2Seq with attention

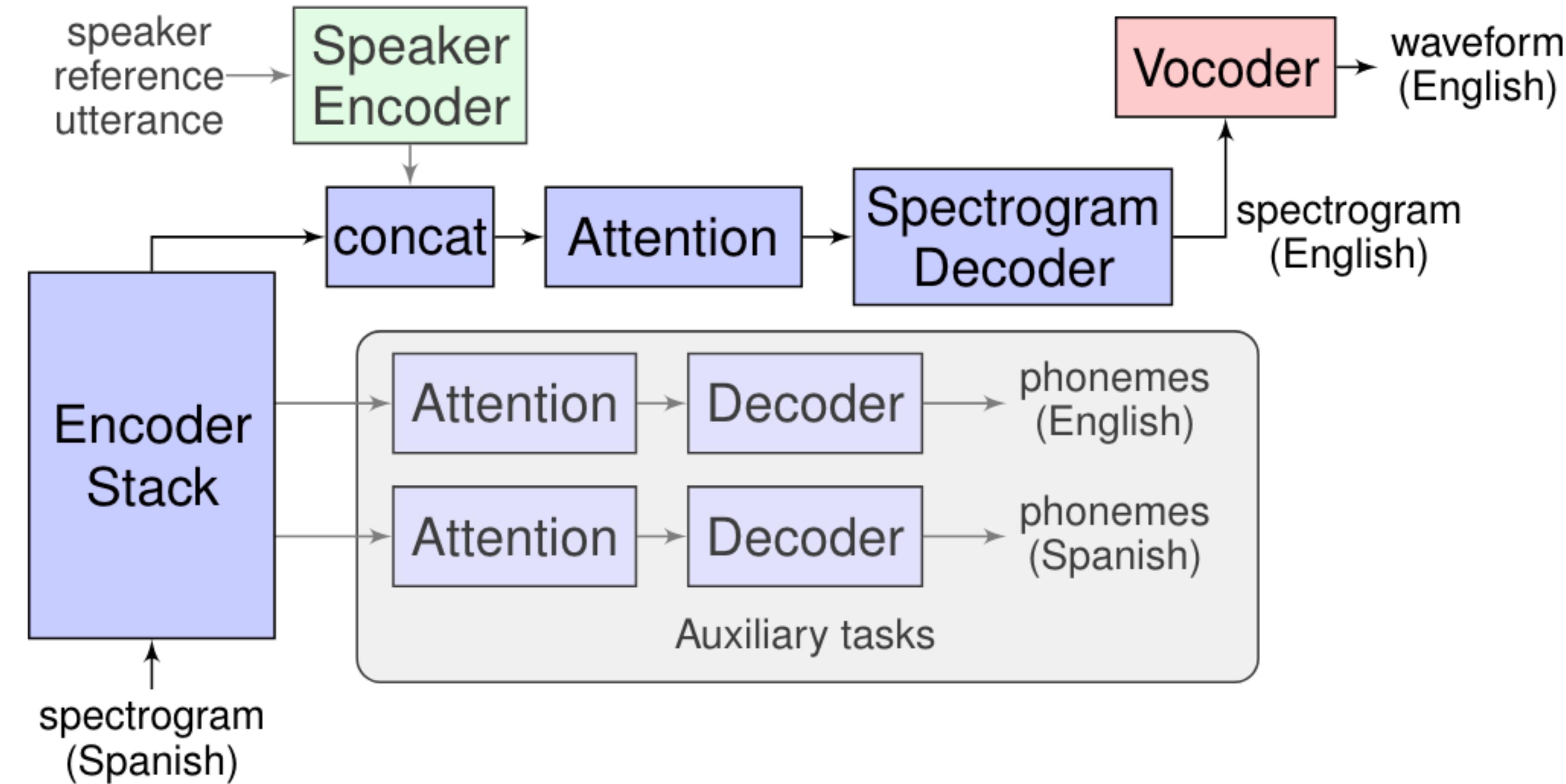


Speech-to-speech translation

- ▶ Some languages don't have a written form

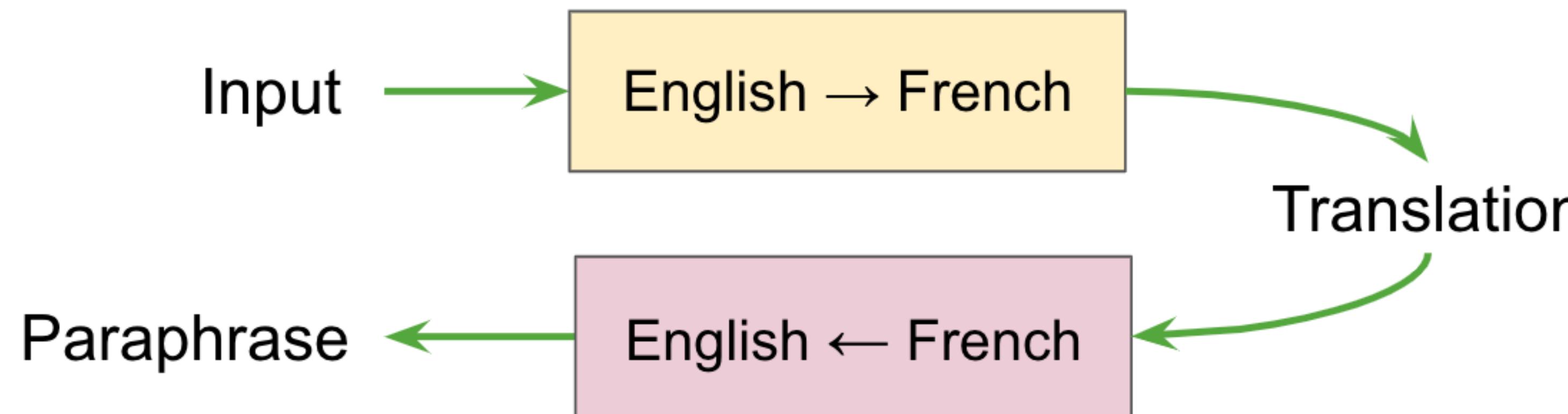


Speech-to-speech translation



Data augmentation: back translation

Previously, tea had been used primarily for Buddhist monks to stay awake during meditation.



In the past, tea was used mostly for Buddhist monks to stay awake during the meditation.

Autrefois, le thé avait été utilisé surtout pour les moines bouddhistes pour rester éveillé pendant la méditation.

Summary

- ▶ Lexical, syntactic, semantic divergences make MT difficult
- ▶ Statistical machine translation and noise channel model
- ▶ End-to-end neural machine translation
- ▶ Speech-to-speech translation for languages without written form

Reading

- ▶ Chapter 10: Machine translation
 - <https://web.stanford.edu/~jurafsky/slp3/13.pdf>