Machine Translation

Philipp Koehn

27 August 2024

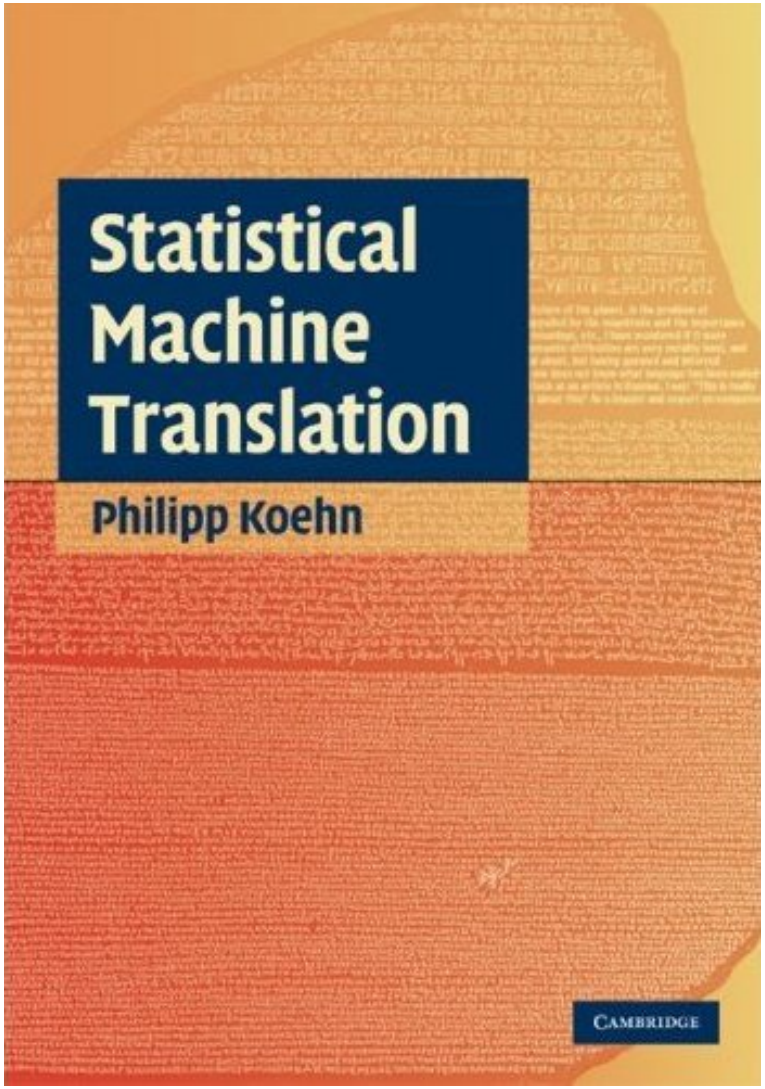
What is This?

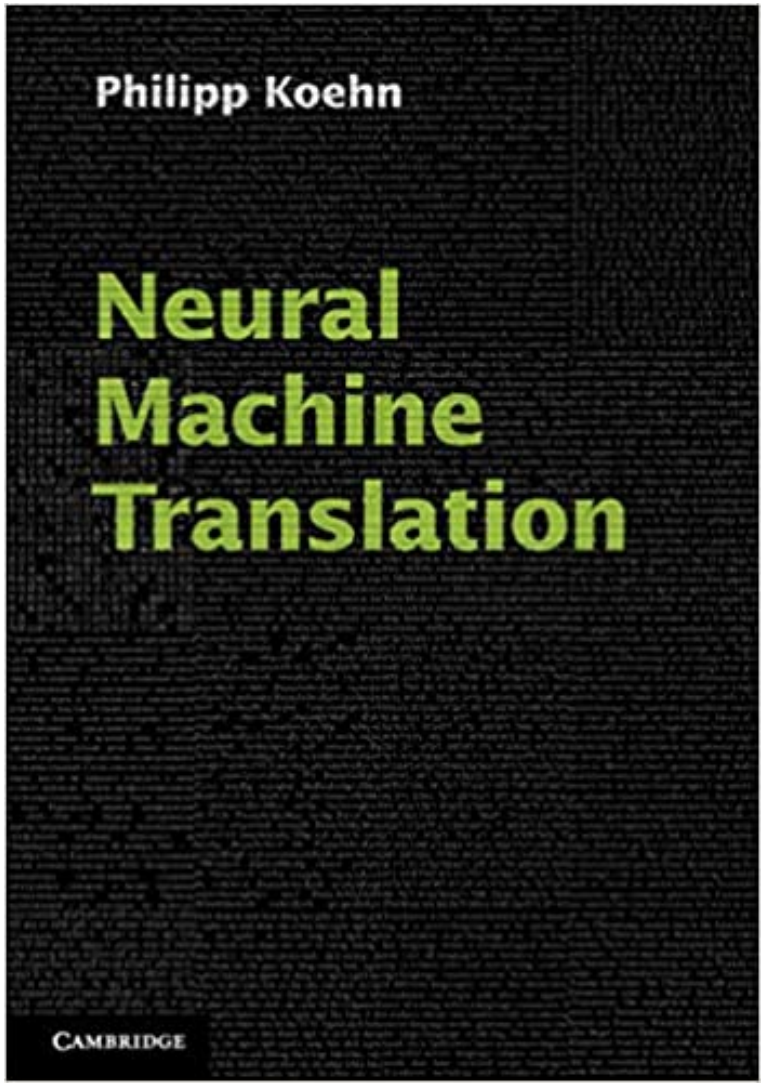
- A class on machine translation
- Taught at Johns Hopkins University, Fall 2023
- Class web site: `http://www.mt-class.org/jhu/`
- Tuesdays and Thursdays, 1:30-2:45, Hodson 210
- Instructor: Philipp Koehn
- TA: Bismarck Odoom
- Grading
  - five programming assignments (12% each)
  - final project (30%)
  - in-class presentation: language in ten minutes (10%)

Why Take This Class?

- Close look at an **artificial intelligence** problem
- Practical introduction to **natural language processing**
- Introduction to **deep learning** for structured prediction

Textbooks


Statistical
Machine
Translation
Philipp Koehn
CAMBRIDGE


Philipp Koehn
Neural
Machine
Translation
CAMBRIDGE

some history

An Old Idea

Warren Weaver on translation
as code breaking (1947):

> *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".*

Early Efforts and Disappointment

- Excited research in 1950s and 1960s

1954
Georgetown experiment
Machine could translate
250 words and
6 grammar rules

- 1966 ALPAC report:
  - only $20 million spent on translation in the US per year
  - no point in machine translation

Rule-Based Systems

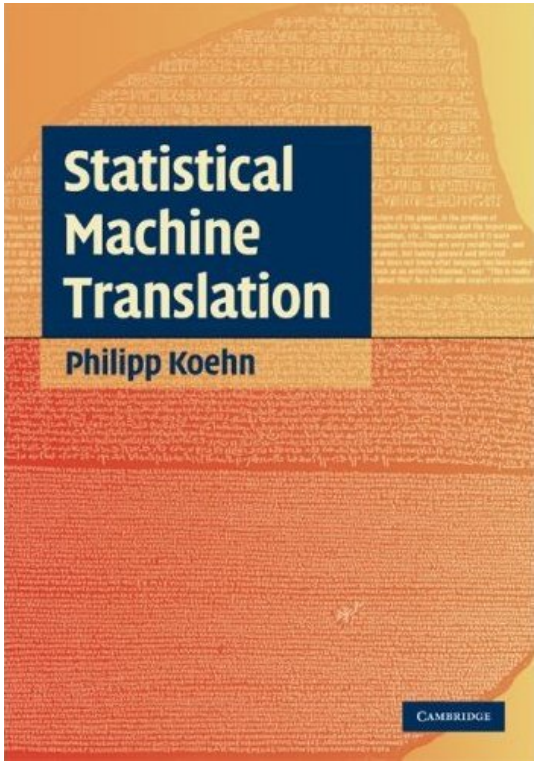
- Rule-based systems
  - build dictionaries
  - write transformation rules
  - refine, refine, refine
- Météo system for weather forecasts (1976)
- Systran (1968), Logos and Metal (1980s)

```
"have" :=

if
   subject(animate)
   and object(owned-by-subject)
then
   translate to "kade...  aahe"
if
   subject(animate)
   and object(kinship-with-subject)
then
   translate to "laa...  aahe"
if
   subject(inanimate)
then
          translate to "madhye...
aahe"
```

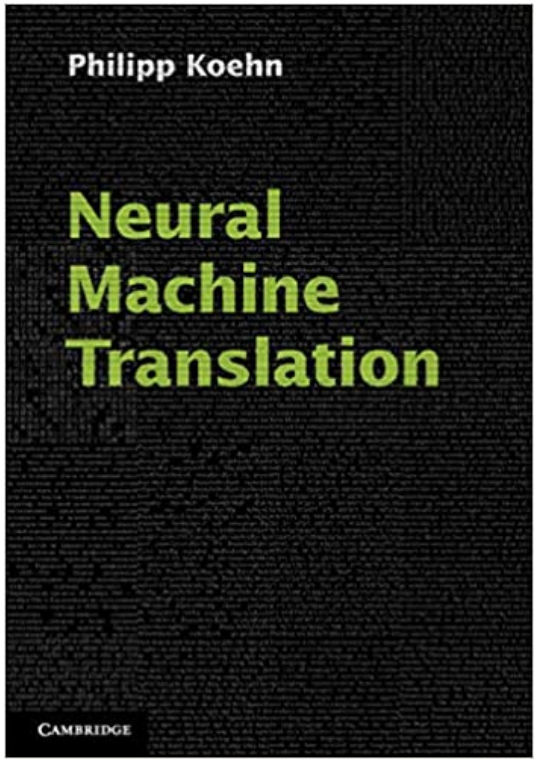
Statistical Machine Translation

- 1980s: IBM
- 1990s: increased research
- Mid 2000s: Phrase-Based MT (Moses, Google)
- Around 2010: commercial viability

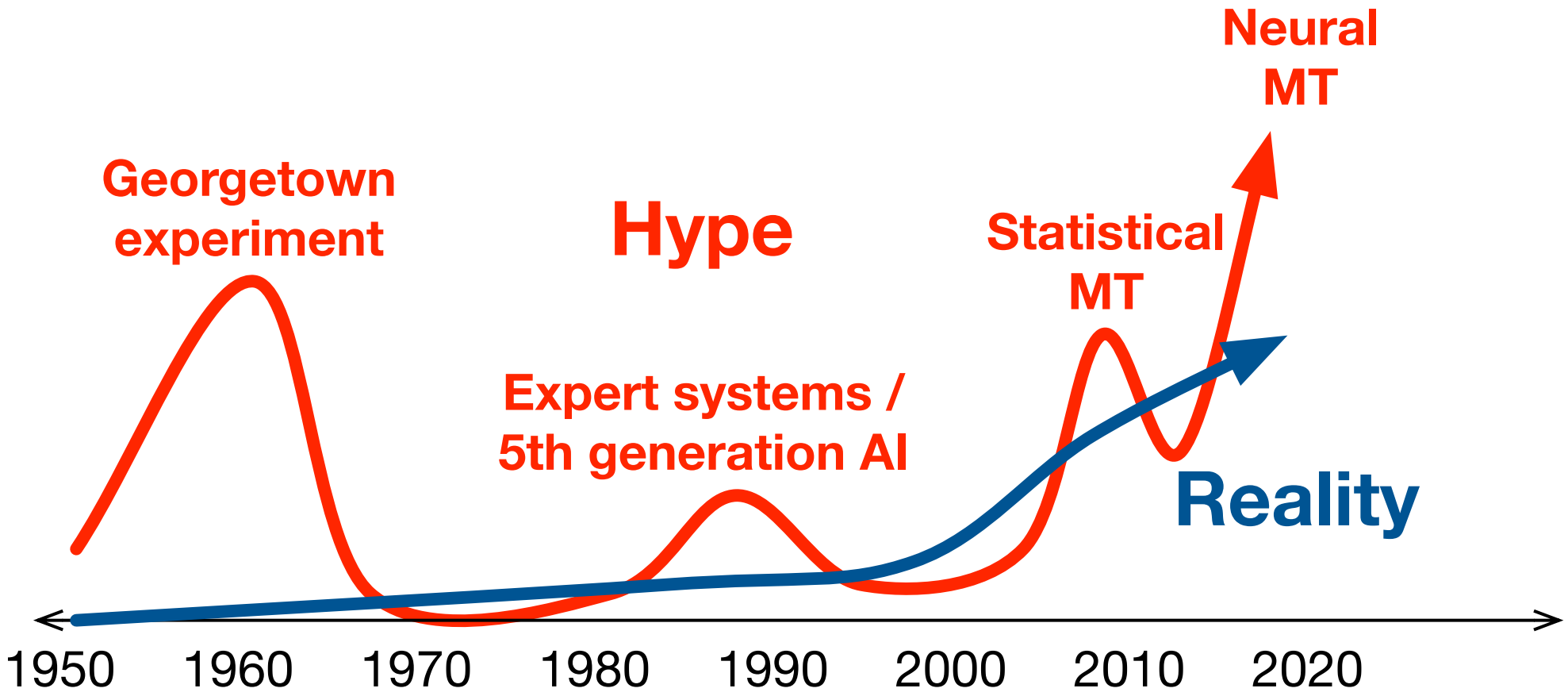


Neural Machine Translation

- Late 2000s: neural models for computer vision
- Since mid 2010s: neural models for machine translation
- 2016: Neural machine translation the new state of the art



Hype


Neural
MT
Georgetown
experiment
Hype
Statistical
MT
Expert systems /
5th generation AI
Reality
1950
1960
1970
1980
1990
2000
2010
2020

how good is machine translation?

Machine Translation: Chinese

记者从环保部了解到，《水十条》要求今年年底前直辖市、省会城市、计划单列市建成区基本解决黑臭水体。截至目前，全国224个地级及以上城市共排查确认黑臭水体2082个，其中34.9%完成整治，28.4%正在整治，22.8%正在开展项目前期。

Reporters learned from the Ministry of Environmental Protection, "Water 10" requirements before the end of this year before the municipality, the provincial capital city, plans to build a separate city to solve the basic black and black water. Up to now, the country's 224 prefecture-level and above cities were identified to confirm the black and white water 2082, of which 34.9% to complete the renovation, 28.4% is remediation, 22.8% is carrying out the project early.

Machine Translation: French

A l'orée de ce débat télévisé inédit dans l'histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c'est la candidate du Front national qui s'est retrouvée au cœur des premières attaques de ses quatre adversaires d'un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d'immigration et de laïcité.

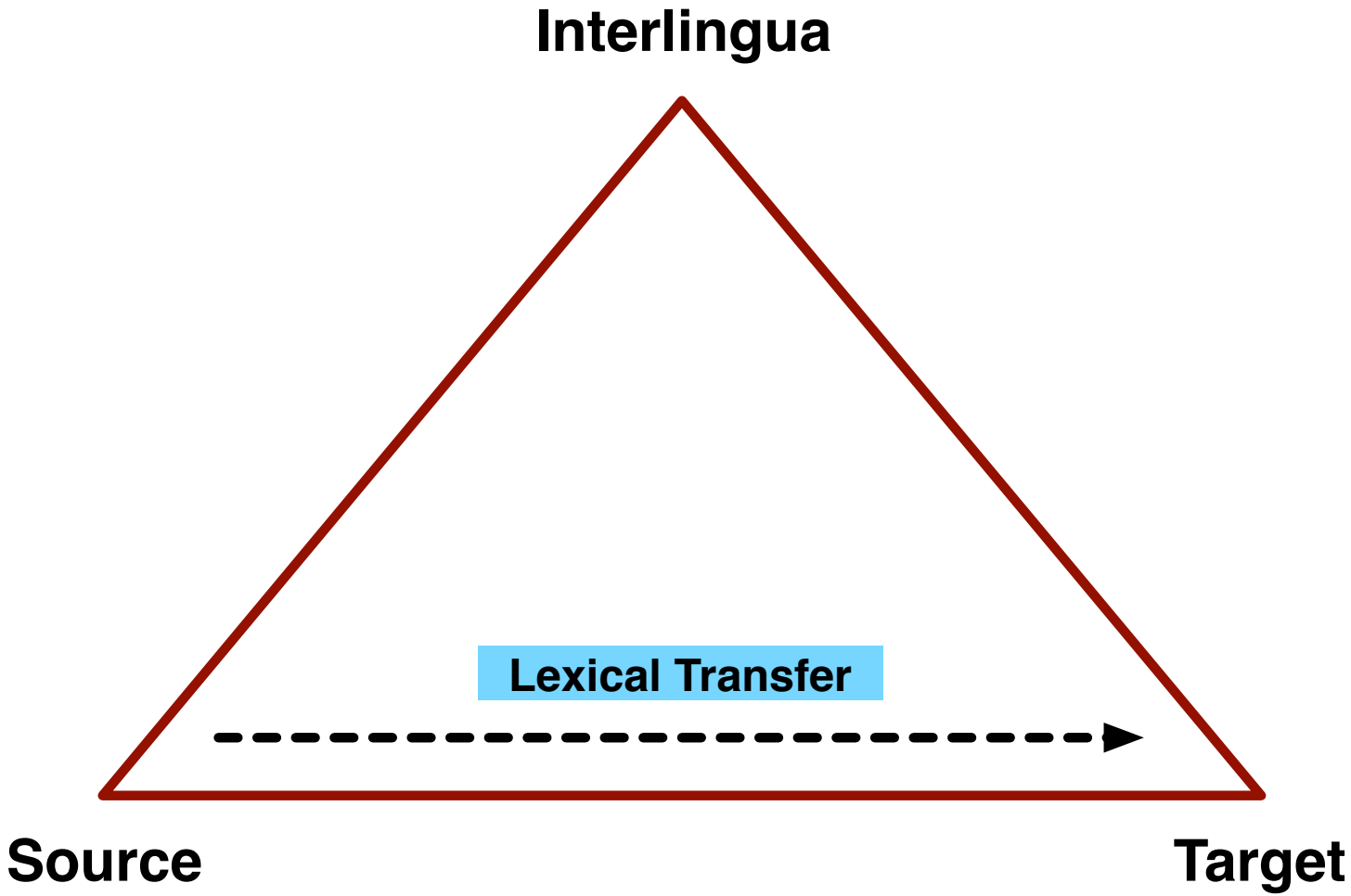
At the beginning of this televised debate, which was unheard of in the history of the Fifth Republic, a "Tous sur Macron" was expected, but it was the candidate of the National Front who found itself at the heart of the first attacks of its four Opponents of one evening, favored by the first theme tackled, the issues of society and thus security, immigration and secularism.

Machine Translation: French (2022)

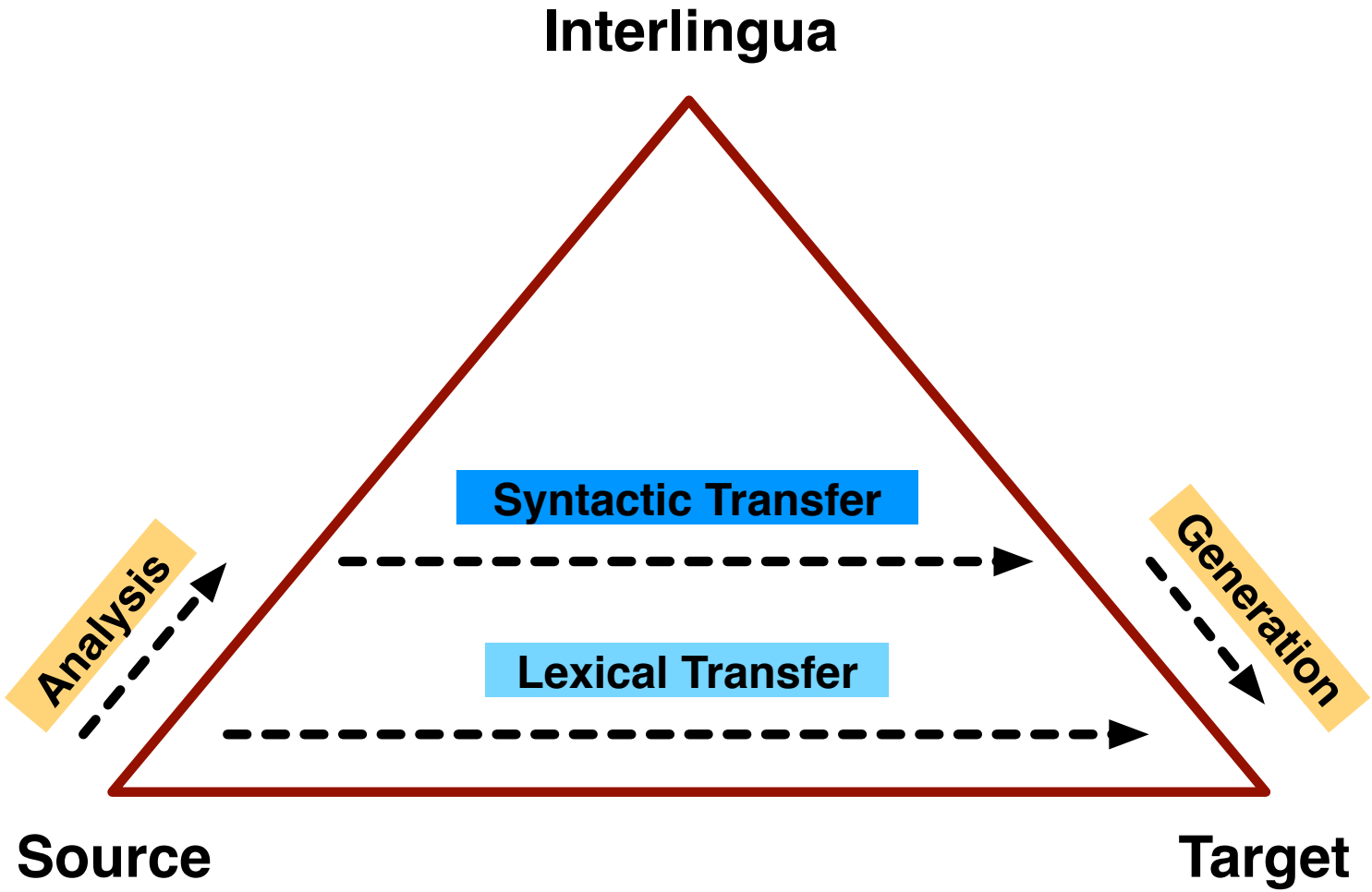
A l'orée de ce débat télévisé inédit dans l'histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c'est la candidate du Front national qui s'est retrouvée au cœur des premières attaques de ses quatre adversaires d'un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d'immigration et de laïcité.

At the start of this unprecedented televised debate in the history of the Vé République, we expected a form of "All on Macron" but it was the Candidate of the National Front who found herself at the heart of the first attacks of her four adversaries for one evening, favored by the first theme addressed, questions of society and therefore of security, immigration and secularism.

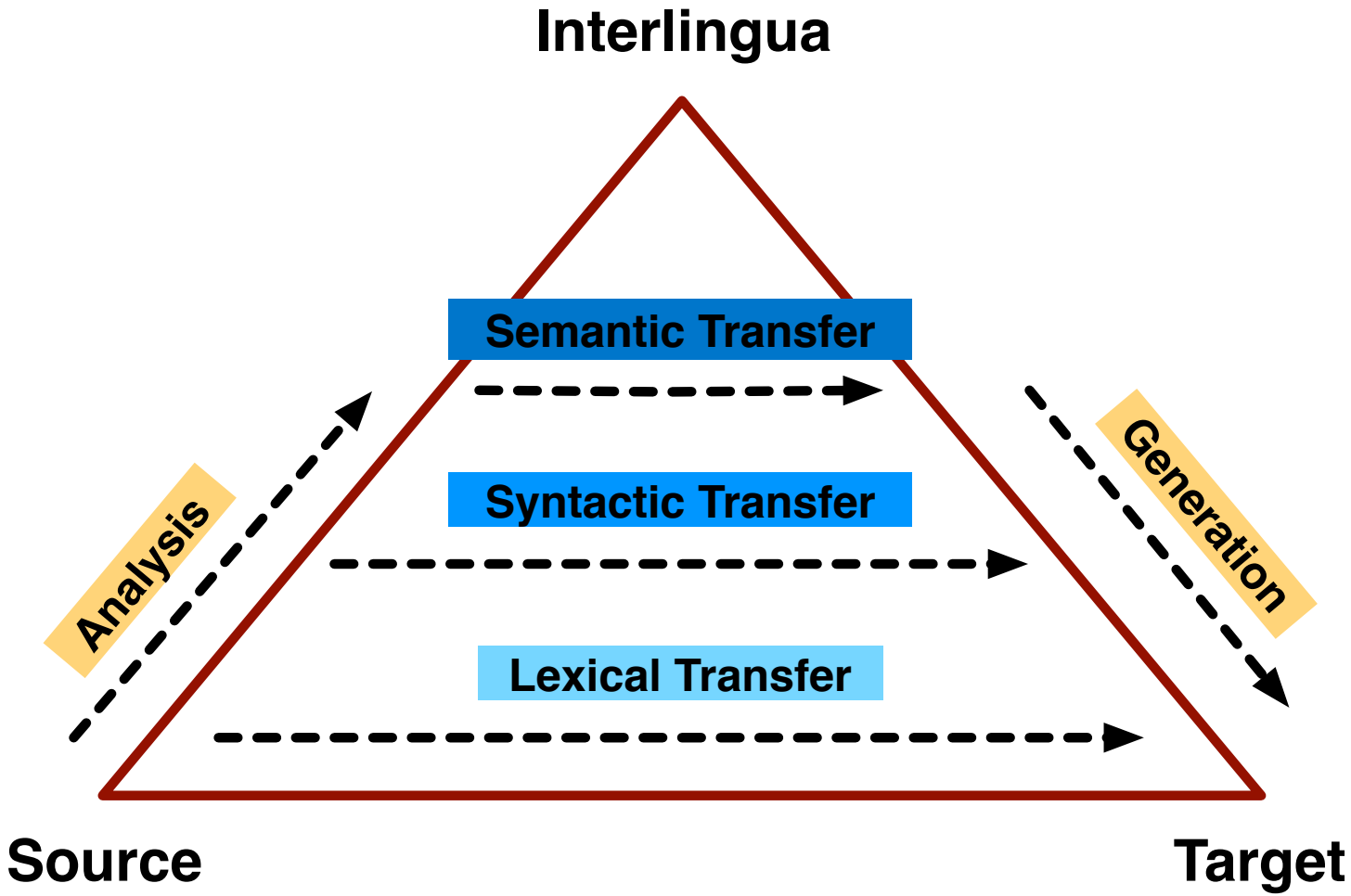
A Clear Plan



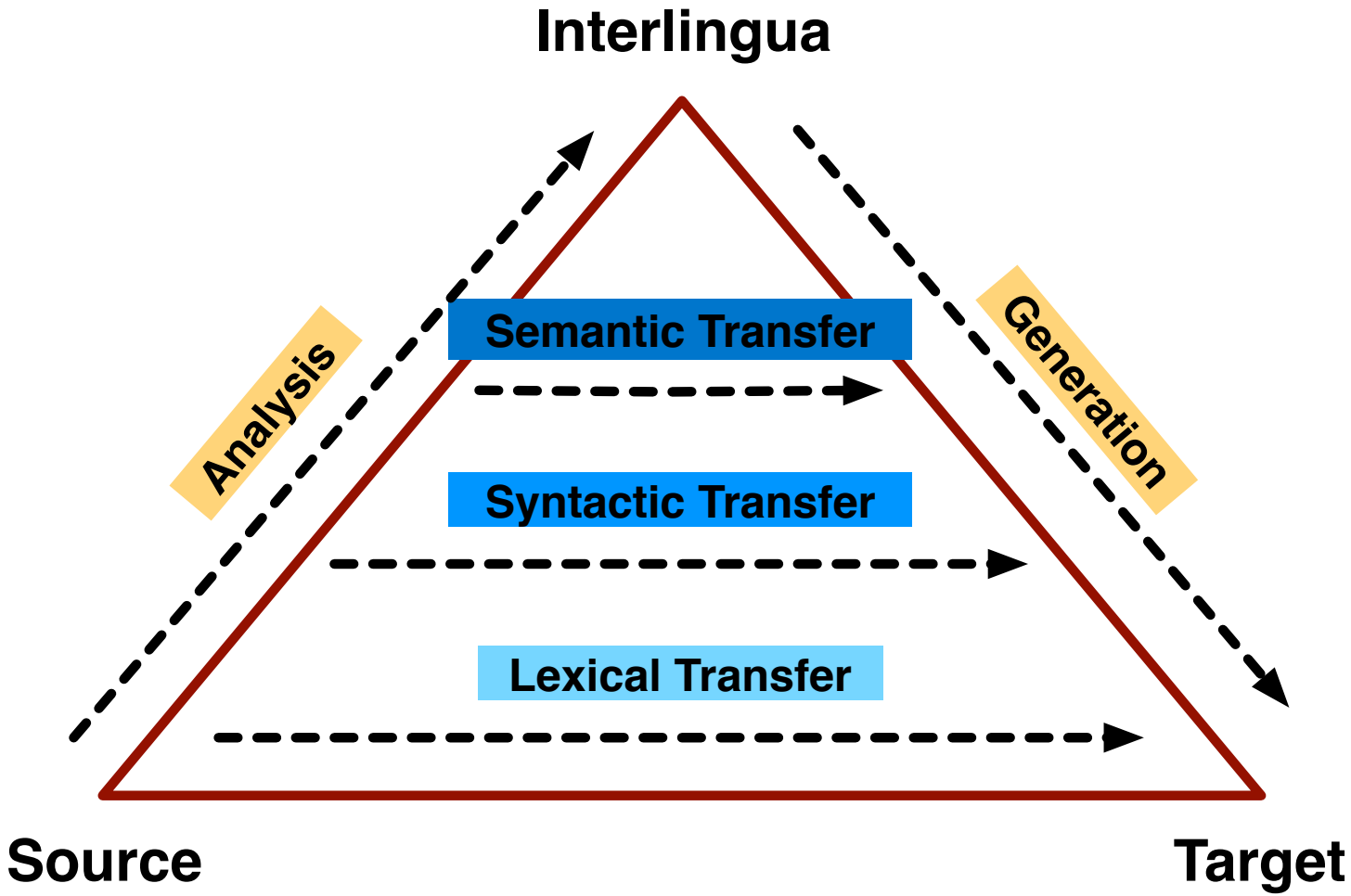
A Clear Plan


Interlingua
Syntactic Transfer
Analysis
Generation
Lexical Transfer
Source
Target

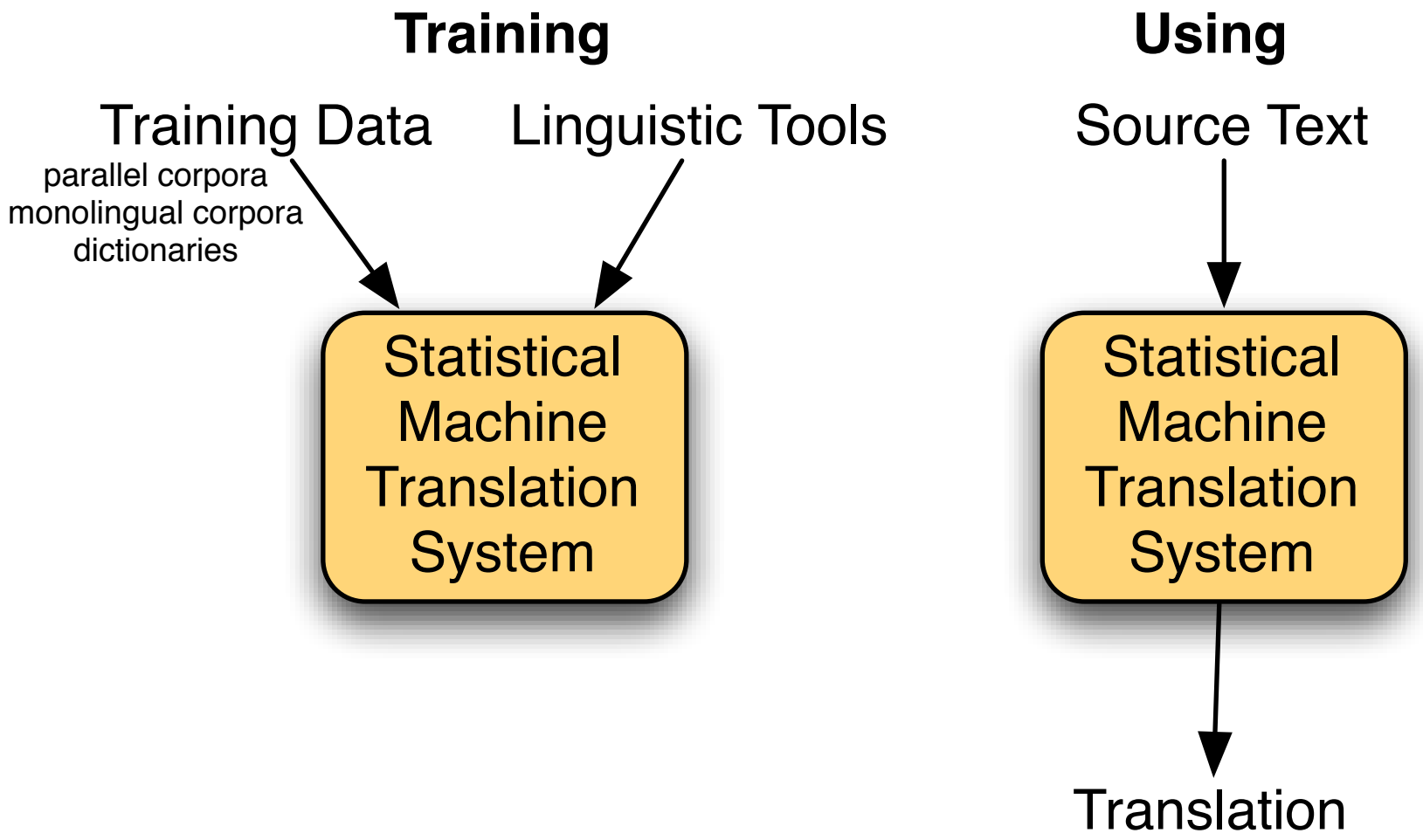
A Clear Plan


Interlingua
Semantic Transfer
Syntactic Transfer
Lexical Transfer
Analysis
Generation
Source
Target

A Clear Plan


Interlingua
Analysis
Semantic Transfer
Generation
Syntactic Transfer
Lexical Transfer
Source
Target

Learning from Data



why is that a good plan?

Word Translation Problems

- Words are ambiguous

He deposited money in a bank account
with a high interest rate.

Sitting on the bank of the Mississippi,
a passing ship piqued his interest.

- How do we find the right meaning, and thus translation?
- Context should be helpful

Syntactic Translation Problems

- Languages have different sentence structure

das	behaupten	sie	wenigstens
this	claim	they	at least
the		she	

- Convert from object-verb-subject (OVS) to subject-verb-object (SVO)

- Ambiguities can be resolved through syntactic analysis
  - the meaning the of das not possible (not a noun phrase)
  - the meaning she of sie not possible (subject-verb agreement)

Semantic Translation Problems

- Pronominal anaphora

I saw the movie and it is good.

- How to translate it into German (or French)?
  - it refers to movie
  - movie translates to Film
  - Film has masculine gender
  - ergo: it must be translated into masculine pronoun er

- We are not handling this very well [Le Nagard and Koehn, 2010]

Semantic Translation Problems

- Coreference

Whenever I visit my uncle and his daughters,
I can't decide who is my favorite cousin.

- How to translate cousin into German? Male or female?

- Complex inference required

Semantic Translation Problems

- Discourse

Since you brought it up, I do not agree with you.

Since you brought it up, we have been working on it.

- How to translated since? Temporal or conditional?

- Analysis of discourse structure — a hard problem

Learning from Data

- What is the best translation?

Sicherheit → security
Sicherheit → safety
Sicherheit → certainty

Learning from Data

- What is the best translation?

Sicherheit → security 14,516
Sicherheit → safety 10,015
Sicherheit → certainty 334

- Counts in European Parliament corpus

Learning from Data

- What is the best translation?

Sicherheit → security 14,516
Sicherheit → safety 10,015
Sicherheit → certainty 334

- Phrasal rules

Sicherheitspolitik → security policy 1580
Sicherheitspolitik → safety policy 13
Sicherheitspolitik → certainty policy 0

Lebensmittelsicherheit → food security 51
Lebensmittelsicherheit → food safety 1084
Lebensmittelsicherheit → food certainty 0

Rechtssicherheit → legal security 156
Rechtssicherheit → legal safety 5
Rechtssicherheit → legal certainty 723

Learning from Data

- What is most fluent?

a problem for translation

a problem of translation

a problem in translation

Learning from Data

- What is most fluent?

a problem for translation 13,000

a problem of translation 61,600

a problem in translation 81,700

- Hits on Google

Learning from Data

- What is most fluent?

a problem for translation 13,000

a problem of translation 61,600

a problem in translation 81,700

a translation problem 235,000

Learning from Data

- What is most fluent?

police disrupted the demonstration

police broke up the demonstration

police dispersed the demonstration

police ended the demonstration

police dissolved the demonstration

police stopped the demonstration

police suppressed the demonstration

police shut down the demonstration

Learning from Data

- What is most fluent?

police disrupted the demonstration 2,140

police broke up the demonstration 66,600

police dispersed the demonstration 25,800

police ended the demonstration 762

police dissolved the demonstration 2,030

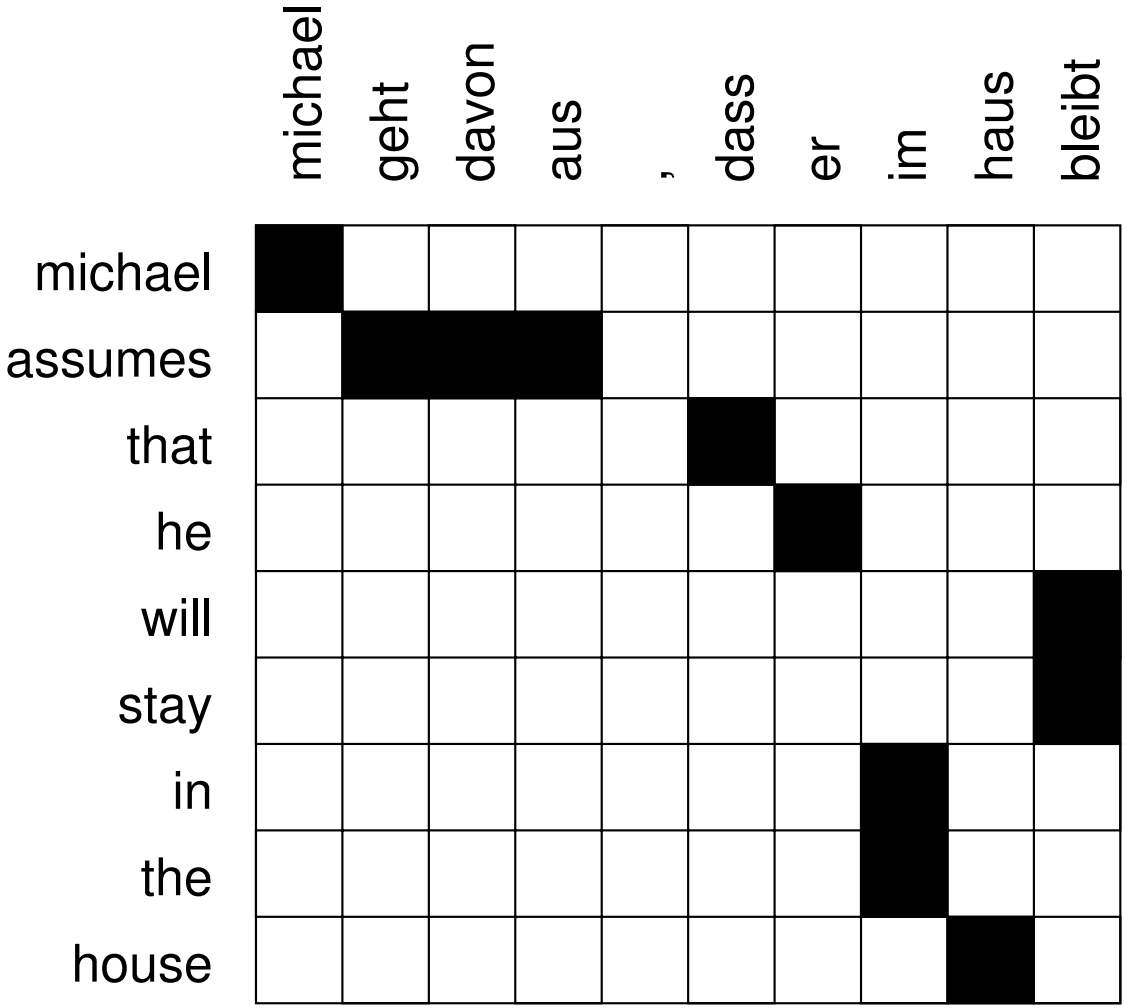
police stopped the demonstration 722,000

police suppressed the demonstration 1,400

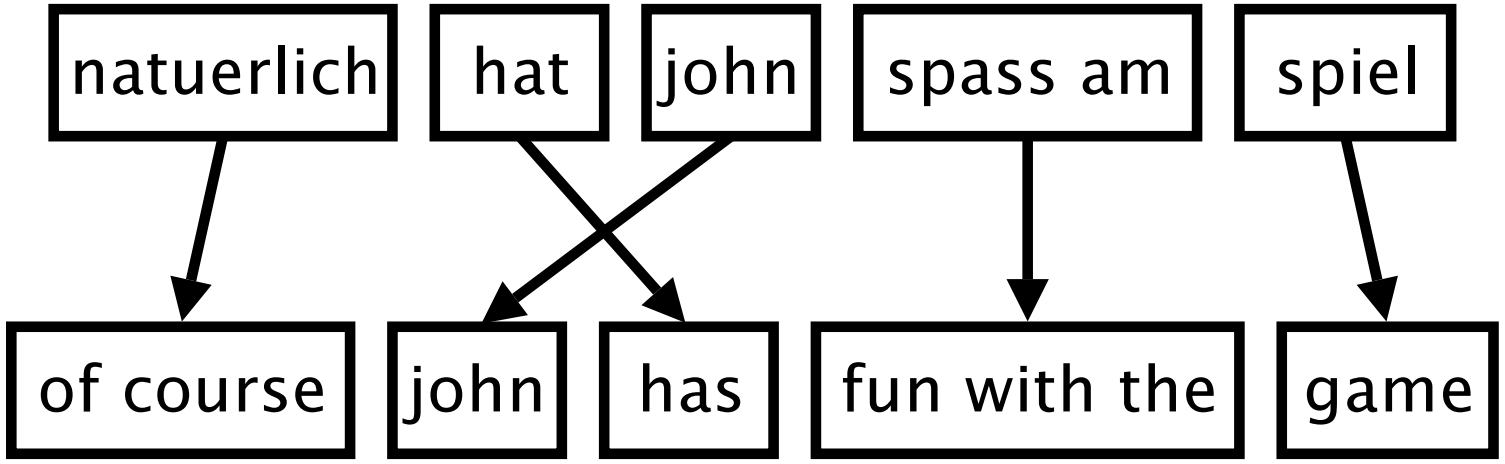
police shut down the demonstration 2,040

where are we now?

Word Alignment



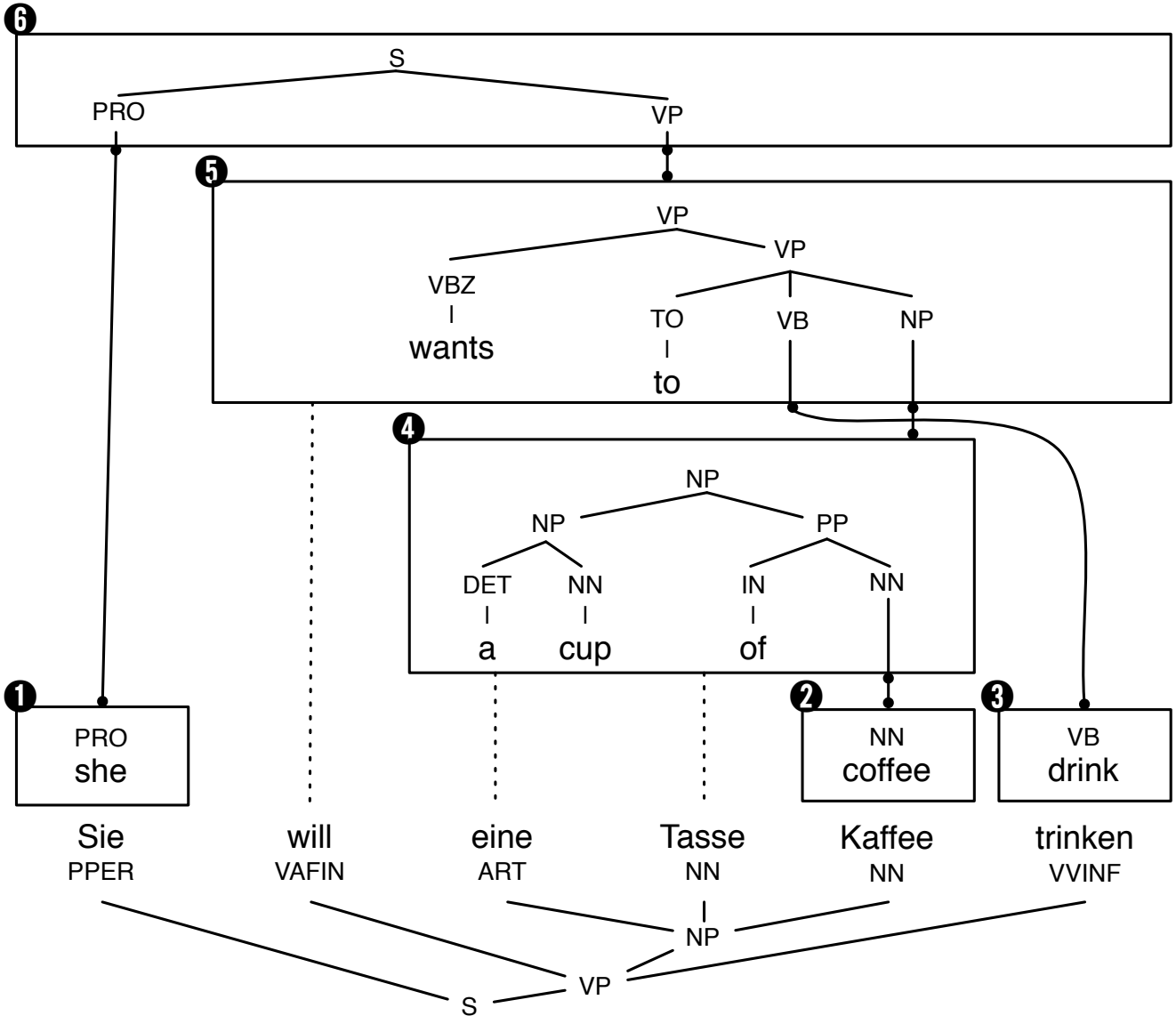
Phrase-Based Model



- Foreign input is segmented in phrases
- Each phrase is translated into English
- Phrases are reordered

- Workhorse of today's statistical machine translation

Syntax-Based Translation



Semantic Translation

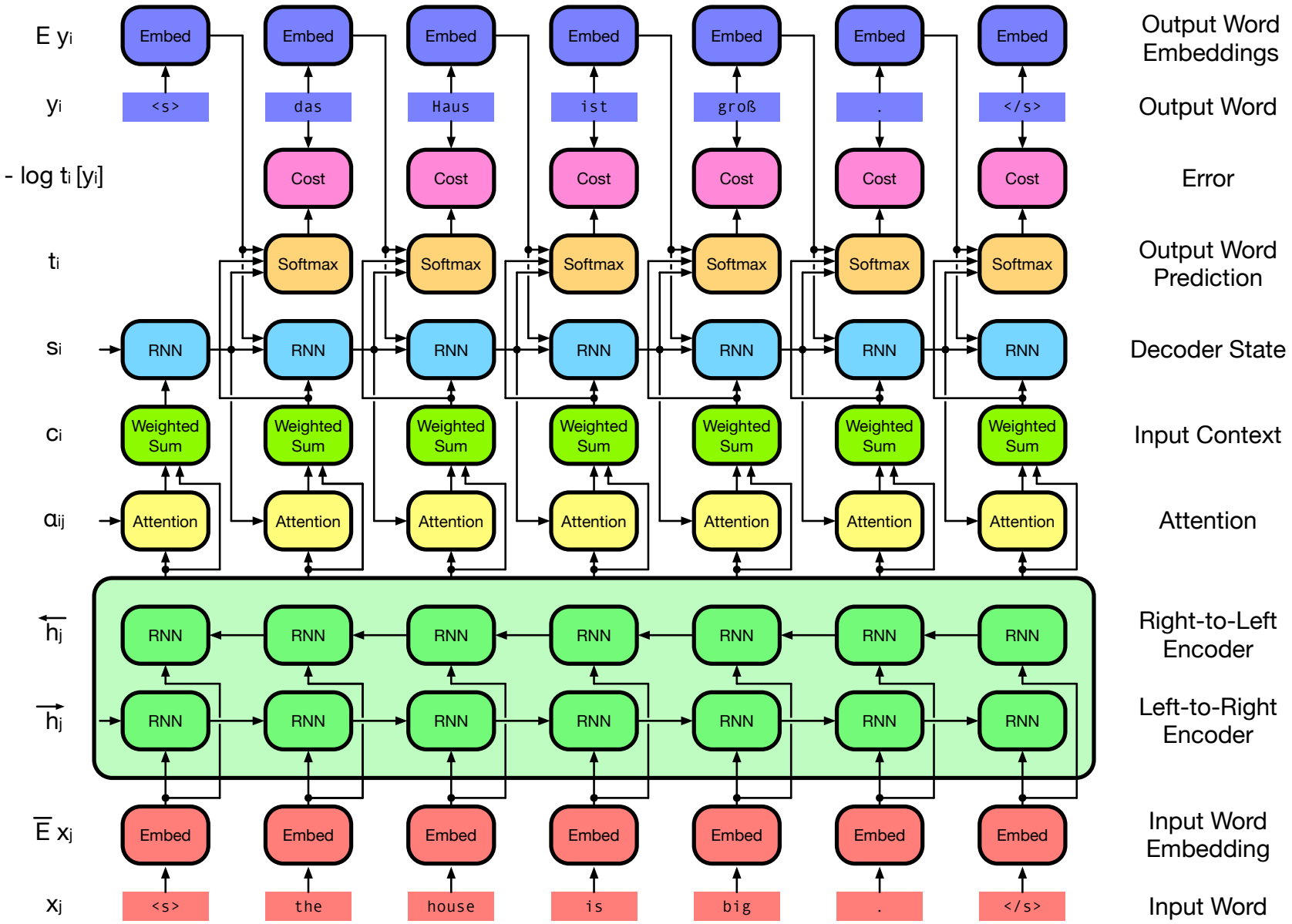
- Abstract meaning representation [Knight et al., ongoing]

```
(w / want-01
   :agent (b / boy)
   :theme (l / love
        :agent (g / girl)
        :patient b))
```

- Generalizes over equivalent syntactic constructs
  (e.g., active and passive)

- Defines semantic relationships
  - semantic roles
  - co-reference
  - discourse relations

Neural Model



Machine Translation and Language Models

- State-of-the-art model for machine translation: Transformer
- Transformer model was also adopted for language modeling
- Currently, large language models being built by major IT companies (GPT4, Llama, Gemini, ...)
- Latest approach: fine-tuning large language models for machine translation

what is it good for?

what is it good *enough* for?

Why Machine Translation?

Assimilation — reader initiates translation, wants to know content

- user is tolerant of inferior quality
- focus of majority of research (GALE program, etc.)

Communication — participants don't speak same language, rely on translation

- users can ask questions, when something is unclear
- chat room translations, hand-held devices
- often combined with speech recognition, IWSLT campaign

Dissemination — publisher wants to make content available in other languages

- high demands for quality
- currently almost exclusively done by human translators

Problem: No Single Right Answer

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Quality

HTER	**assessment**
0%	
	publishable
10%	
	editable
20%	
30%	gistable
40%	triagable
50%	

(scale developed in preparation of DARPA GALE programme)

Applications

HTER	assessment	application examples
0%		Seamless bridging of language divide
	publishable	Automatic publication of official announcements
10%		
	editable	Increased productivity of human translators
20%		Access to official publications
		Multi-lingual communication (chat, social networks)
30%	gistable	Information gathering
		Trend spotting
40%	triagable	Identifying relevant documents
50%		

Current State of the Art

HTER	assessment	language pairs and domains
0%		French-English restricted domain
	publishable	French-English news stories
10%		German-English news stories
	editable	Chinese-English news stories
20%		
30%	gistable	Swahili–English news stories
40%	triagable	Uyghur–English news stories
50%		

(informal rough estimates by presenter)

Thank You

questions?