Machine Translation

Course web site

mt-class.org/penn

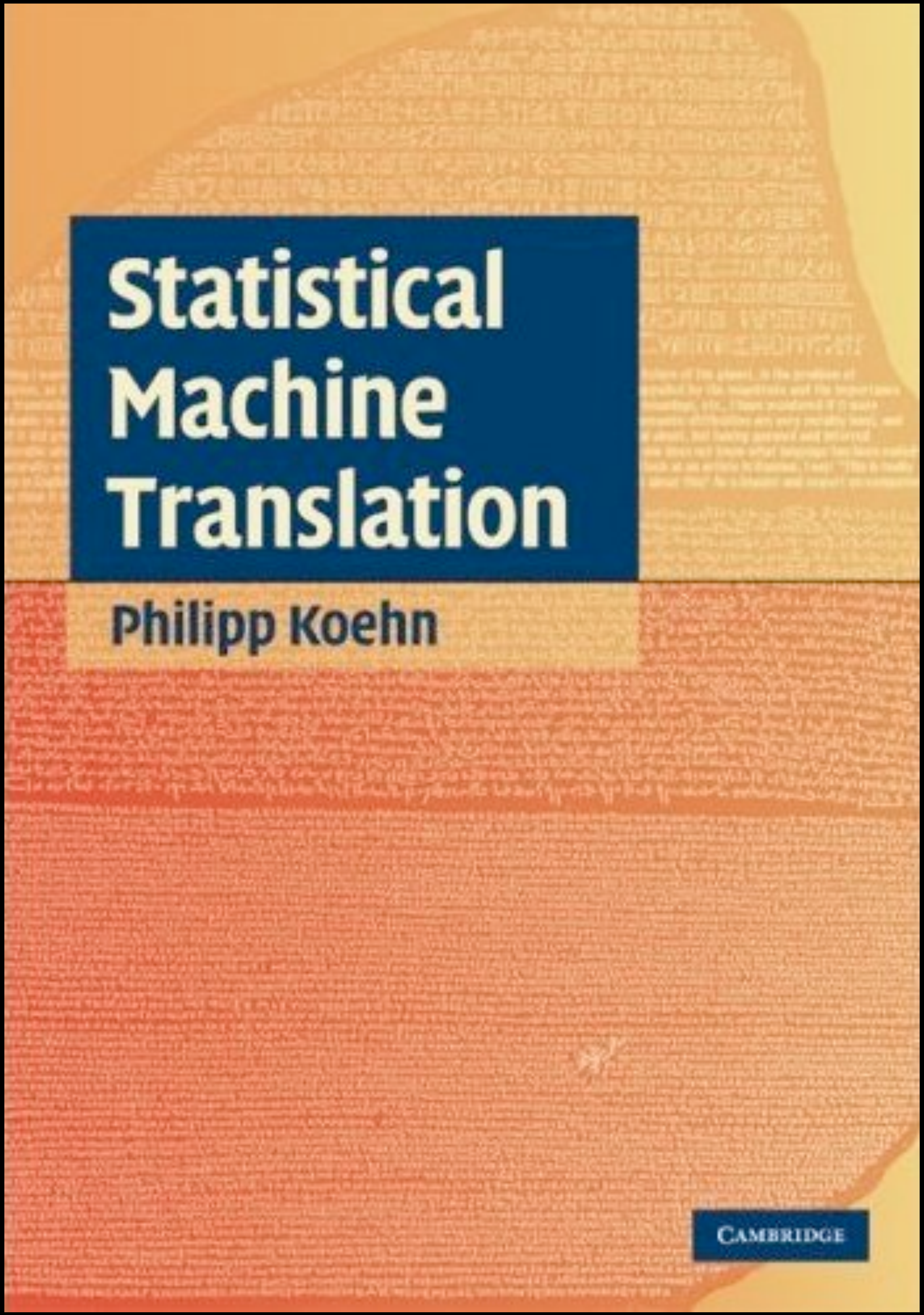
Course materials developed with

Adam Lopez
Edinburgh

Matt Post
JHU

Chris Dyer
DeepMind

Textbook


Statistical Machine Translation
Philipp Koehn
CAMBRIDGE

Draft Textbook

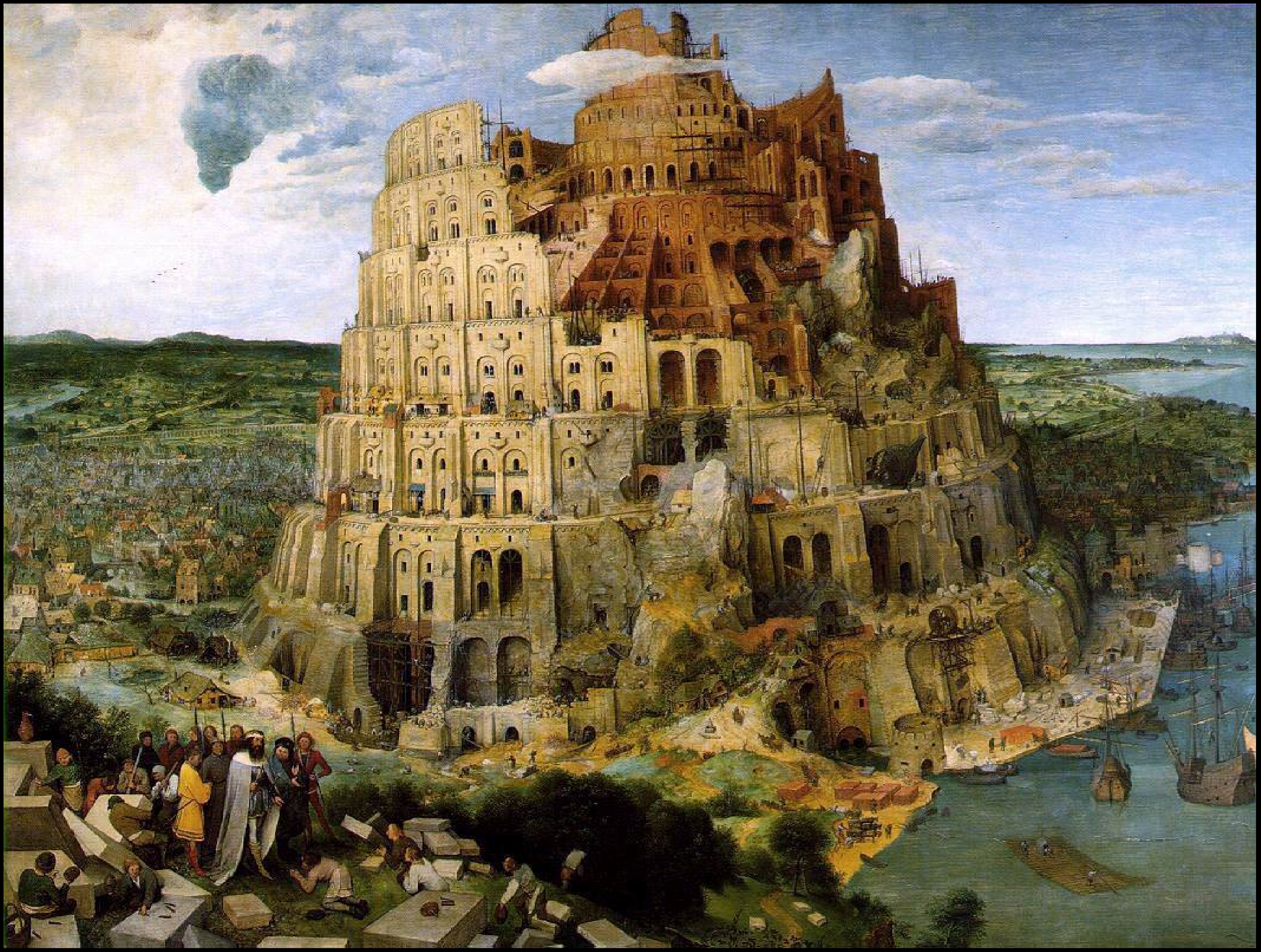
Neural Machine Translation
Philipp Koehn

September 25, 2017

http://mt-class.org/jhu/assets/nmt-book.pdf

نائب امریکی صدر ڈک چینی کا کہنا ہے کہ میں اسامہ بن لادن کو زندہ یا مردہ دیکھنا چاہتا ہوں۔

American Vice President Dick Cheney has said that he wants to see Osama bin Laden dead or alive.



The Tower of Babel

Pieter Brueghel the Elder (1563)

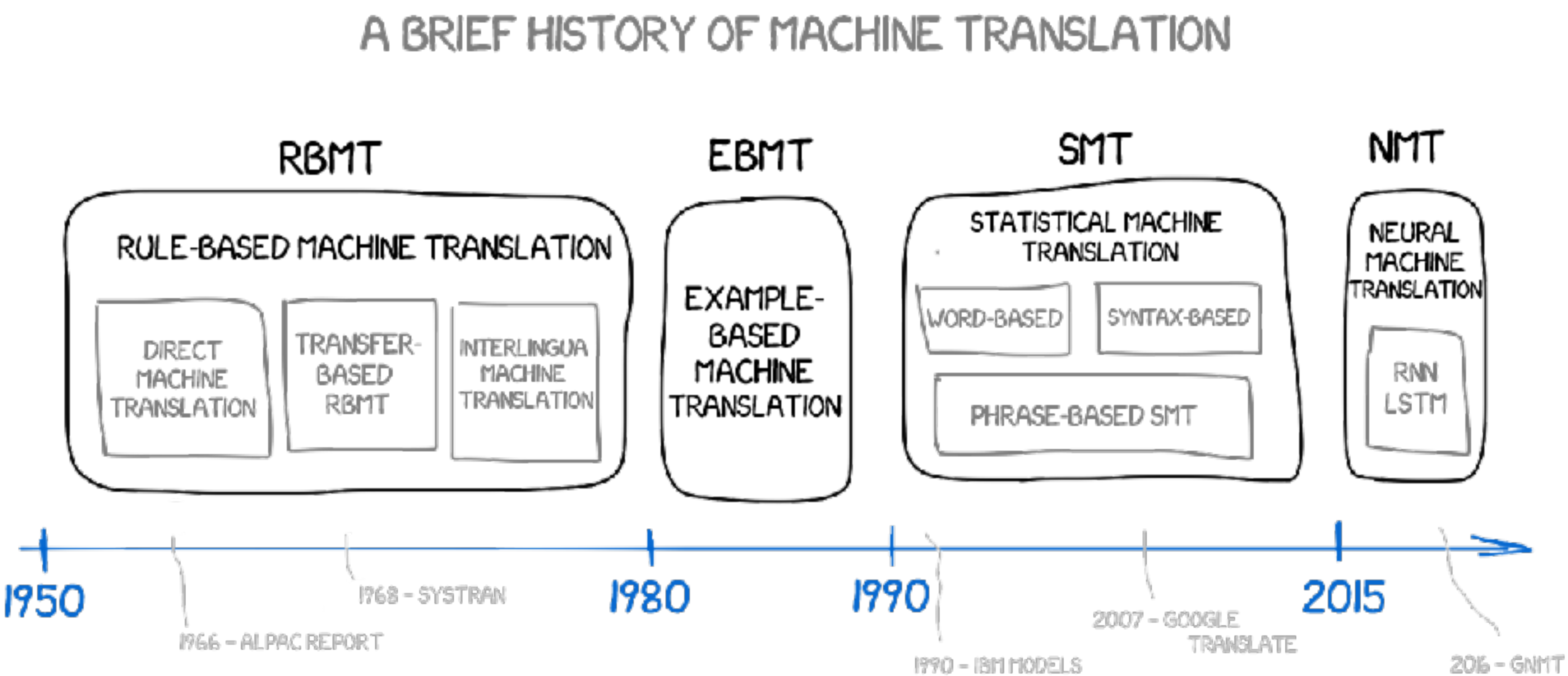
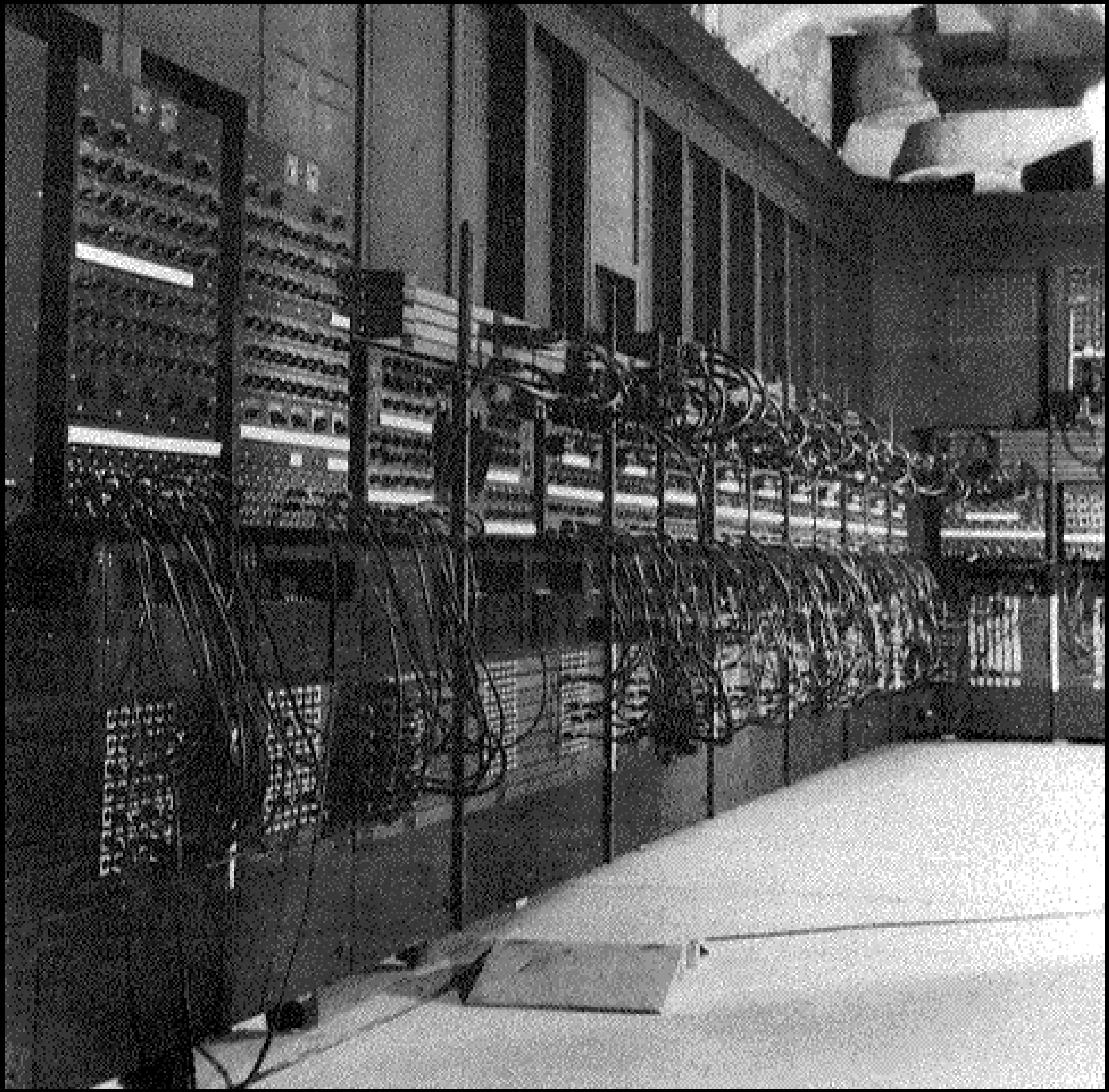


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov



ENIAC (1946)

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 (1949)

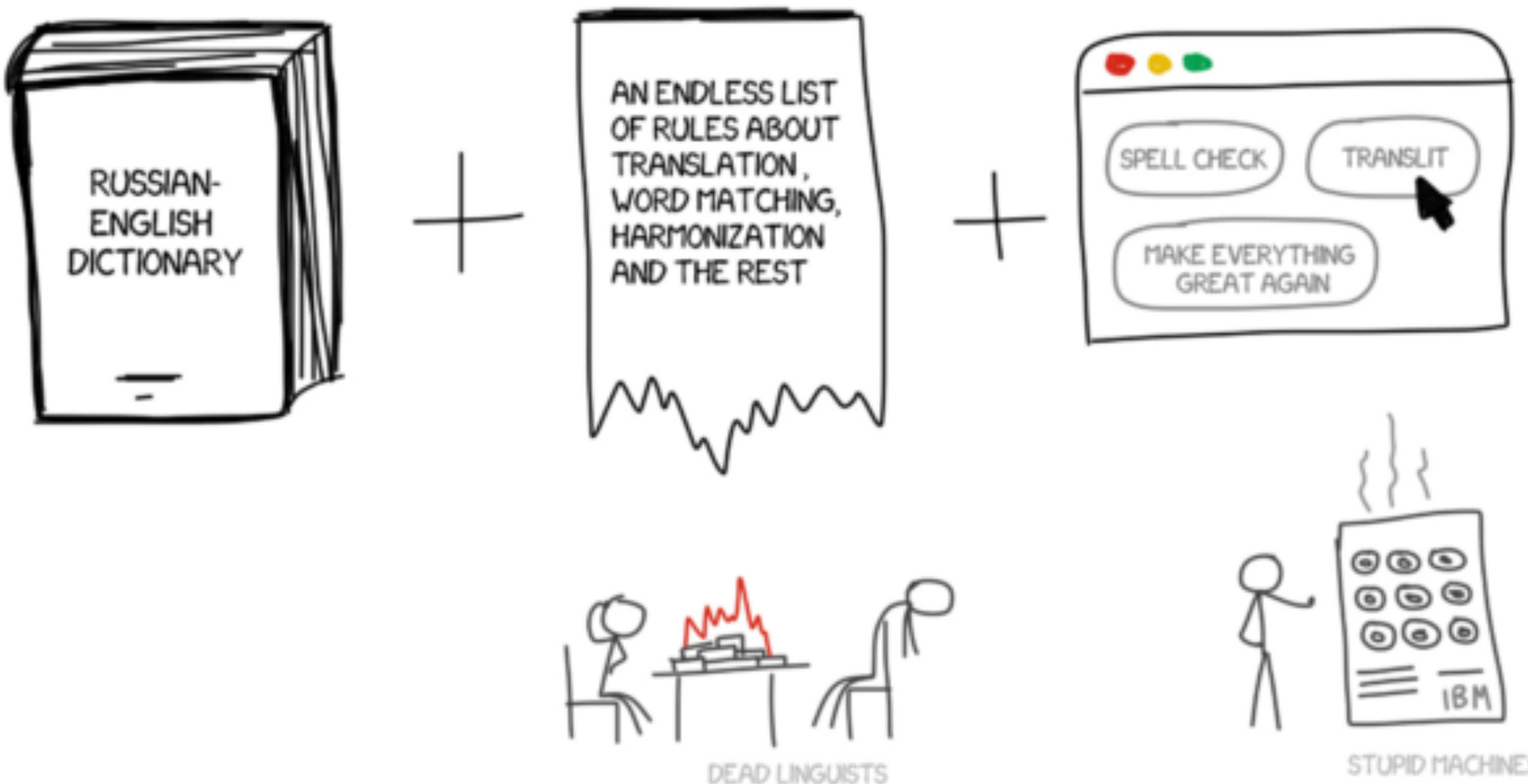


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

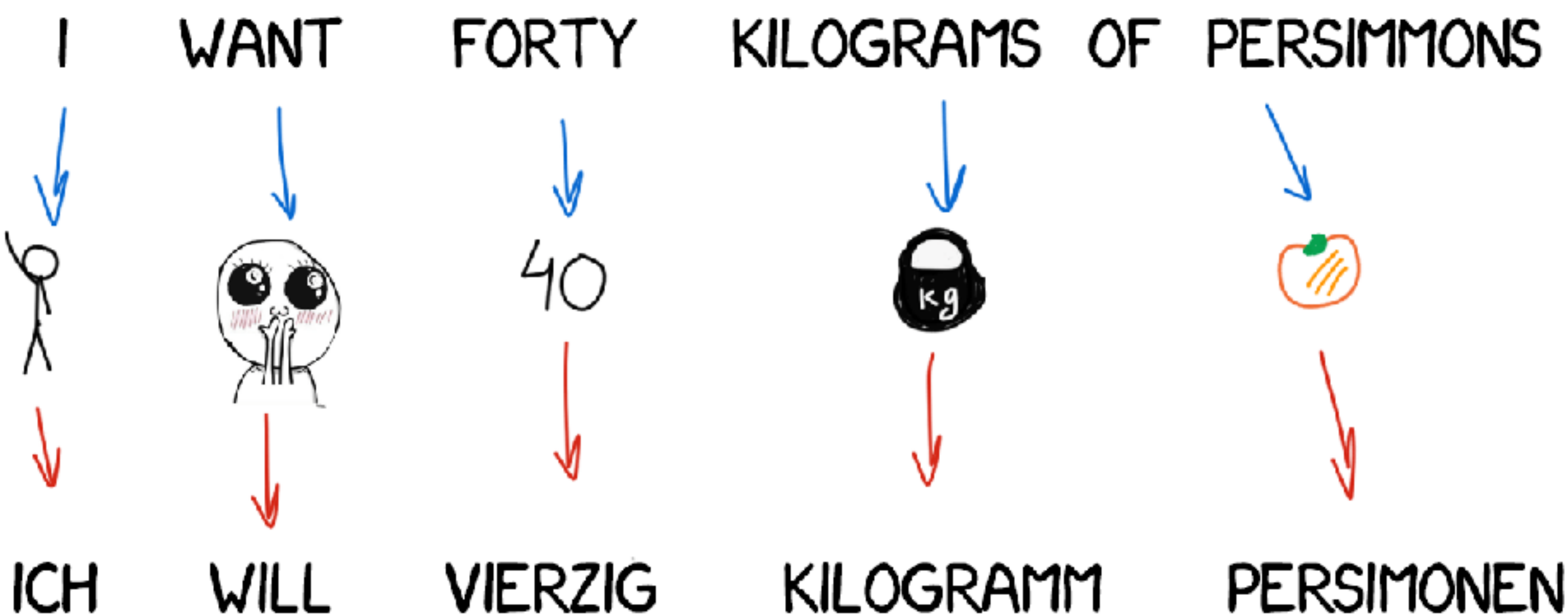


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

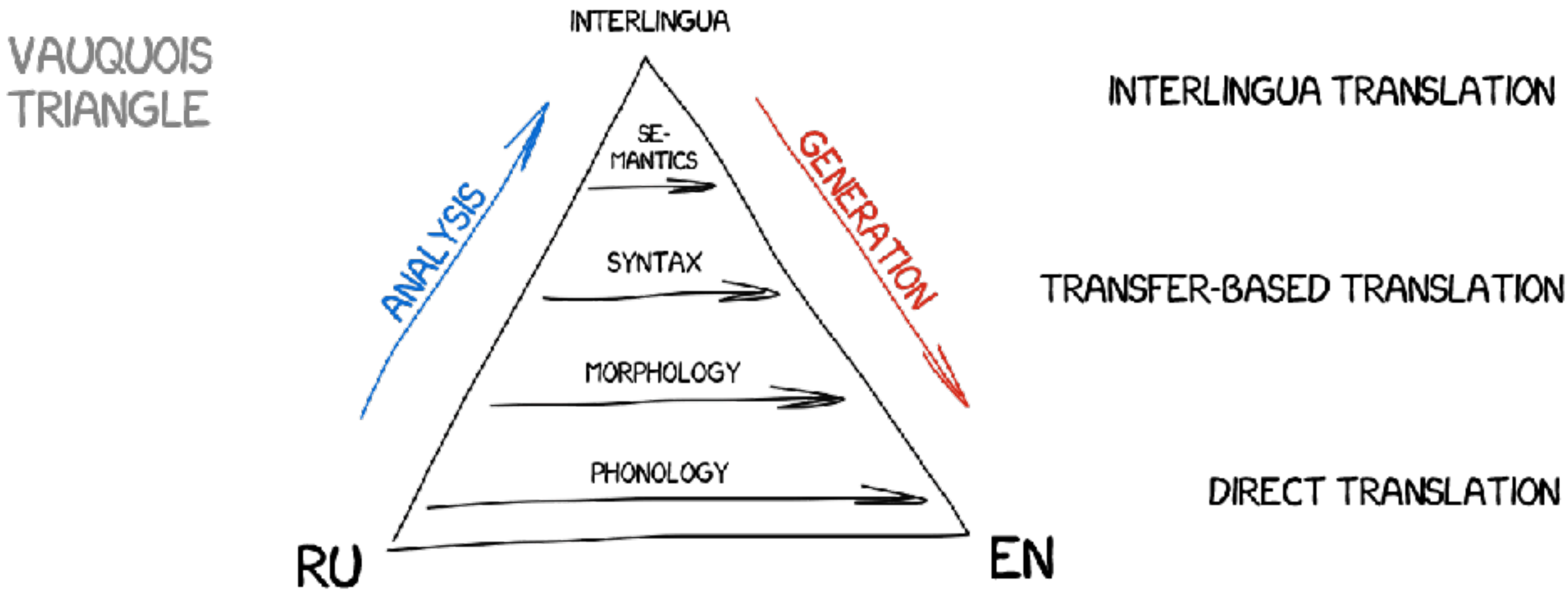
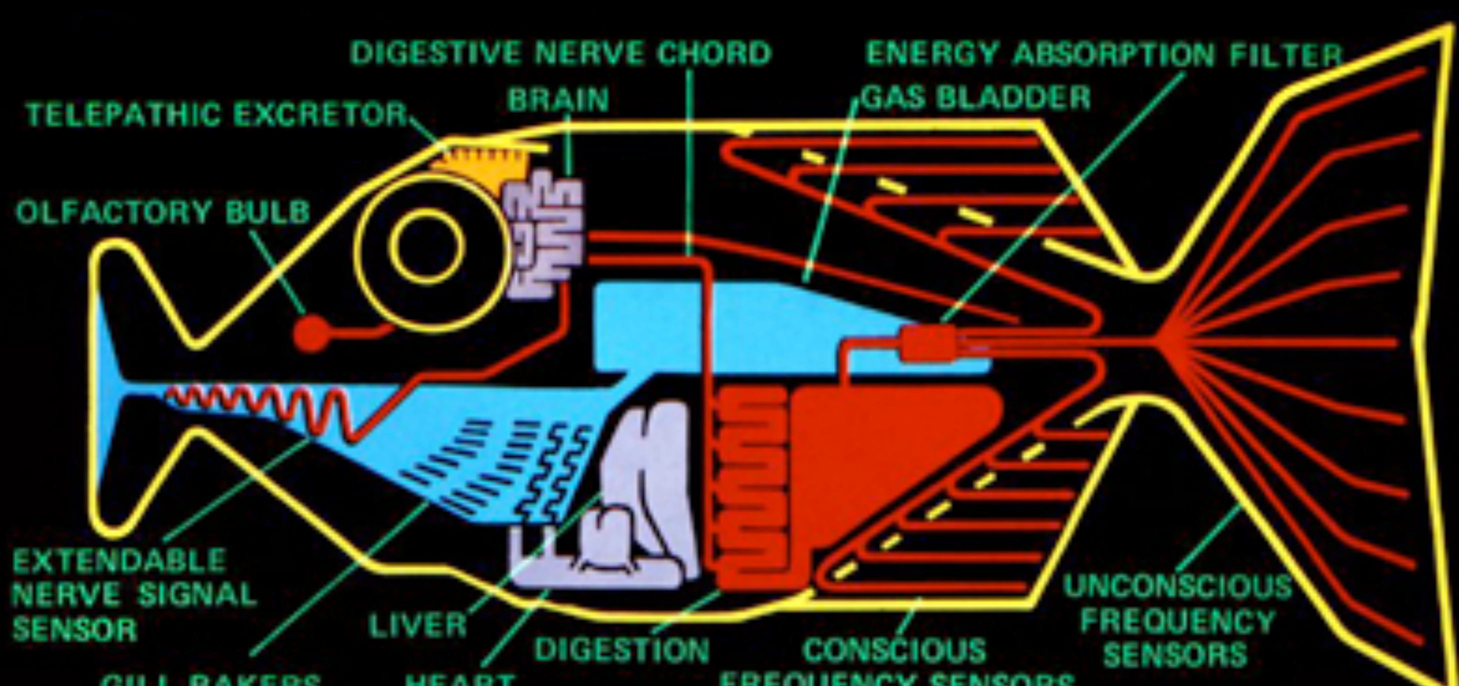


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

Star Trek



Hitchhiker's Guide to the Galaxy

Example-based Machine Translation (EBMT)

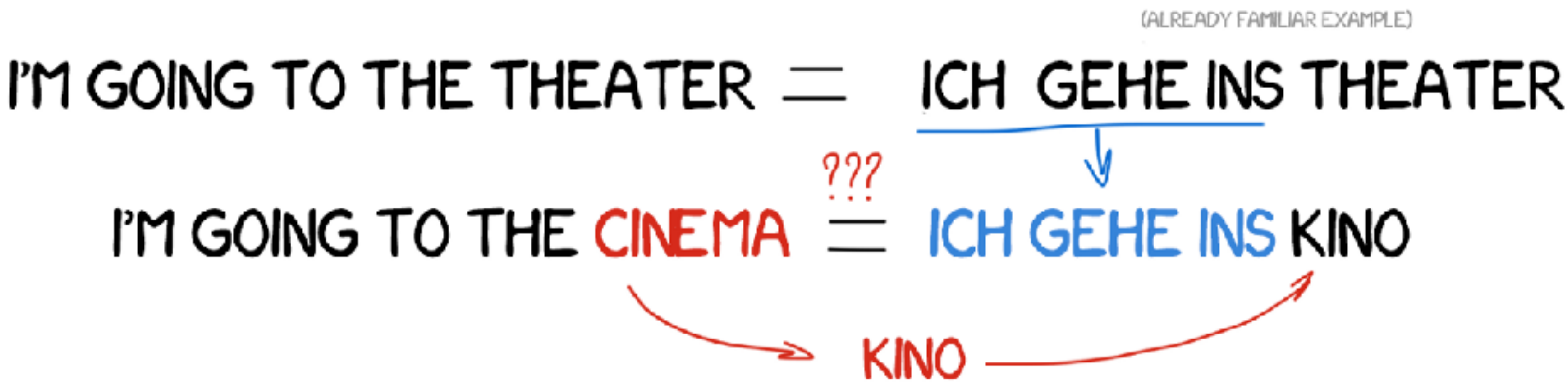


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

Statistical Machine Translation (SMT)

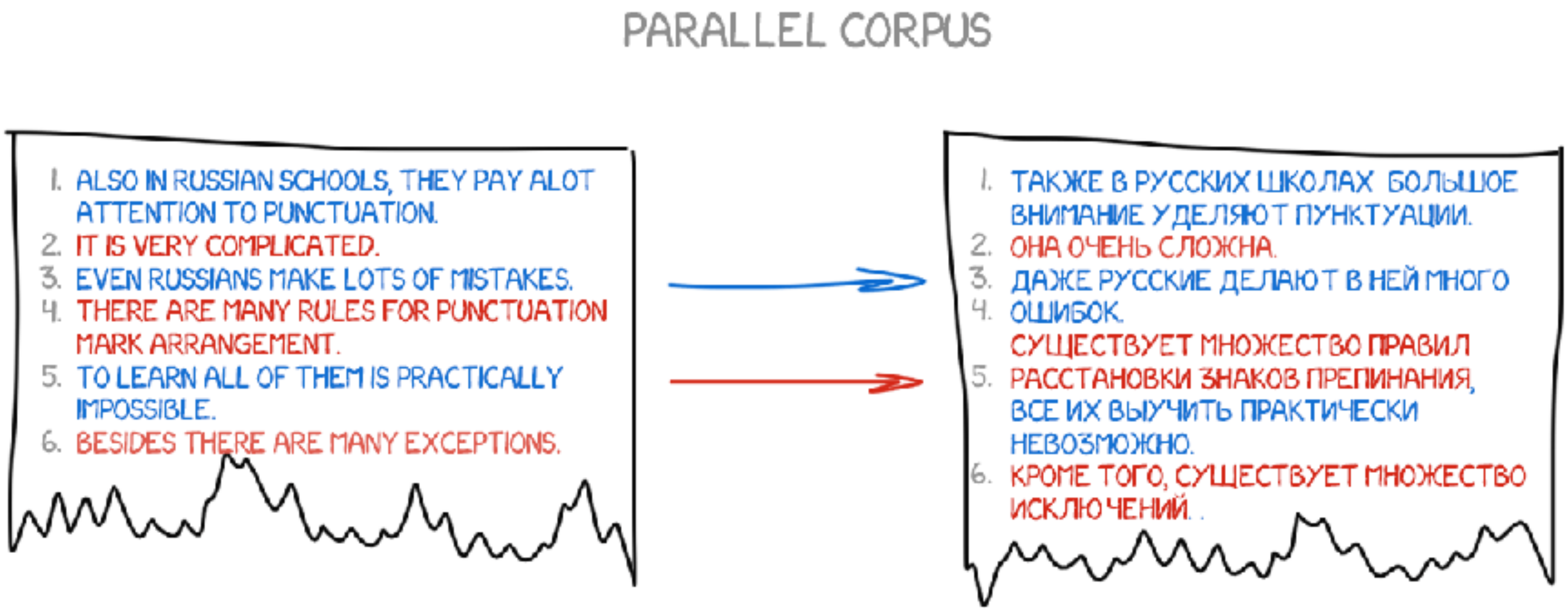


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

Statistical Machine Translation (SMT)

Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

Statistical Machine Translation Live

4/28/2006
Franz Och

Because we want to provide everyone with access to all the world's information, including information written in every language, one of the exciting projects at Google Research is machine translation... Now you can see the results for yourself. We recently launched an online version of our system for Arabic-English and English-Arabic. Try it out!

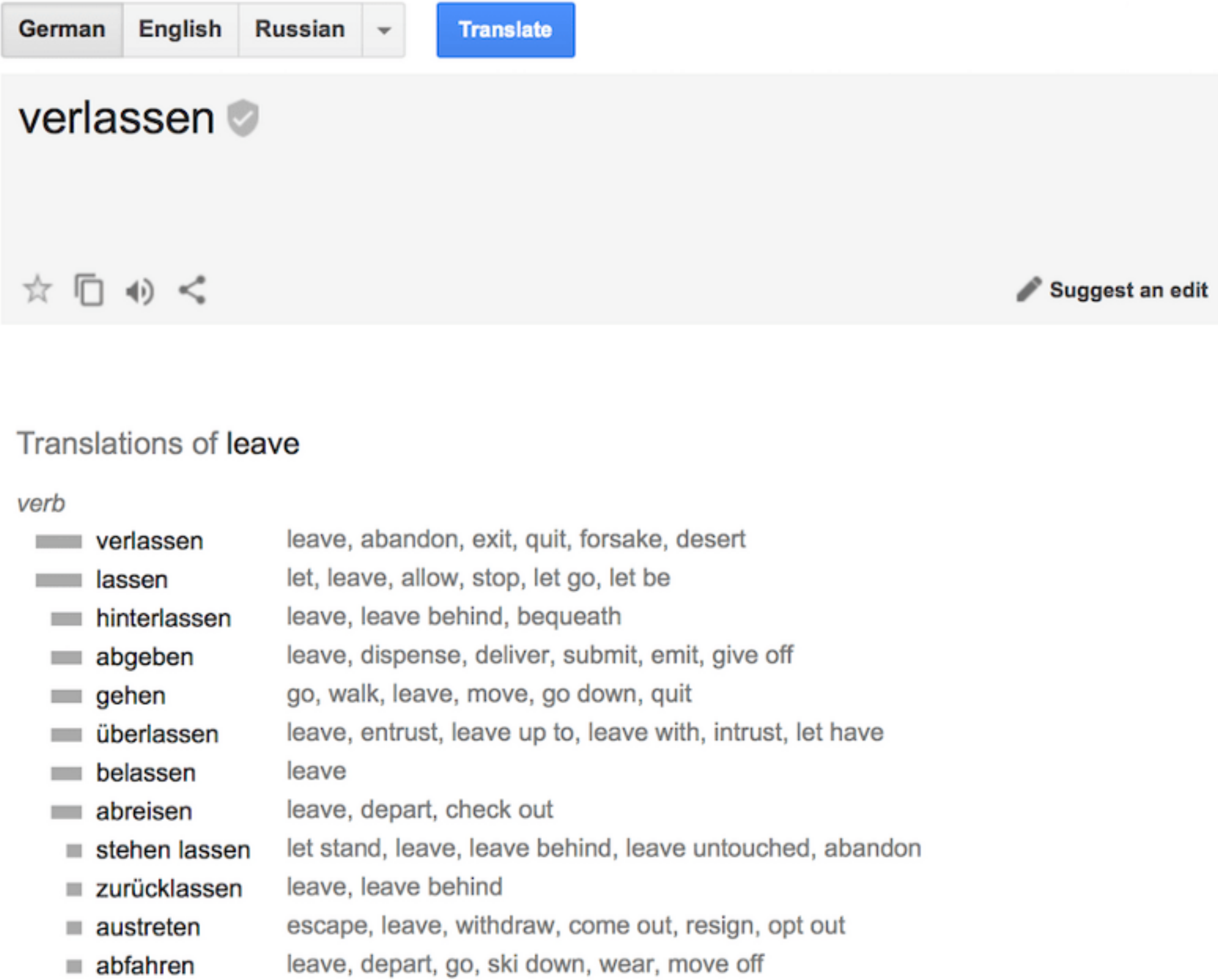


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

Neural Machine Translation

EMNLP 2014
Kyunghyn Cho et al

The source text is encoded by one neural network, and then another neural network decodes it back to the text, but, in another language. The decoder only knows its language. Both have no idea about the each other, and each of them knows only its own language. Interlingua is back.

Neural Machine Translation (NMT)

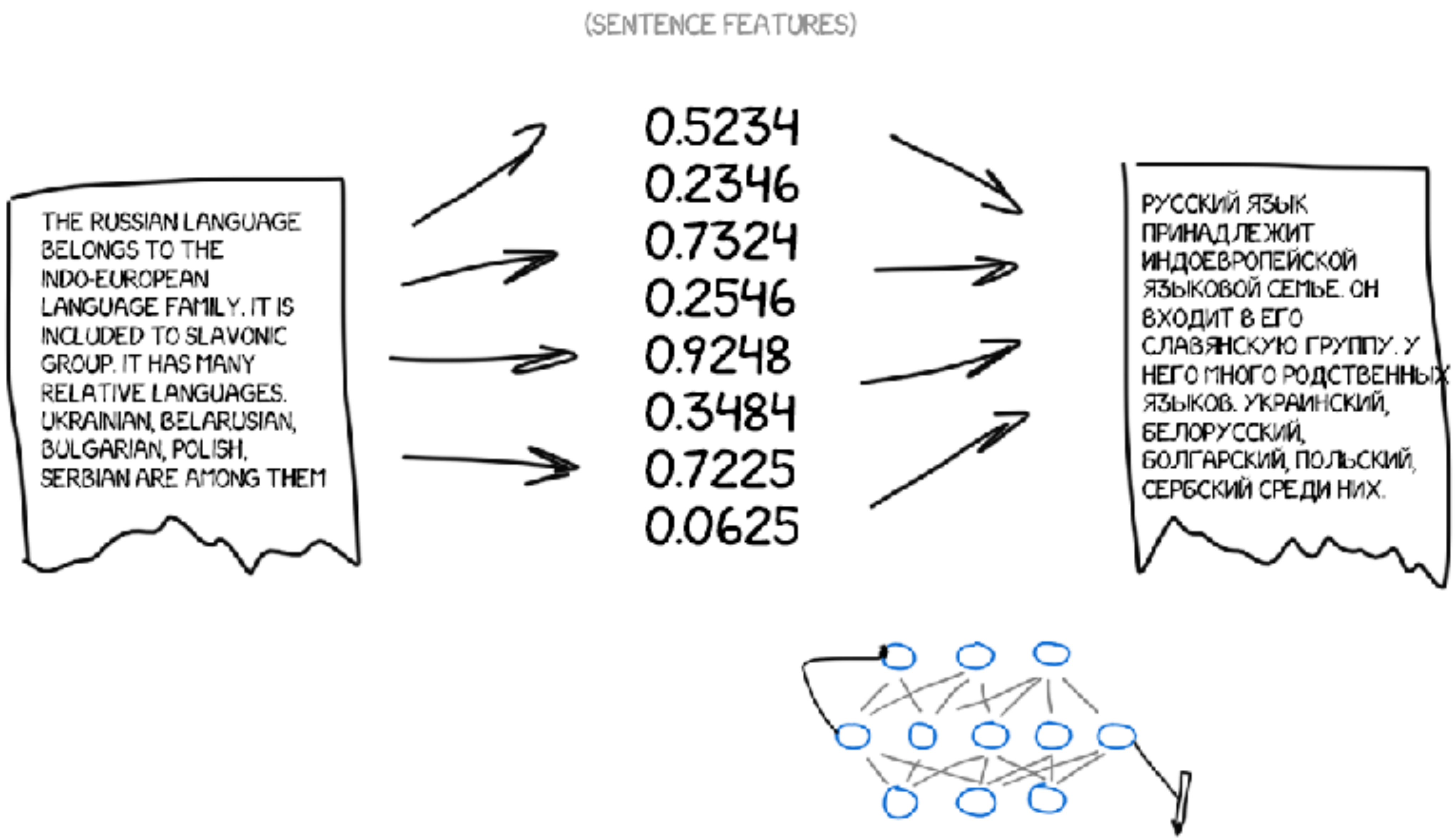


Figure from "A history of machine translation from the Cold War to deep learning" by Ilya Pestov

Poetry by Google Neural Machine Translation

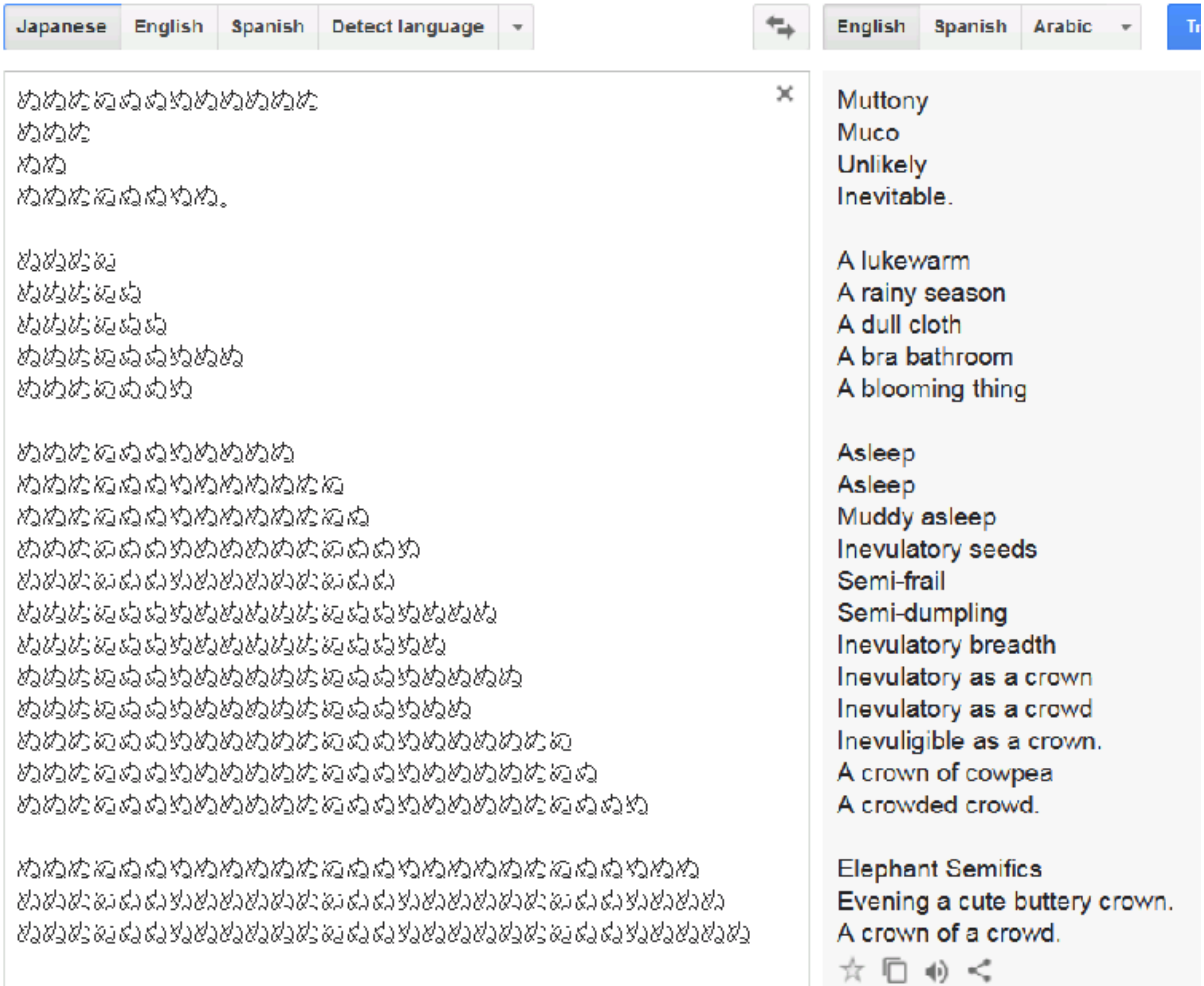


Figure from "Elephant Semifics" by Mark Liberman on the Language Log

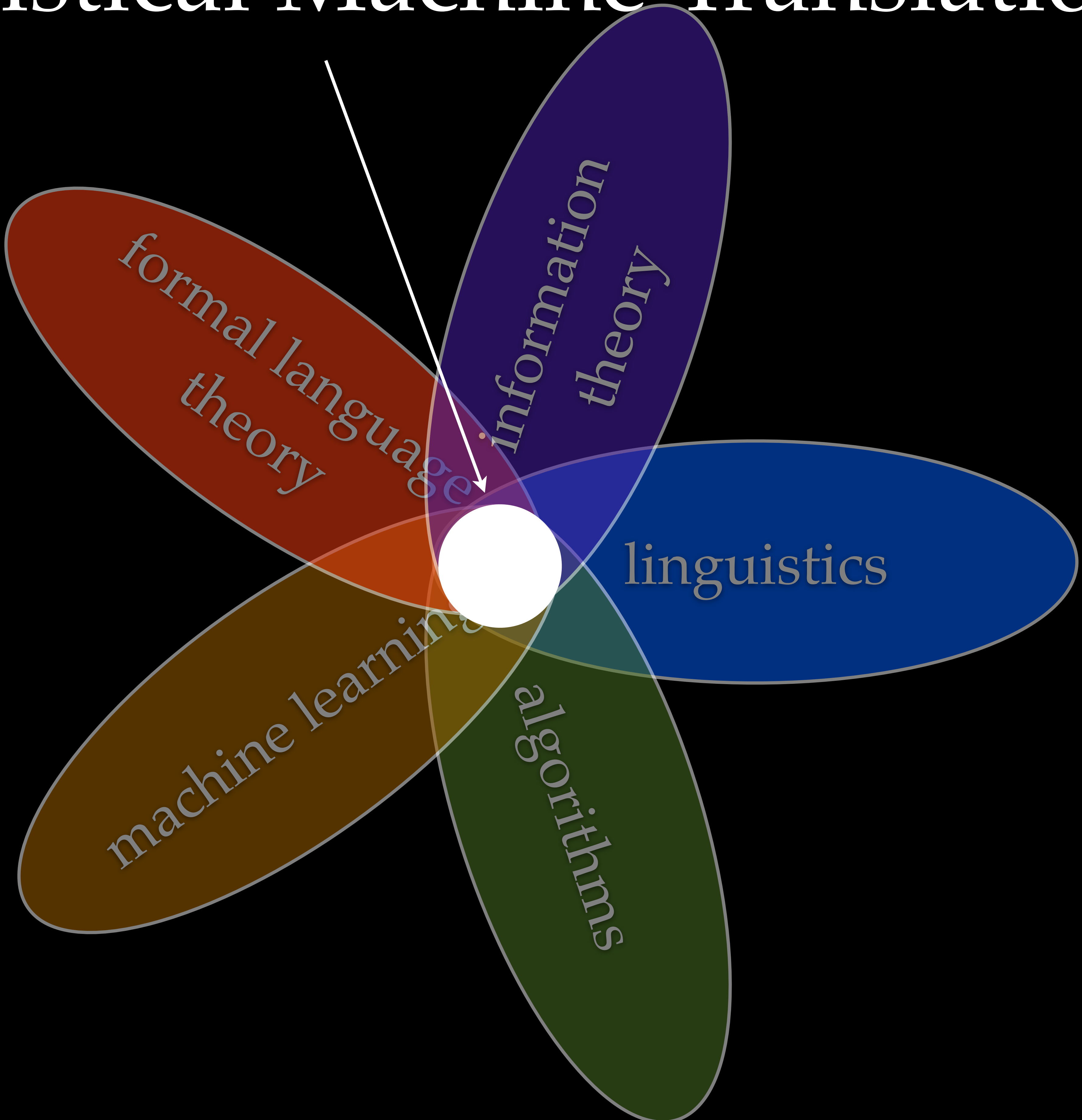

www.un.org
http://www.un.org/english/
中文简体
NBS
www.stats.gov.cn
News and Coming
What's New
Statistical Data
Monthly
Yearly
Census
Others
Related links
Chinese Version


联合国主页
http://www.un.org/chinese/
中文繁体
ENGLISH
http://www.stats.gov.cn/
中华人民共和国国家
National Bureau of Statistics

Statistical Machine Translation

Develop a statistical $\boxed{\textit{model}}$ of translation that can be $\boxed{\textit{learned}}$ from $\boxed{\textit{data}}$ and used to $\boxed{\textit{predict}}$ the correct English translation of new Chinese sentences.

Statistical Machine Translation



In-class exercise

Word aligner

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

Word aligner

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

Lexical Translation

- How do we translate a word? Look it up in the dictionary

  Haus : *house*, *home*, *shell*, *household*

- Multiple translations
  - Different word senses, different registers, different inflections (?)
  - *house*, *home* are common
  - *shell* is specialized (the Haus of a snail is a shell)

How common is each?

Translation	Count
house	5000
home	2000
shell	100
household	80

MLE

$$\hat{p}_{\mathrm{MLE}}(e \mid \texttt{Haus}) = \begin{cases} 0.696 & \text{if } e = \texttt{house} \\ 0.279 & \text{if } e = \texttt{home} \\ 0.014 & \text{if } e = \texttt{shell} \\ 0.011 & \text{if } e = \texttt{household} \\ 0 & \text{otherwise} \end{cases}$$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences

$$\mathbf{e} = \langle e_1, e_2, \ldots, e_m \rangle \qquad \mathbf{f} = \langle f_1, f_2, \ldots, f_n \rangle$$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where **e** and **f** are complete English and Foreign sentences
- Lexical translation makes the following ***assumptions***:
  - Each word e_i in **e** is generated from exactly one word in **f**
  - Thus, we have an *alignment* a_i that indicates which word e_i "came from", specifically it came from f_{a_i} .
  - Given the alignments **a**, translation decisions are conditionally independent of each other and depend *only* on the aligned source word f_{a_i} .

Lexical Translation

- Putting our assumptions together, we have:

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in [0,n]^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

Alignment $\times$ Translation | Alignment

Lexical Translation

$$p(e_i \mid f_{a_i})$$

$p(\texttt{house} \mid \texttt{Haus})$ $p(\texttt{shell} \mid \texttt{Haus})$

Remember bigram models...

Lexical Translation

- Putting our assumptions together, we have:

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in [0,n]^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

Alignment $\times$ Translation | Alignment

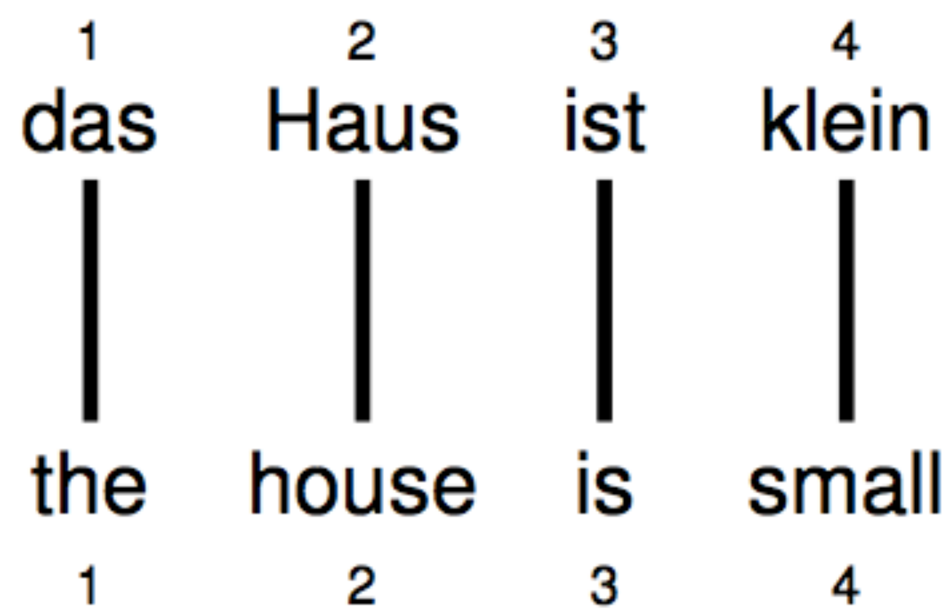
Alignment

$$p(\mathbf{a} \mid \mathbf{f}, m)$$

Most of the action for the first 10 years of SMT was here. Words weren't the problem, word *order* was hard.

Alignment

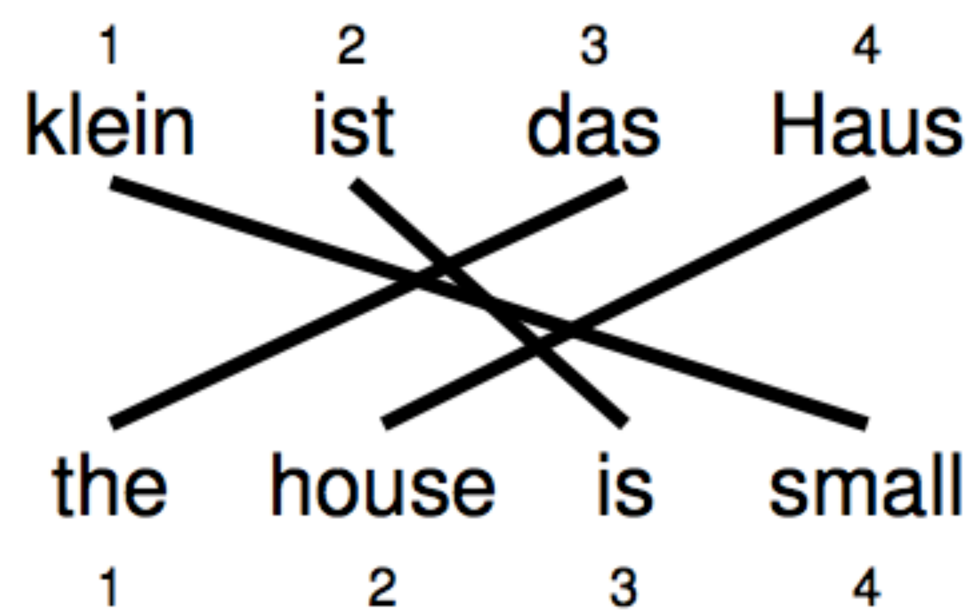
- Alignments can be visualized in by drawing links between two sentences, and they are represented as vectors of positions:



$$\mathbf{a} = (1, 2, 3, 4)^\top$$

Reordering

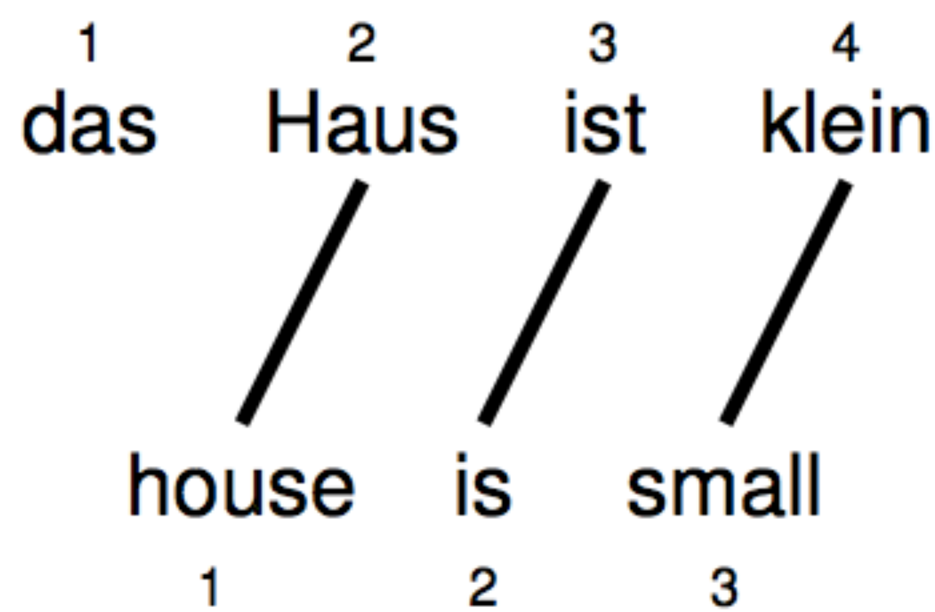
- Words may be reordered during translation.



$$\mathbf{a} = (3, 4, 2, 1)^\top$$

Word Dropping

- A source word may not be translated at all

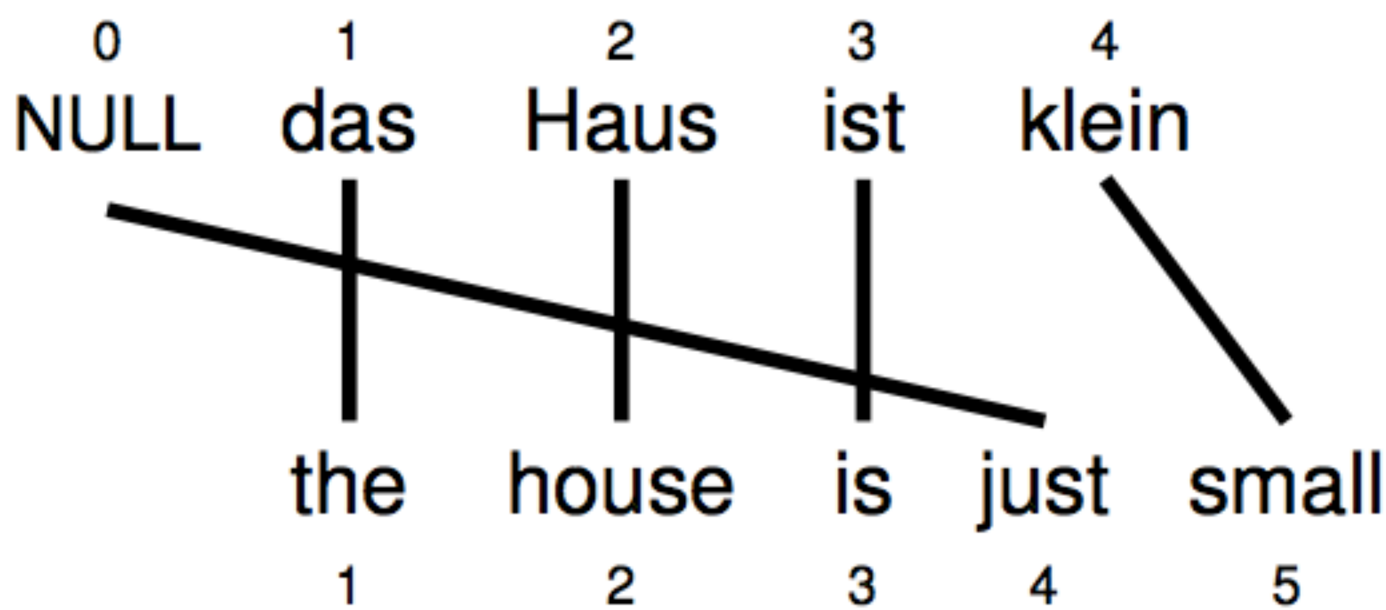


$$\mathbf{a} = (2, 3, 4)^\top$$

Word Insertion

- Words may be inserted during translation
  English just does not have an equivalent

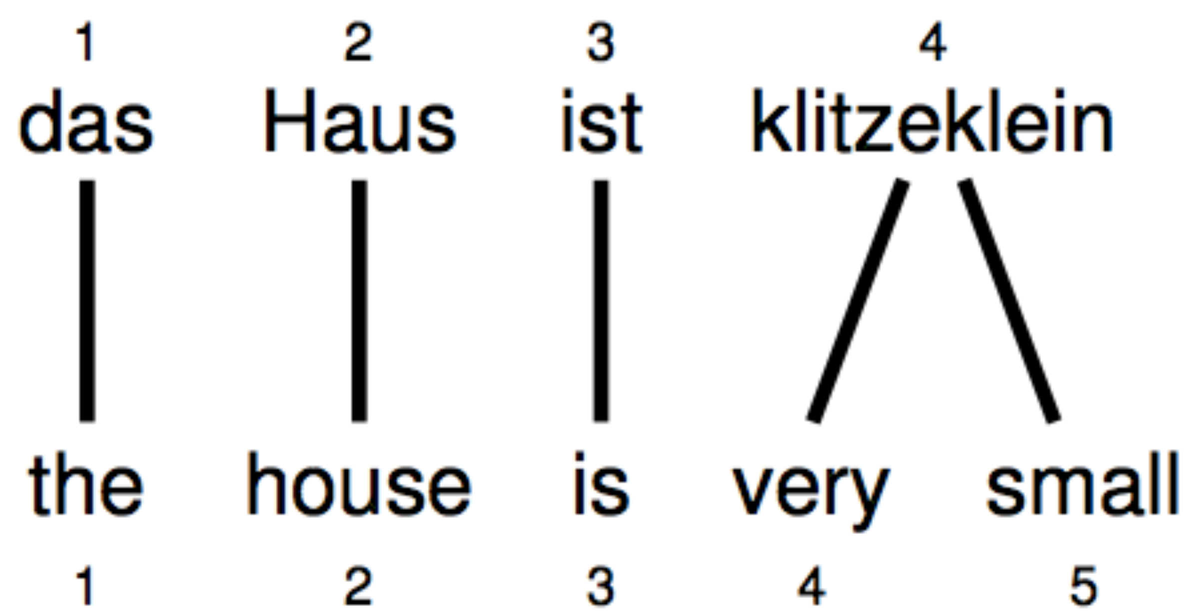
  But it must be explained - we typically assume every source sentence contains a NULL token



$$\mathbf{a} = (1, 2, 3, 0, 4)^\top$$

One-to-many Translation

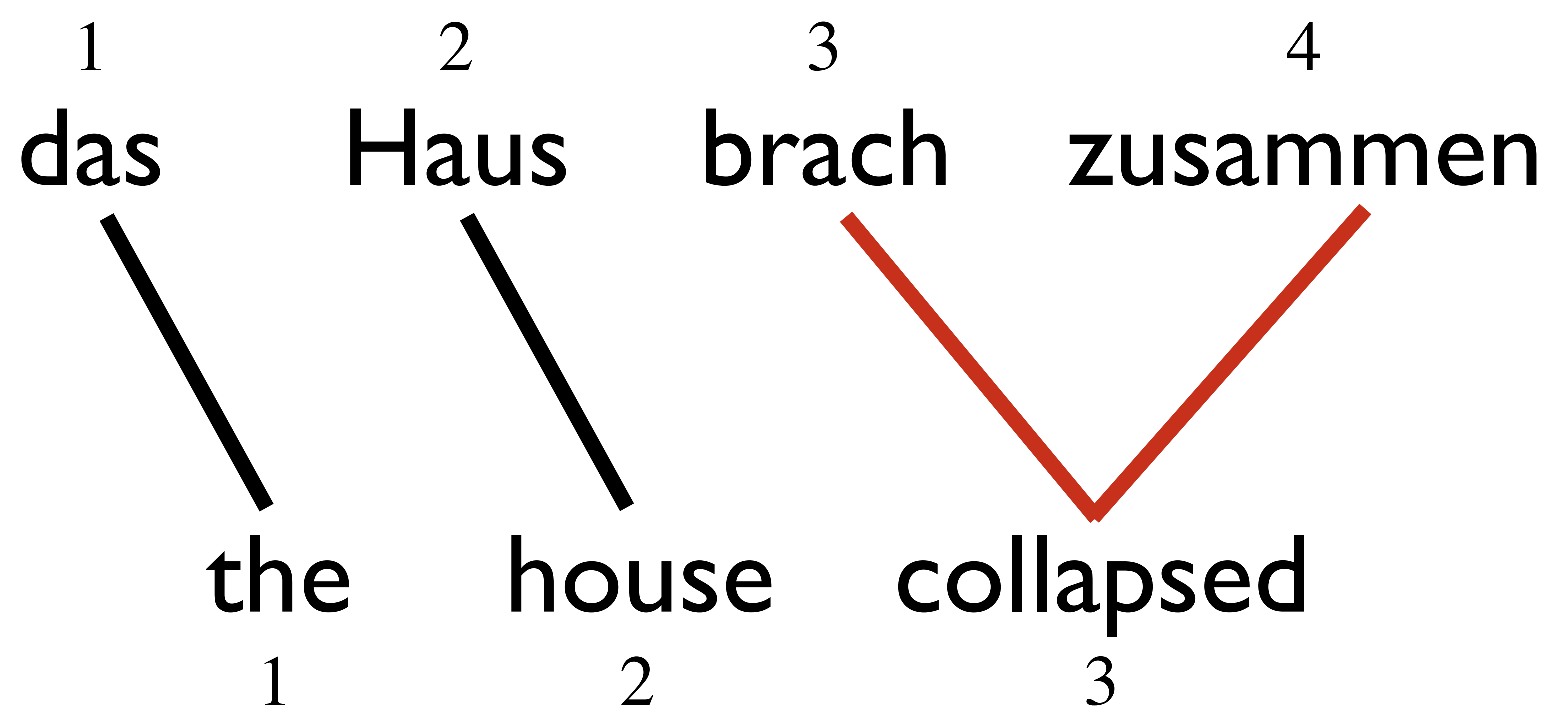
- A source word may translate into **more than one** target word



$$\mathbf{a} = (1, 2, 3, 4, 4)^\top$$

Many-to-one Translation

- **More than one source word** may **not** translate as a unit in lexical translation



$$\mathbf{a} = ???$$

$$\mathbf{a} = (1, 2, (3, 4)^\top)^\top \quad ?$$

IBM Model 1

- Simplest possible lexical translation model
- Additional assumptions
  - The m alignment decisions are independent
  - The alignment distribution for each a_i is uniform over all source words and NULL

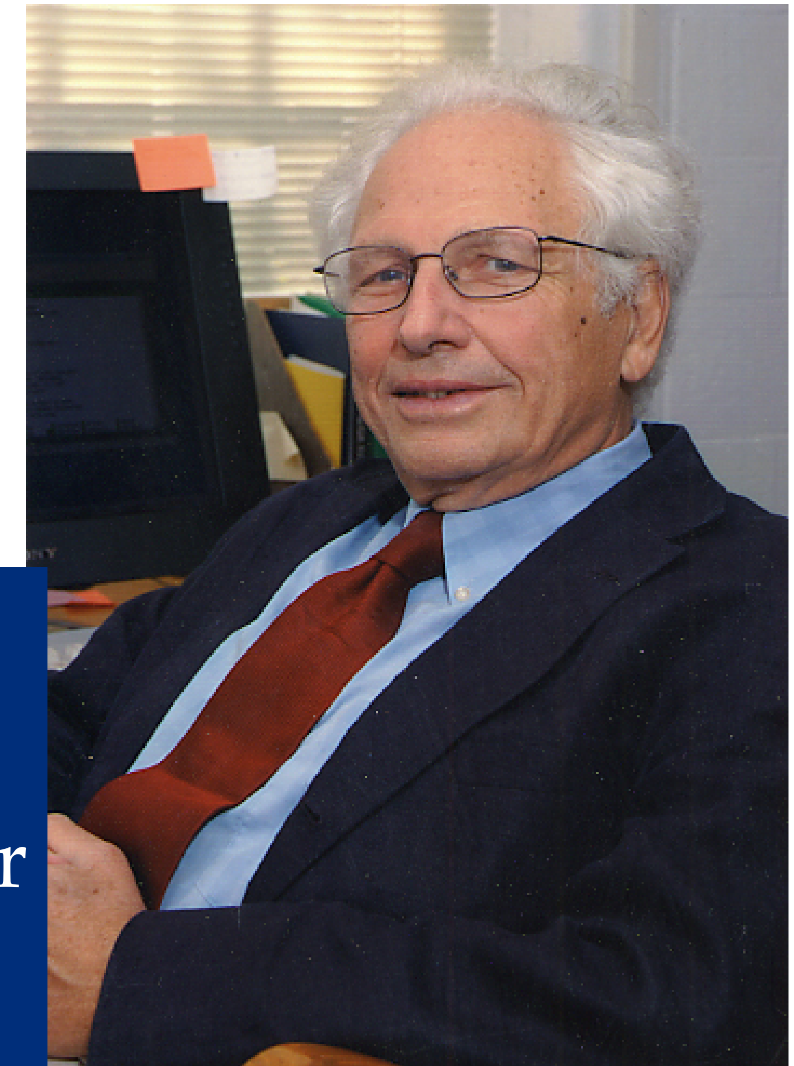
$$
\begin{aligned}
&\text{for each } i \in [1, 2, \ldots, m] \\
&\qquad a_i \sim \mathrm{Uniform}(0, 1, 2, \ldots, n) \\
&\qquad e_i \sim \mathrm{Categorical}(\boldsymbol{\theta}_{f_{a_i}})
\end{aligned}
$$

Historical Note

IBM Models

Renaissance

“The validity of a statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950 (cf. Hutchins, MT – Past, Present, Future, Ellis Horwood, 1986, p. 30ff and references therein). The crude force of computers is not science. The paper is simply beyond the scope of COLING.”



Fred Jelinek (1932-2010)

The Center For Language and Speech Processing at the Johns Hopkins University

1. Introduction

The growing availability of bilingual, machine-readable texts has stimulated interest

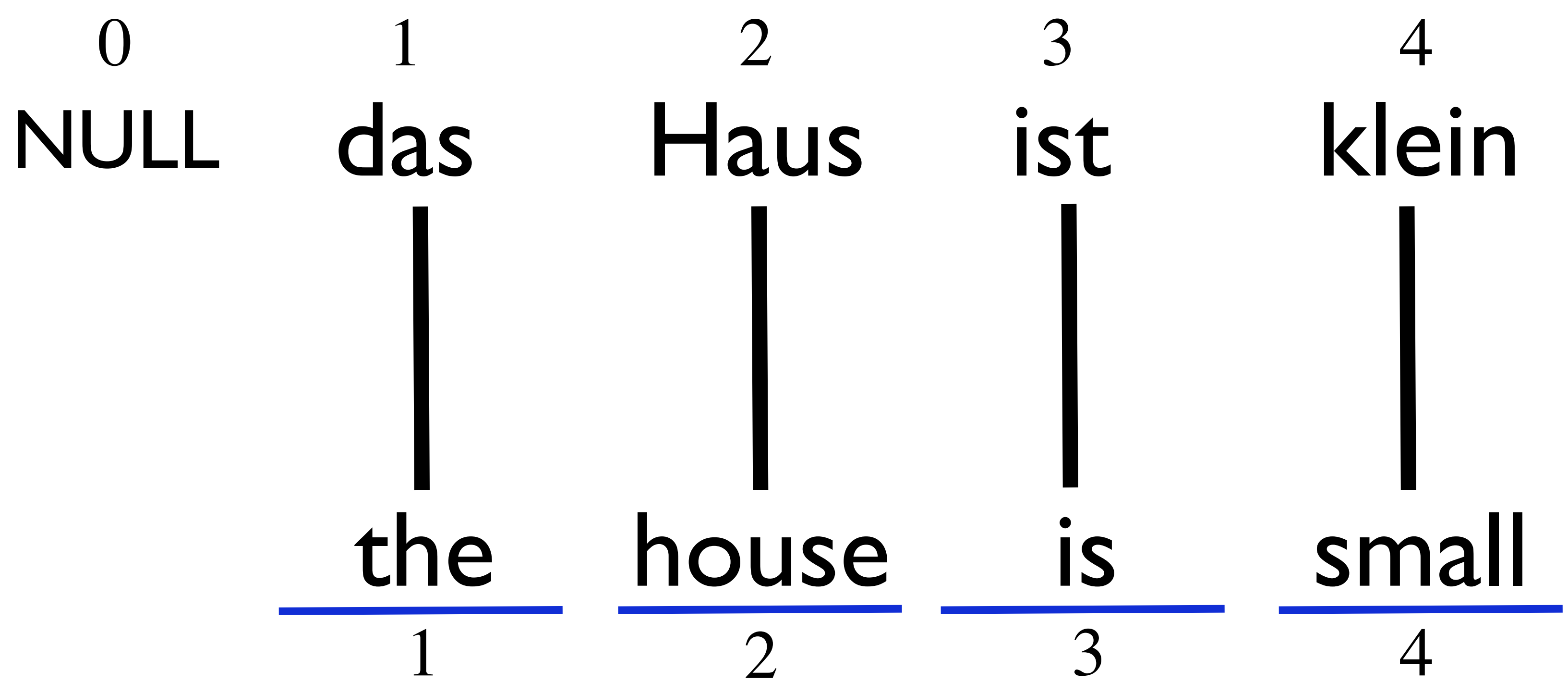
Example

0	1	2	3	4
NULL	das	Haus	ist	klein

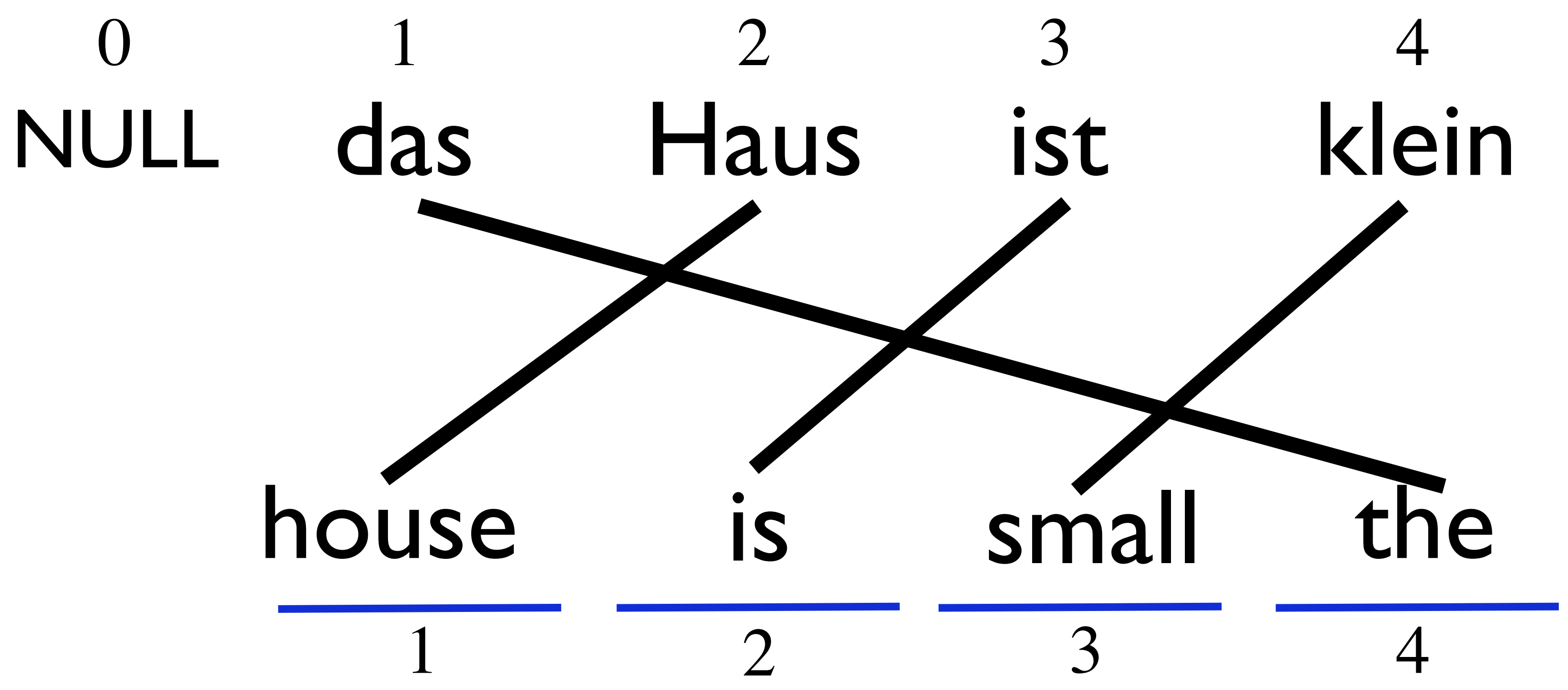
___	___	___	___
1	2	3	4

Start with a foreign sentence and a target length.

Example



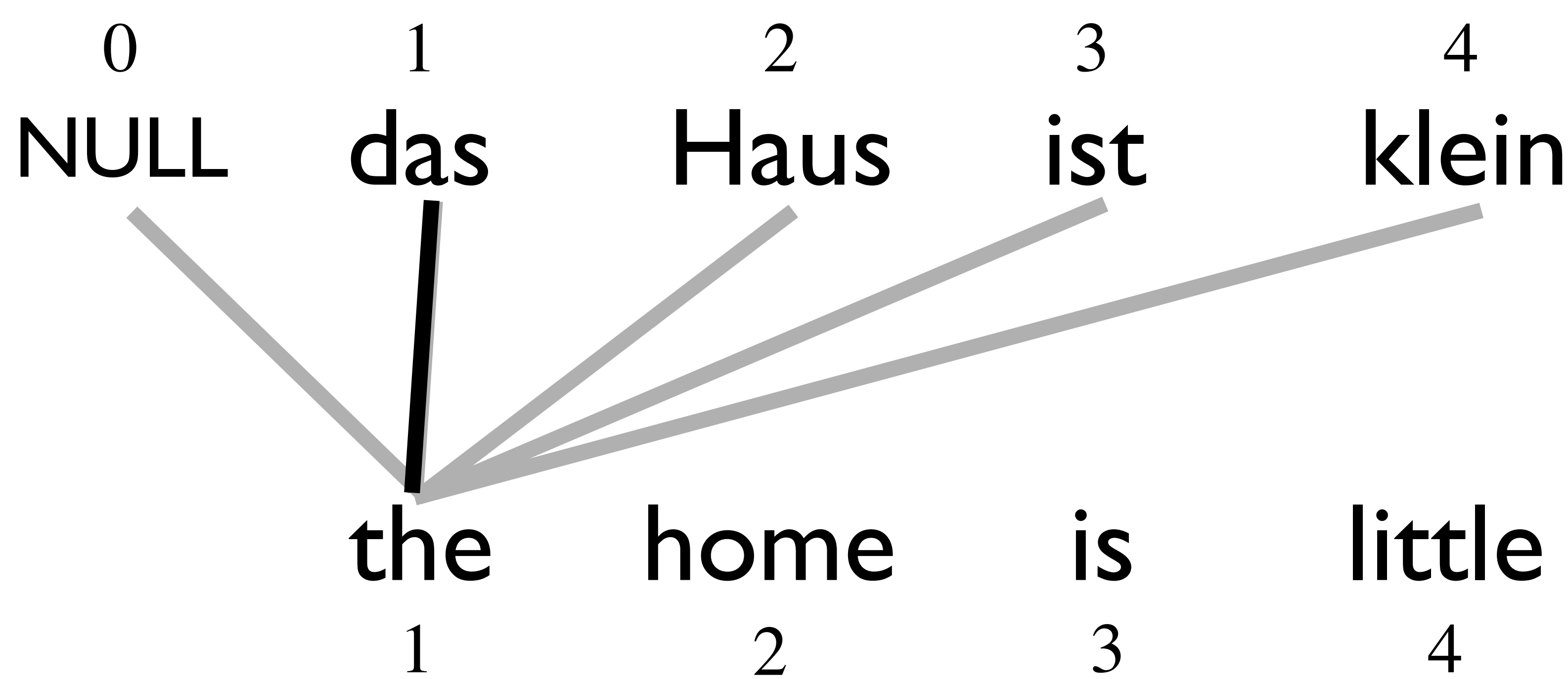
Example



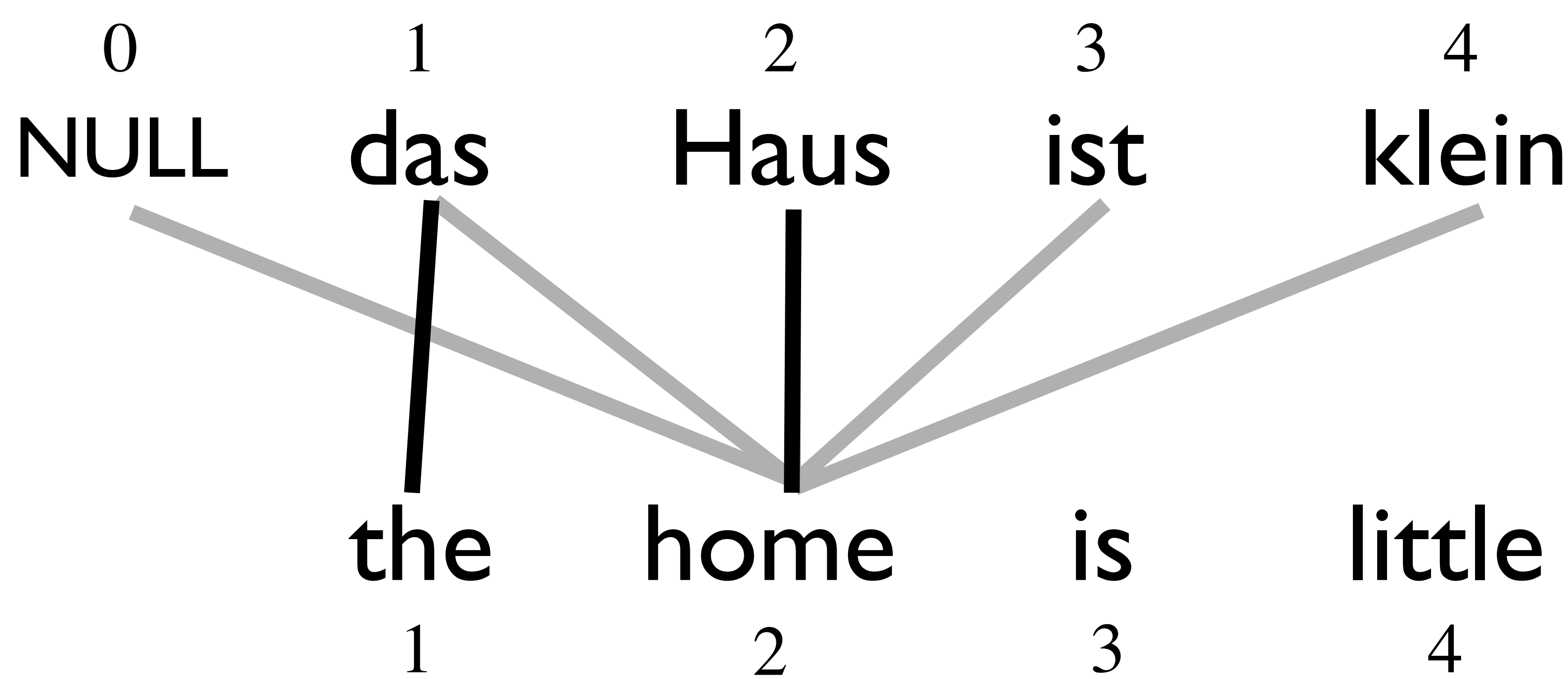
Finding the Viterbi Alignment

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in [0,1,\ldots,n]^m} p(\mathbf{a} \mid \mathbf{e}, \mathbf{f})$$

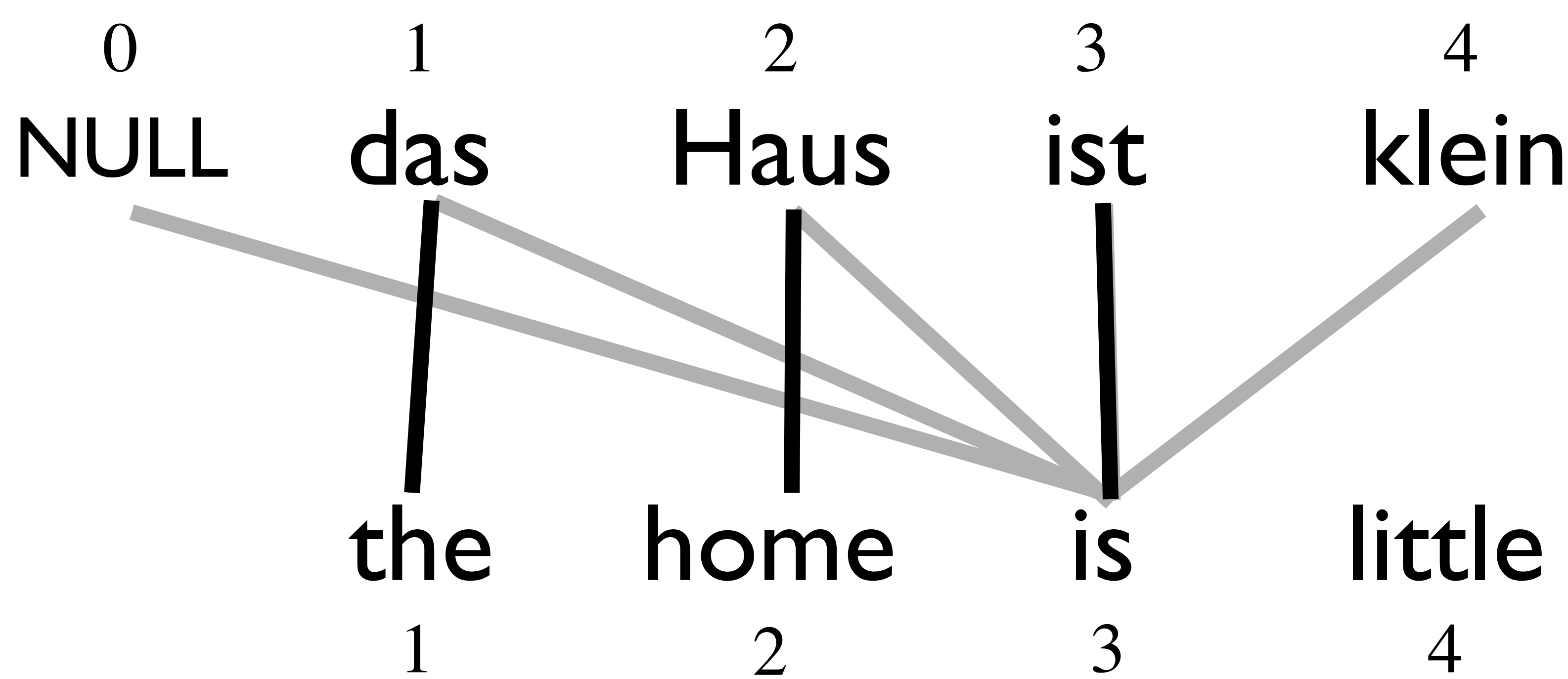
Finding the Viterbi Alignment



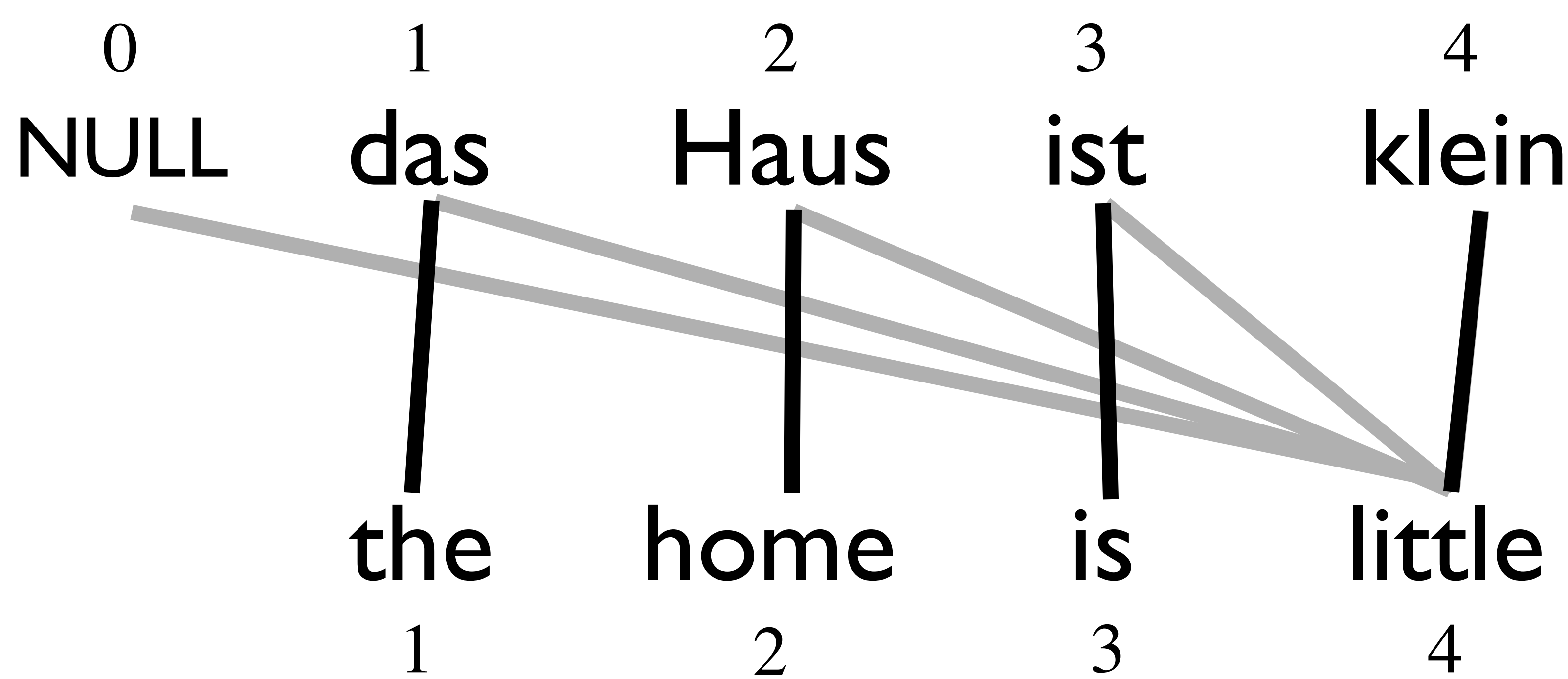
Finding the Viterbi Alignment



Finding the Viterbi Alignment



Finding the Viterbi Alignment



Historical Note #2

The **Viterbi algorithm** is a dynamic programming algorithm for finding the most likely sequence of hidden states – called the **Viterbi path** – that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models.

Andrew Viterbi

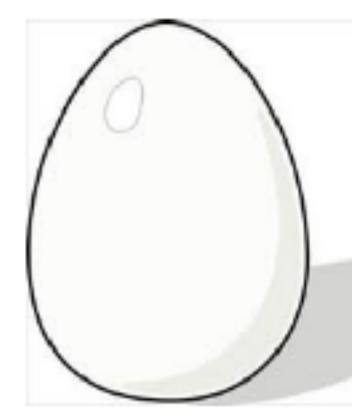
Professor at USC

co-founder of Qualcomm

classmates with Fred Jelinek

Learning Lexical Translation Models

- How do we learn the parameters $p(e \mid f)$
- "Chicken and egg" problem
  - If we had the alignments, we could estimate the parameters (MLE)
  - If we had parameters, we could find the most likely alignments



EM Algorithm

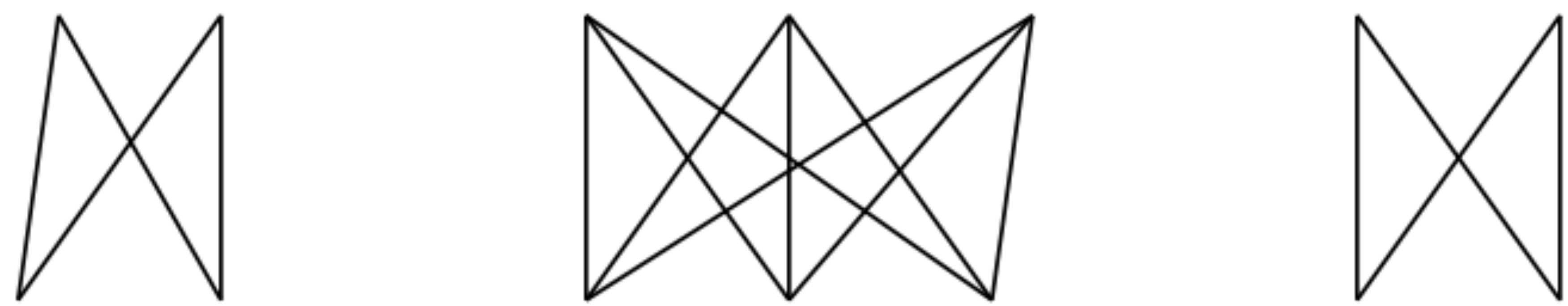
- pick some random (or uniform) parameters
- Repeat until you get bored (~ 5 iterations for lexical translation models)
  - using your current parameters, compute “expected” alignments for every target word token in the training data

    $$p(a_i \mid \mathbf{e}, \mathbf{f})$$

  - keep track of the expected number of times f translates into e throughout the whole corpus
  - keep track of the expected number of times that f is used as the source of any translation
  - use these expected counts as if they were “real” counts in the standard MLE equation

EM for Model 1

... la maison ... la maison blue ... la fleur ...

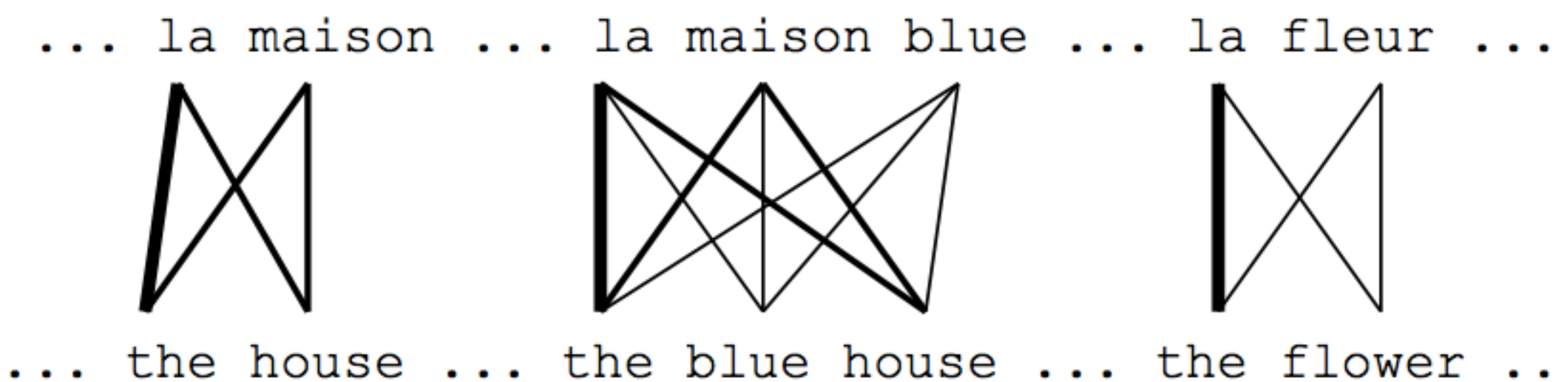


... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely

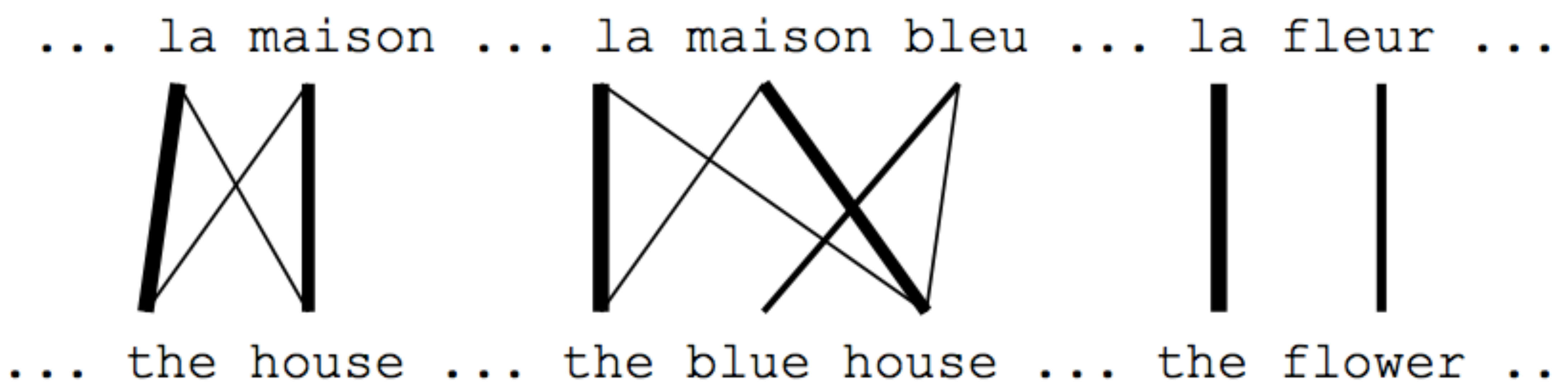
- Model learns that, e.g., la is often aligned with the

EM for Model 1



- After one iteration
- Alignments, e.g., between la and the are more likely

EM for Model 1



- After another iteration
- It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)

EM for Model 1

```
... la maison ... la maison bleu ... la fleur ...
```

```
... the house ... the blue house ... the flower ...
```

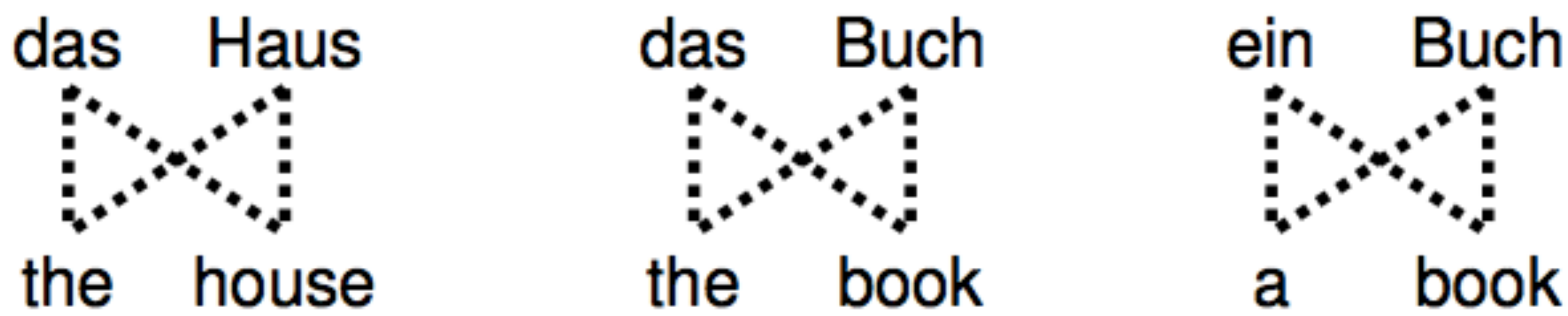
- Convergence
- Inherent hidden structure revealed by EM

EM for Model 1


```
p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...
```

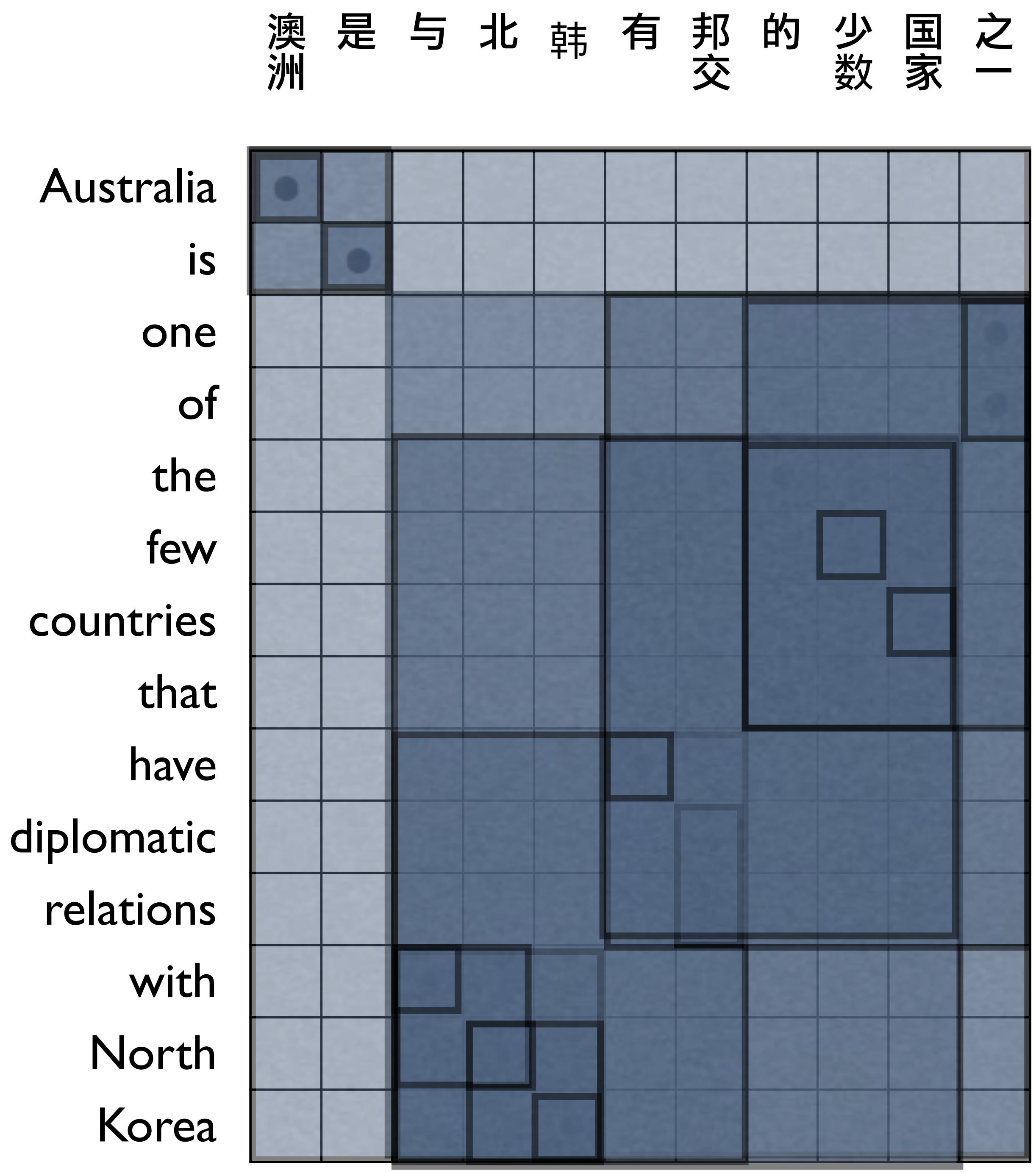
- Parameter estimation from the aligned corpus

Convergence



e	f	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

Phrase Extractor



北 韩

North Korea

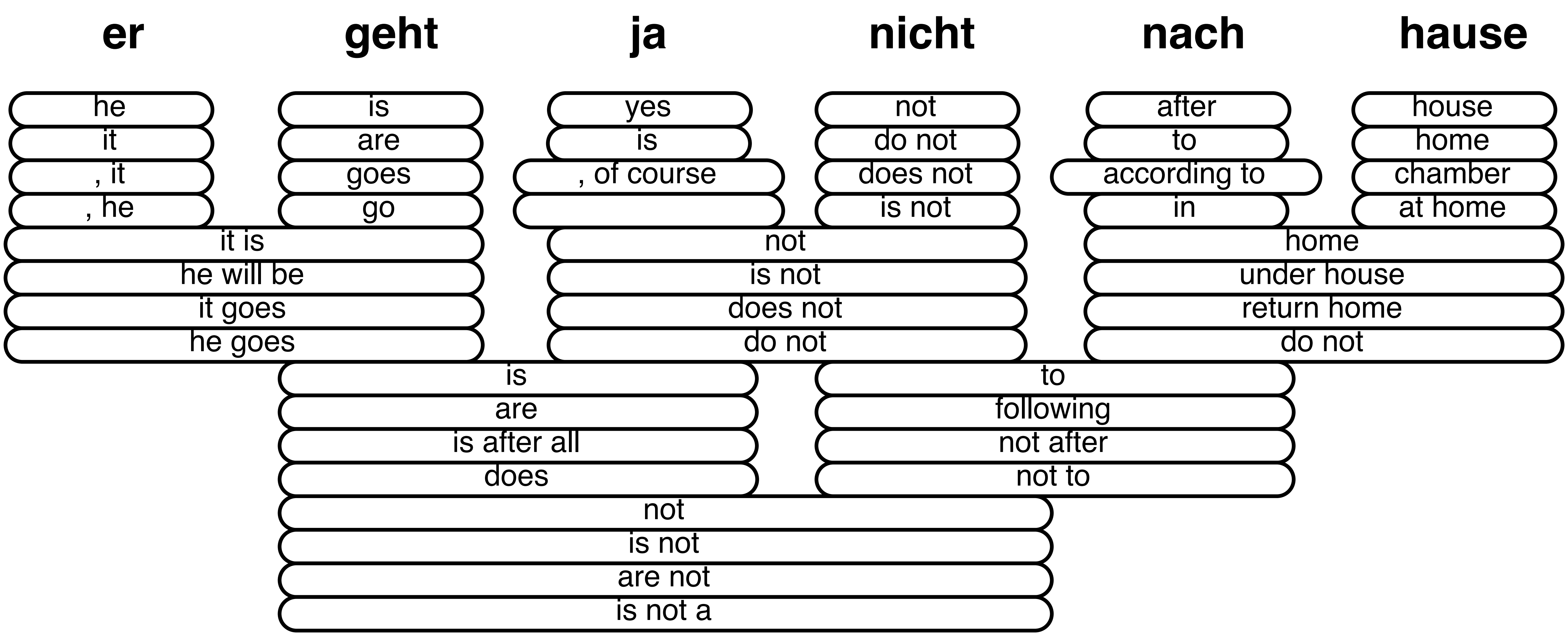
邦交

diplomatic relations

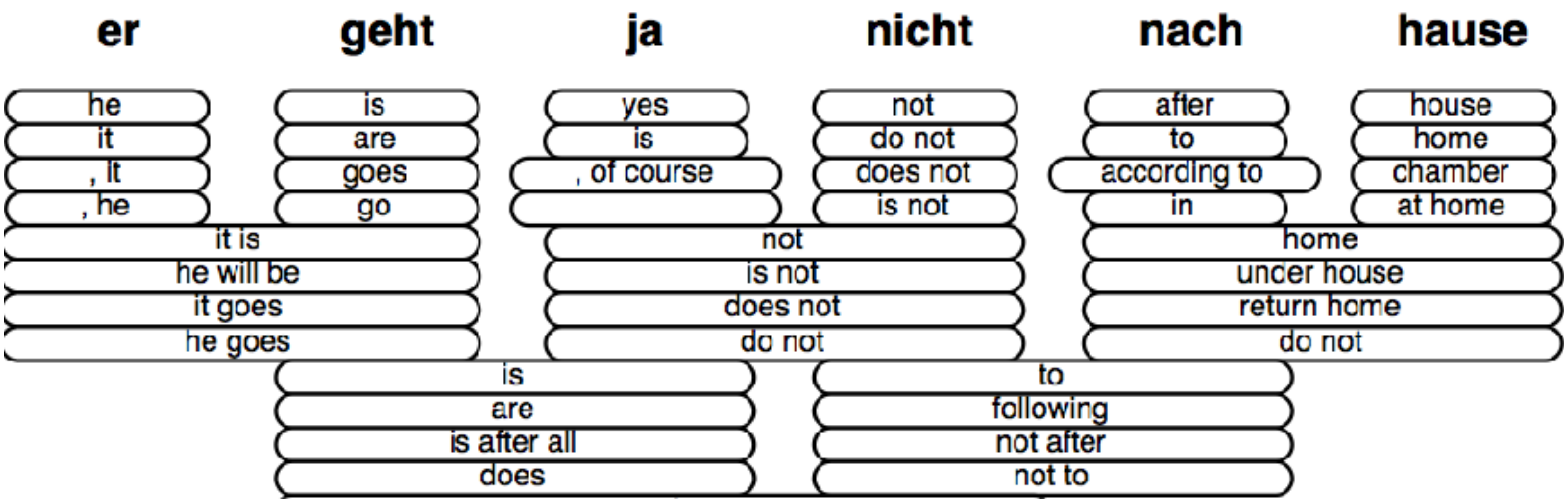
与 北 韩 有 邦交

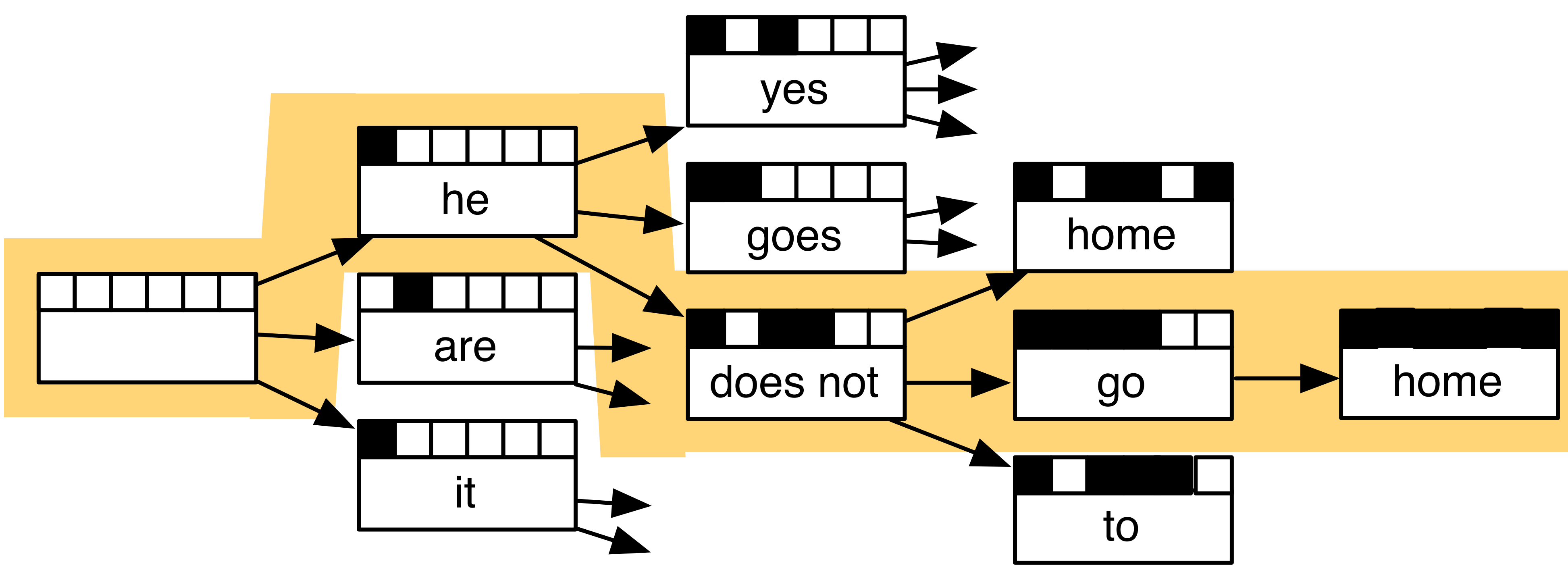
have diplomatic relations with North Korea

Phrase-based Decoder



Phrase-based Decoder





Discriminative Re-Ranking

	LM	TM1	TM2	Lex
12 cartoons insulting the prophet mohammad	4.5	3.0	9.0	6.0
12 cartoons attack the prophet mohammad	10.1	2.0	7.0	17.6
twelve comics offensive to the prophet mohammad	8.0	15.4	45.0	7.0
several drawings mocking the prophet mohammad	5.5	23.2	26.0	9.4

Every time I fire a linguist my performance goes up

- Longstanding debate about wheth information can help statistical tra
- Two camps

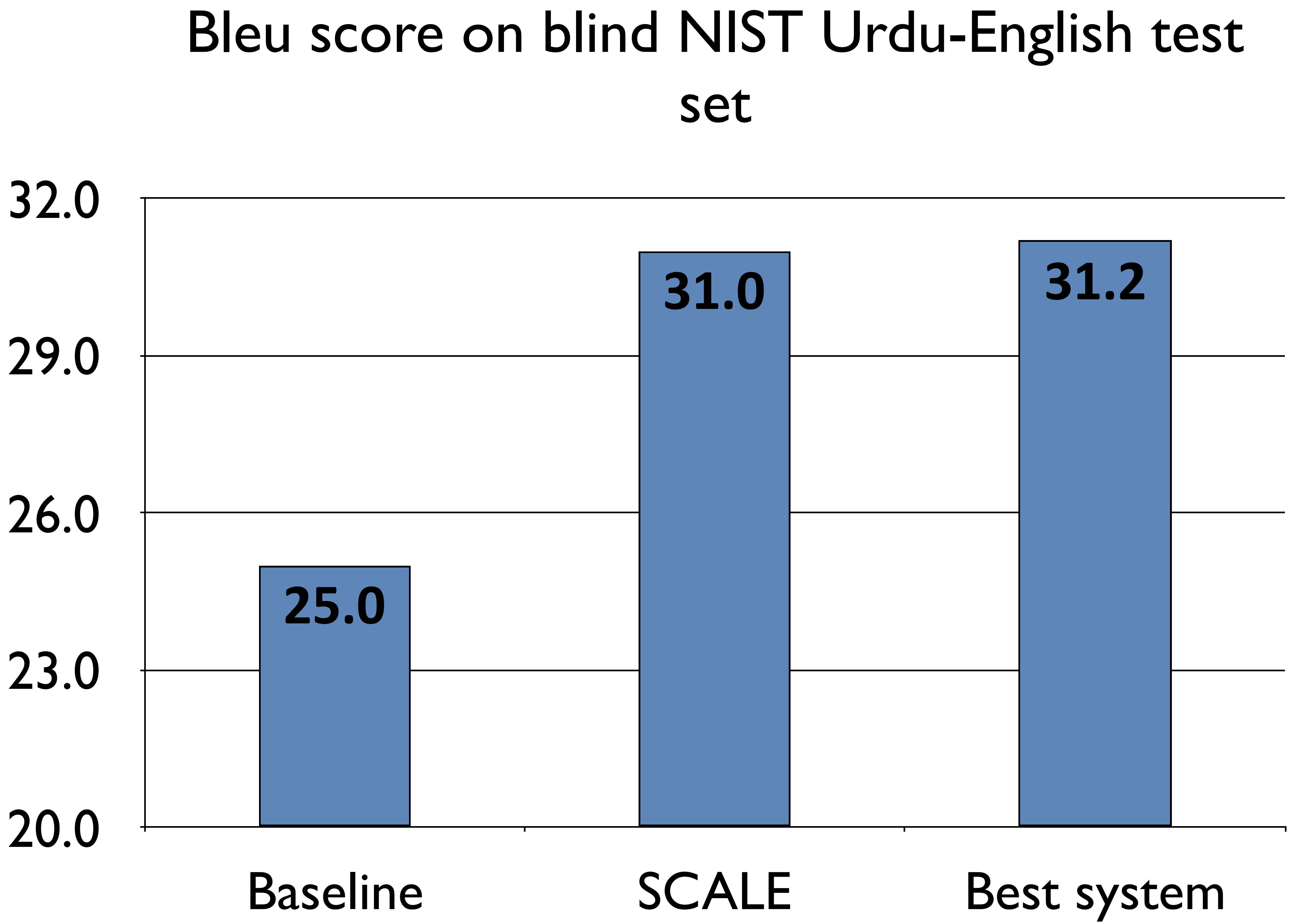
Syntax will improve translation

Simpler data-driven models will always win

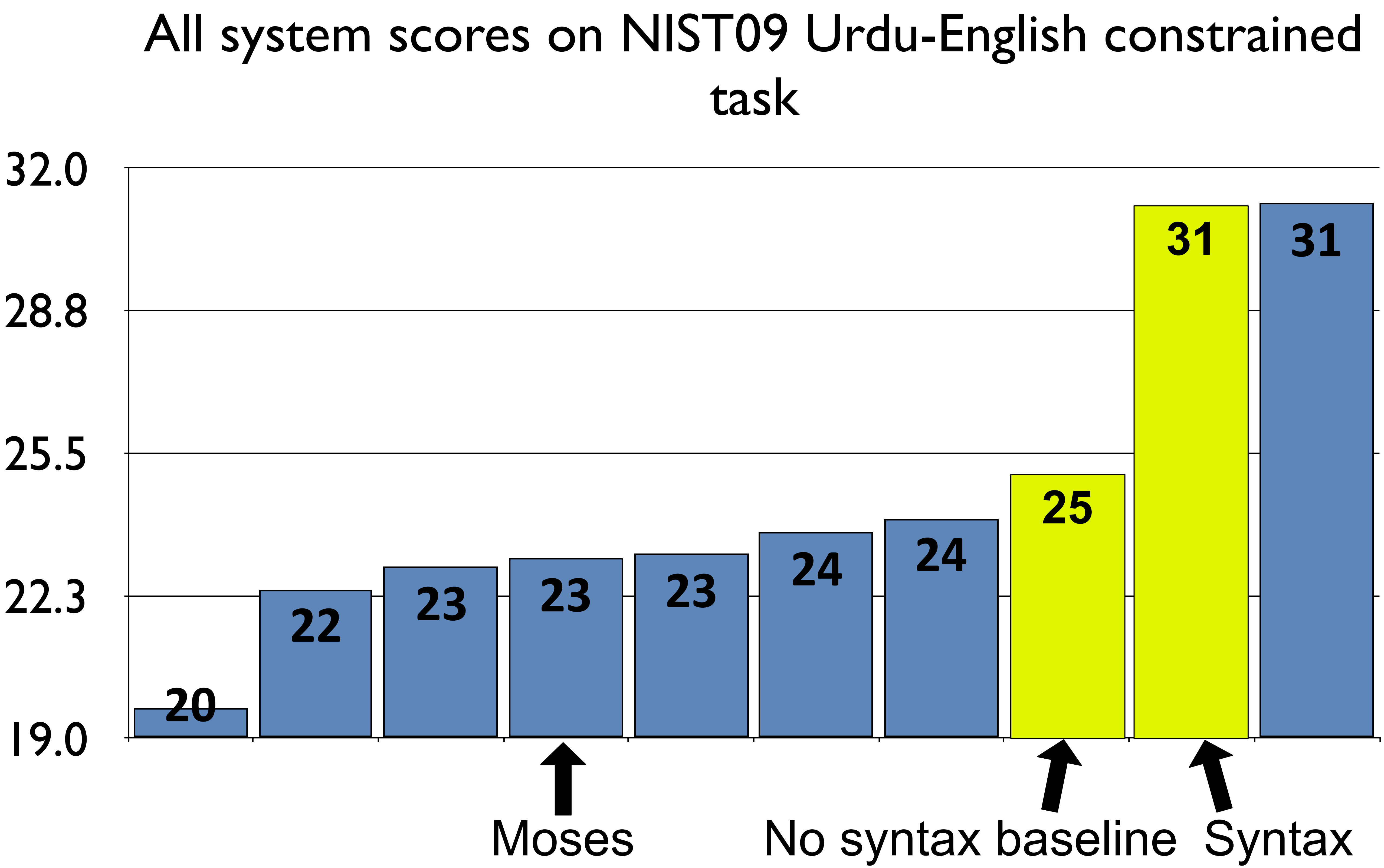
Urdu-English Machine Translation

- Results from the SCALE summer workshop
- 20 researchers working on the problem from many different angles
- I took a gamble on syntax
- My first foray into syntactic machine translation
- Worked well. Very, very well.

Results first



State of the Art Urdu Results



Joshua Decoder

- Synchronous context free grammars generate pairs of corresponding strings
- Can be used to describe translation and re-ordering between languages
- Because Joshua uses SCFGs, it translates sentences by parsing them

Synchronous Context Free Grammar

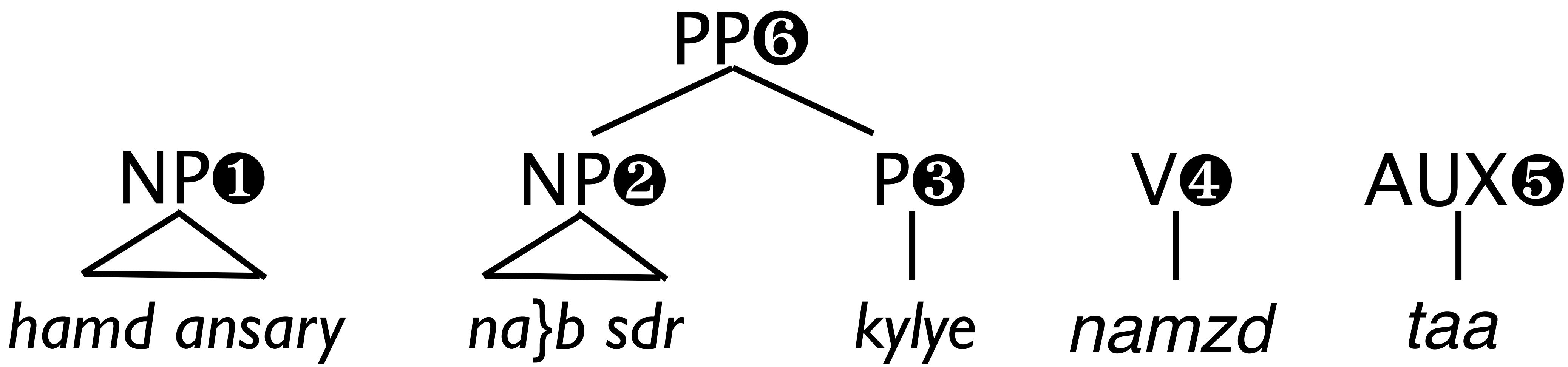
	Urdu	English
S →	NP① VP②	NP① VP②
VP →	PP① VP②	VP② PP①
VP →	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	*hamd ansary*	*Hamid Ansari*
NP →	*na}b sdr*	*Vice President*
V →	*namzd*	*nominated*
P →	*kylye*	*for*
AUX →	*taa*	*was*

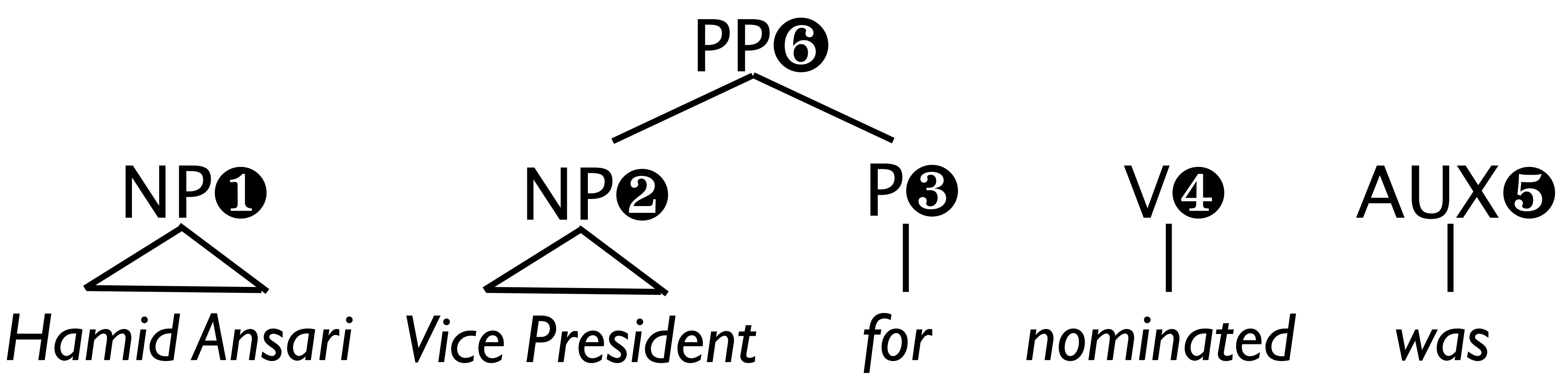
NP❶ NP❷ P❸ V❹ AUX❺

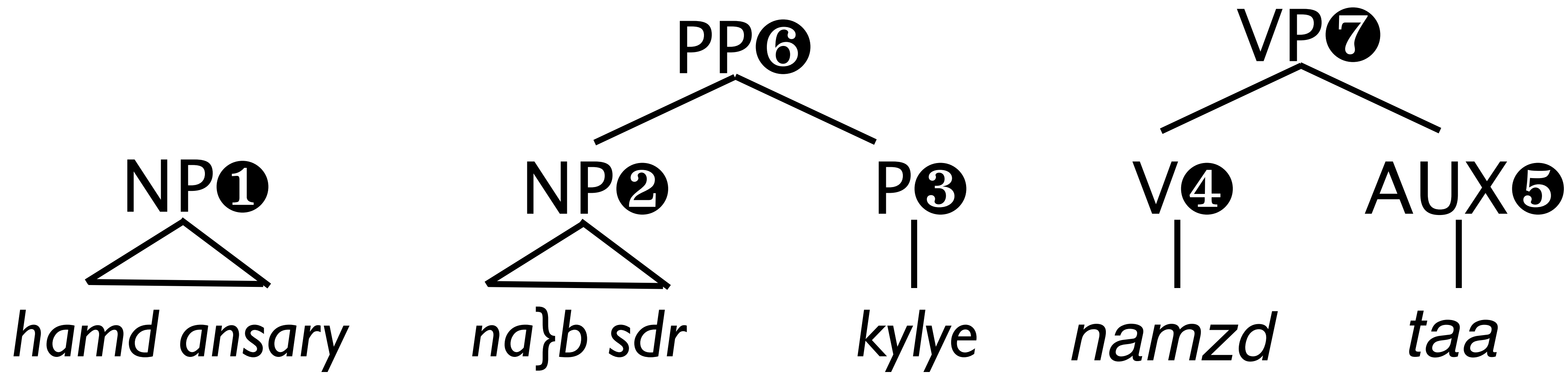
hamd ansary *na}b sdr* *kylye* *namzd* *taa*

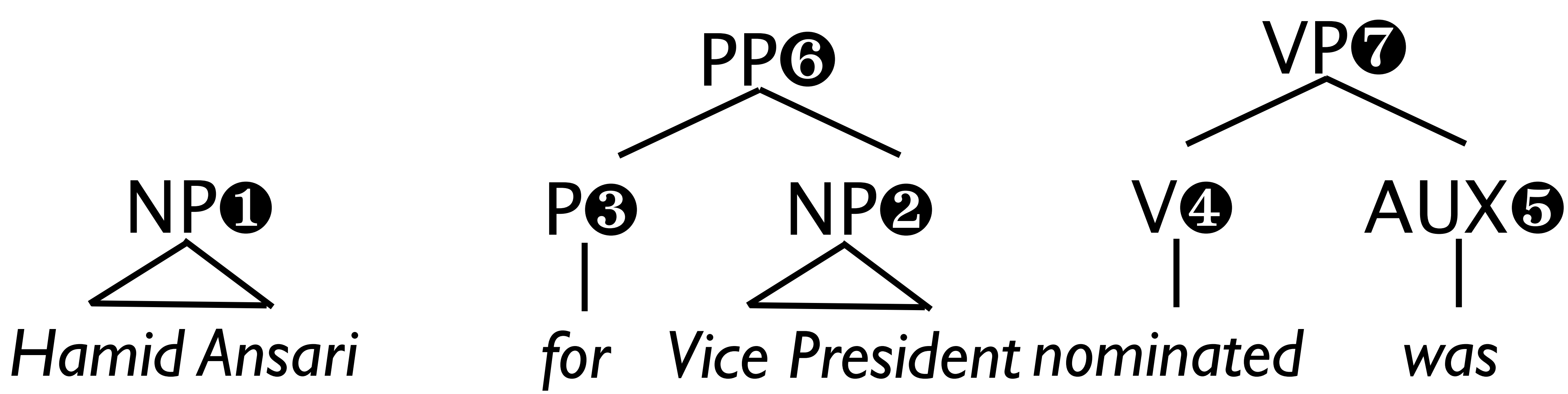
NP❶ NP❷ P❸ V❹ AUX❺

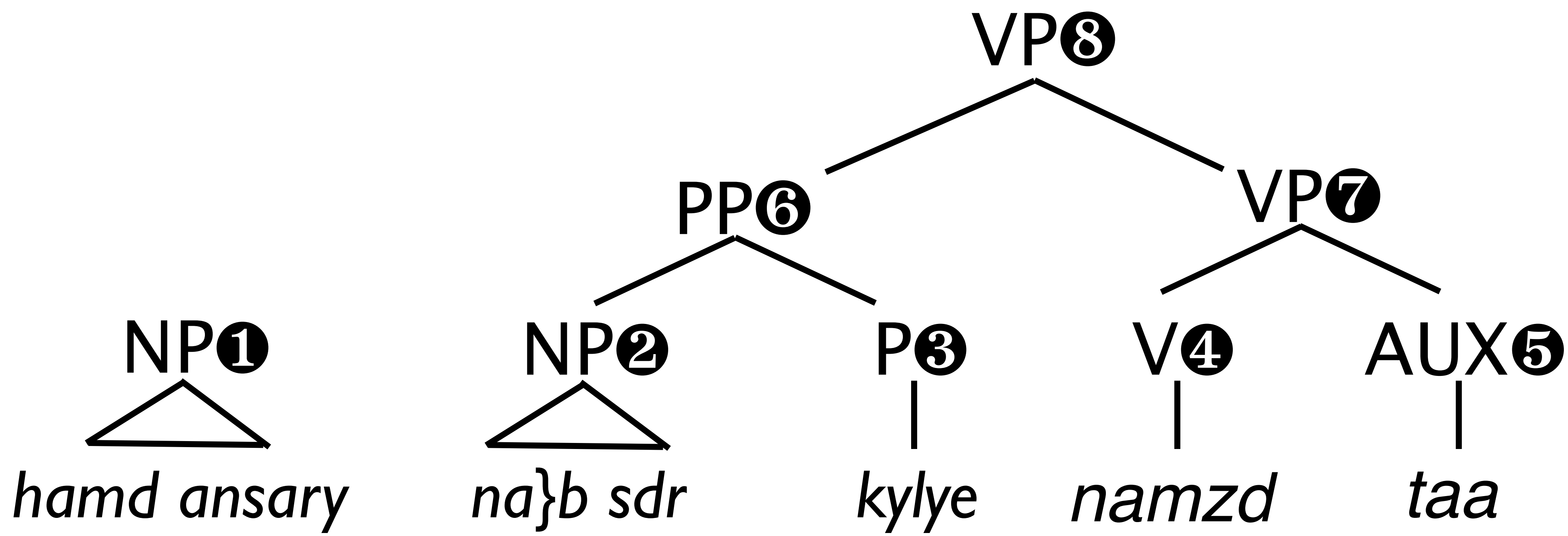
Hamid Ansari *Vice President* *for* *nominated* *was*

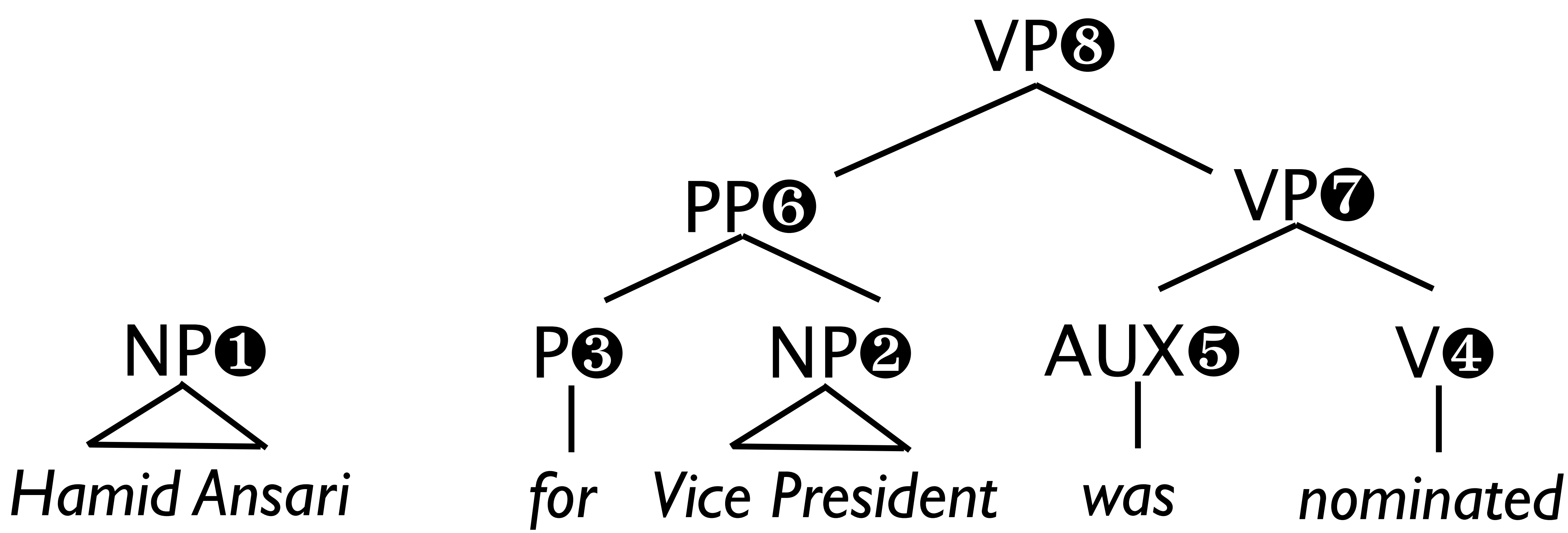

PP❻
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

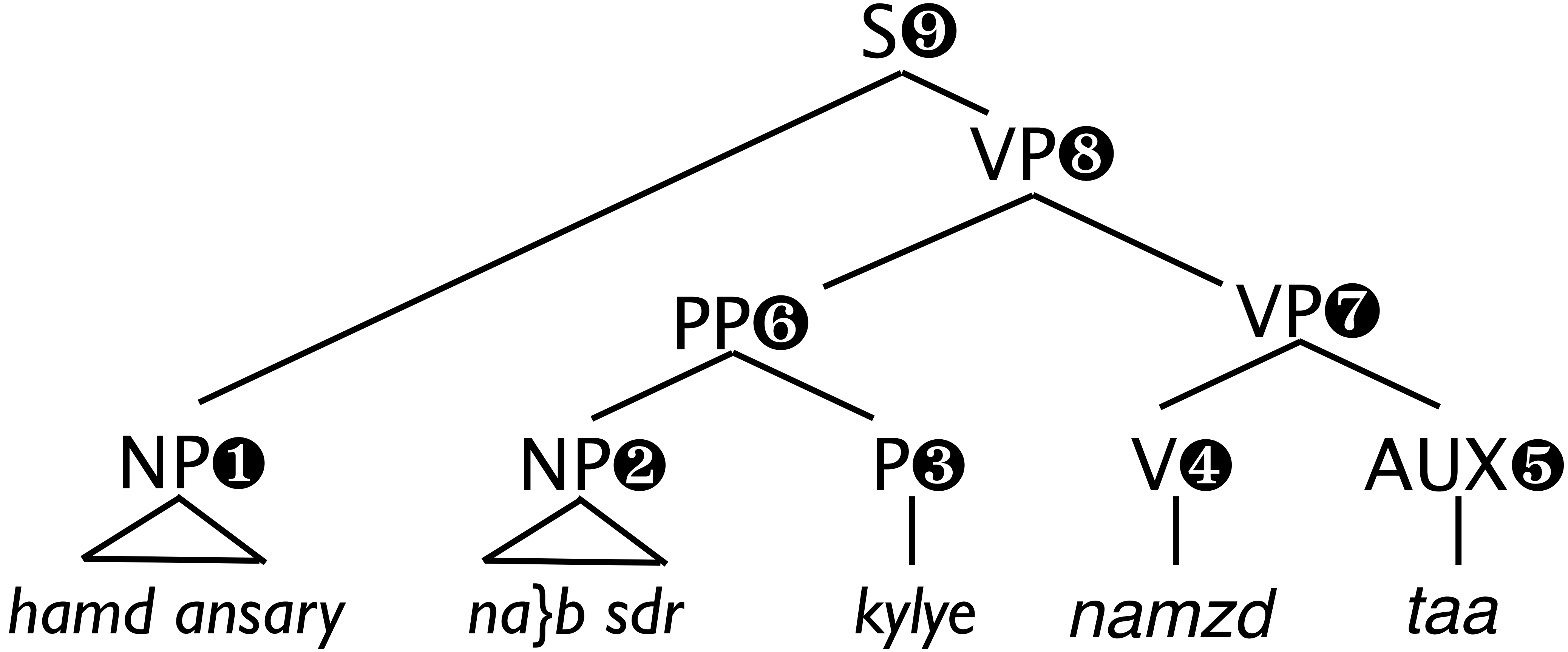

PP❻
NP❶
NP❷
P❸
V❹
AUX❺
Hamid Ansari
Vice President
for
nominated
was

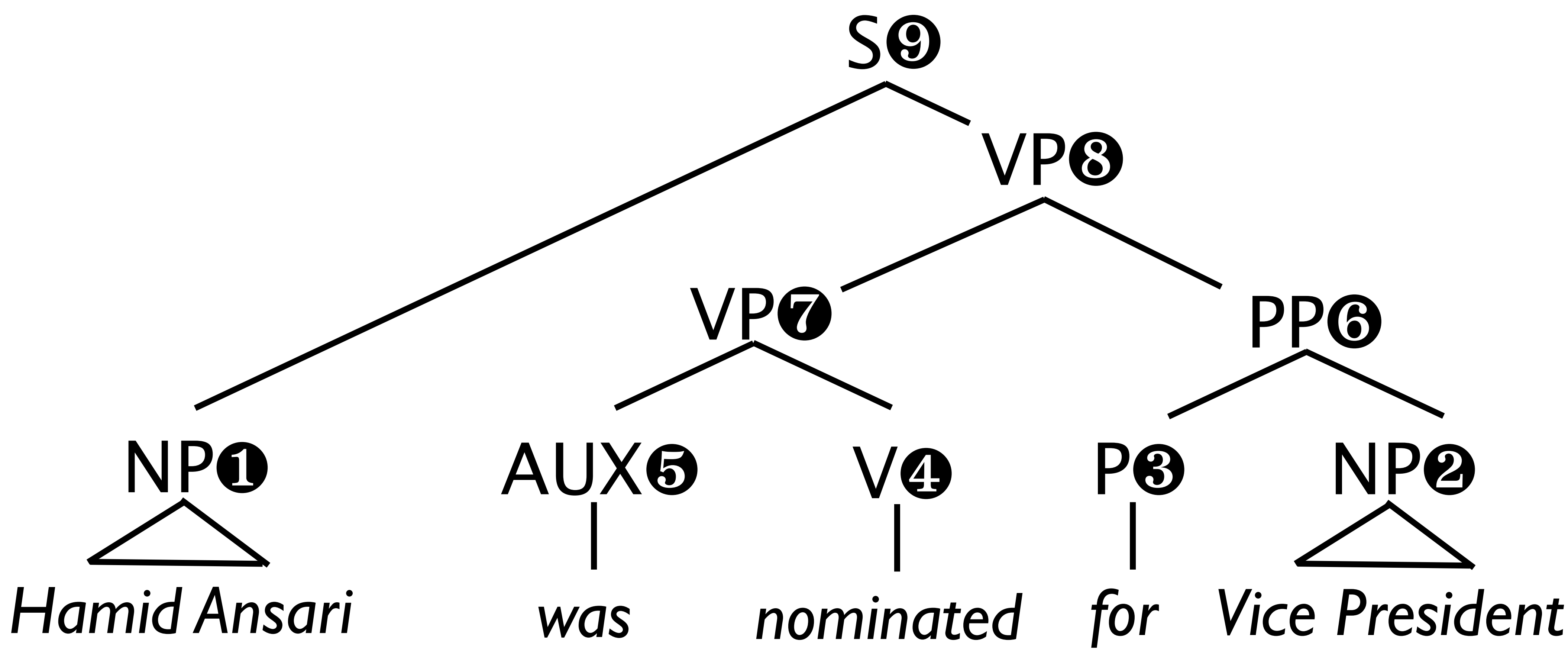

PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


PP❻
VP❼
NP❶
P❸
NP❷
V❹
AUX❺
Hamid Ansari
for
Vice President
nominated
was

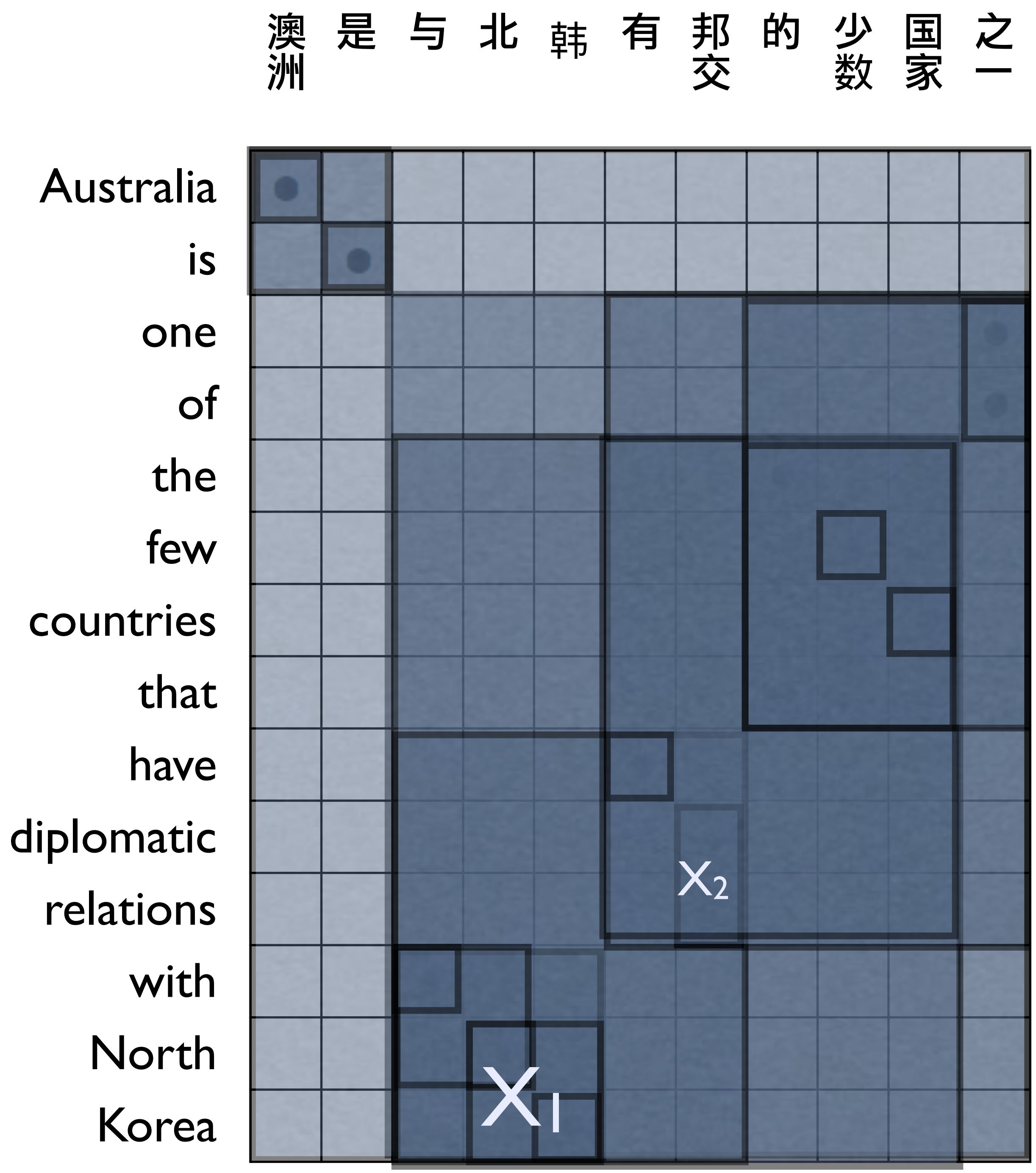

VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


VP❽
PP❻
VP❼
NP❶
P❸
NP❷
AUX❺
V❹
Hamid Ansari
for
Vice President
was
nominated


S❾
VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


S❾
VP❽
VP❼
PP❻
NP❶
AUX❺
V❹
P❸
NP❷
Hamid Ansari
was
nominated
for
Vice President

Extracting Hiero rules



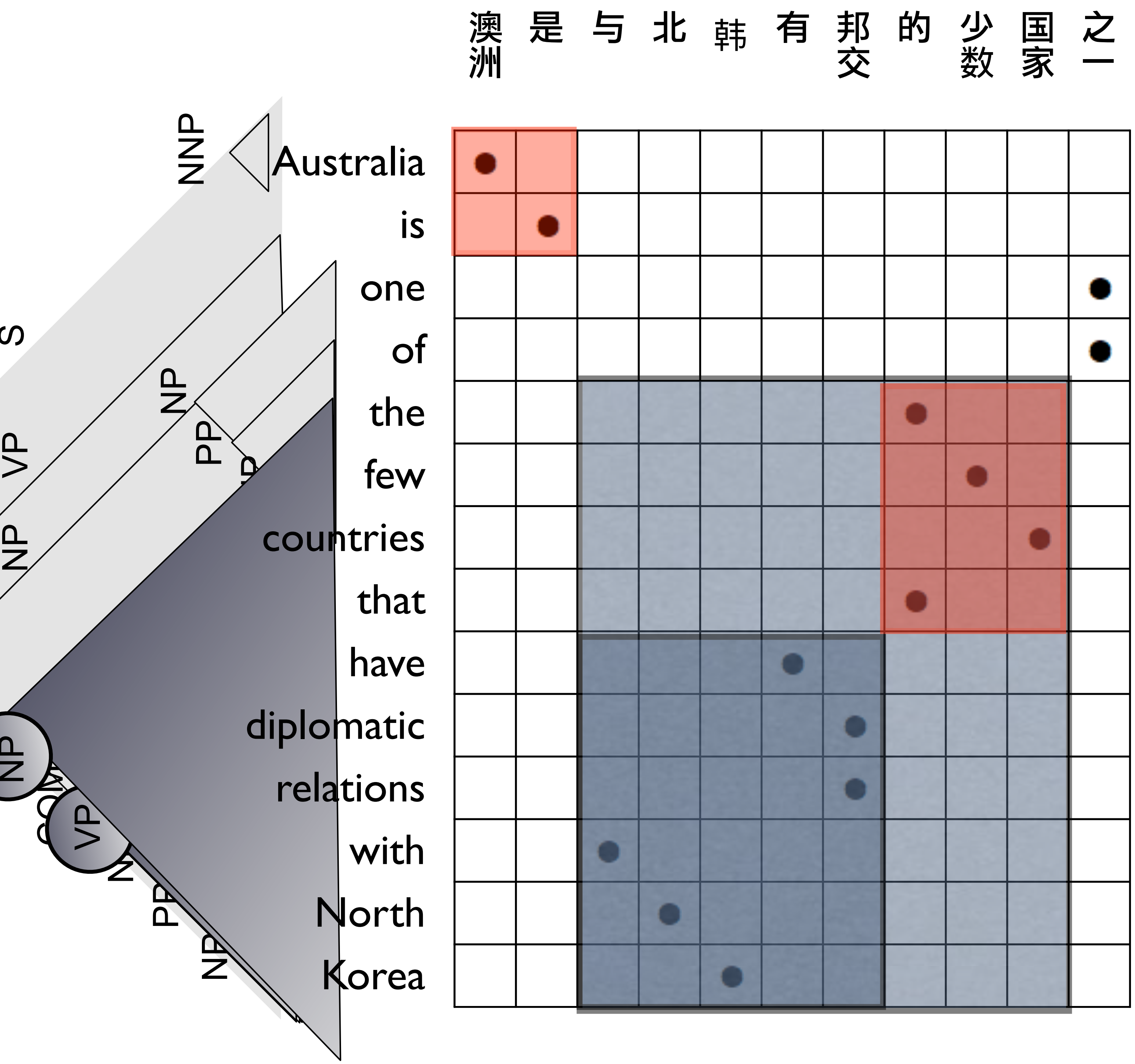
X → 与 北 韩 有 邦交,
have diplomatic relations
with North Korea

X → 邦交,
diplomatic relations

X → 北 韩,
North Korea

X → 与 X_1 有 X_2,
have X_2 with X_1

Extracting Syntactic Rules



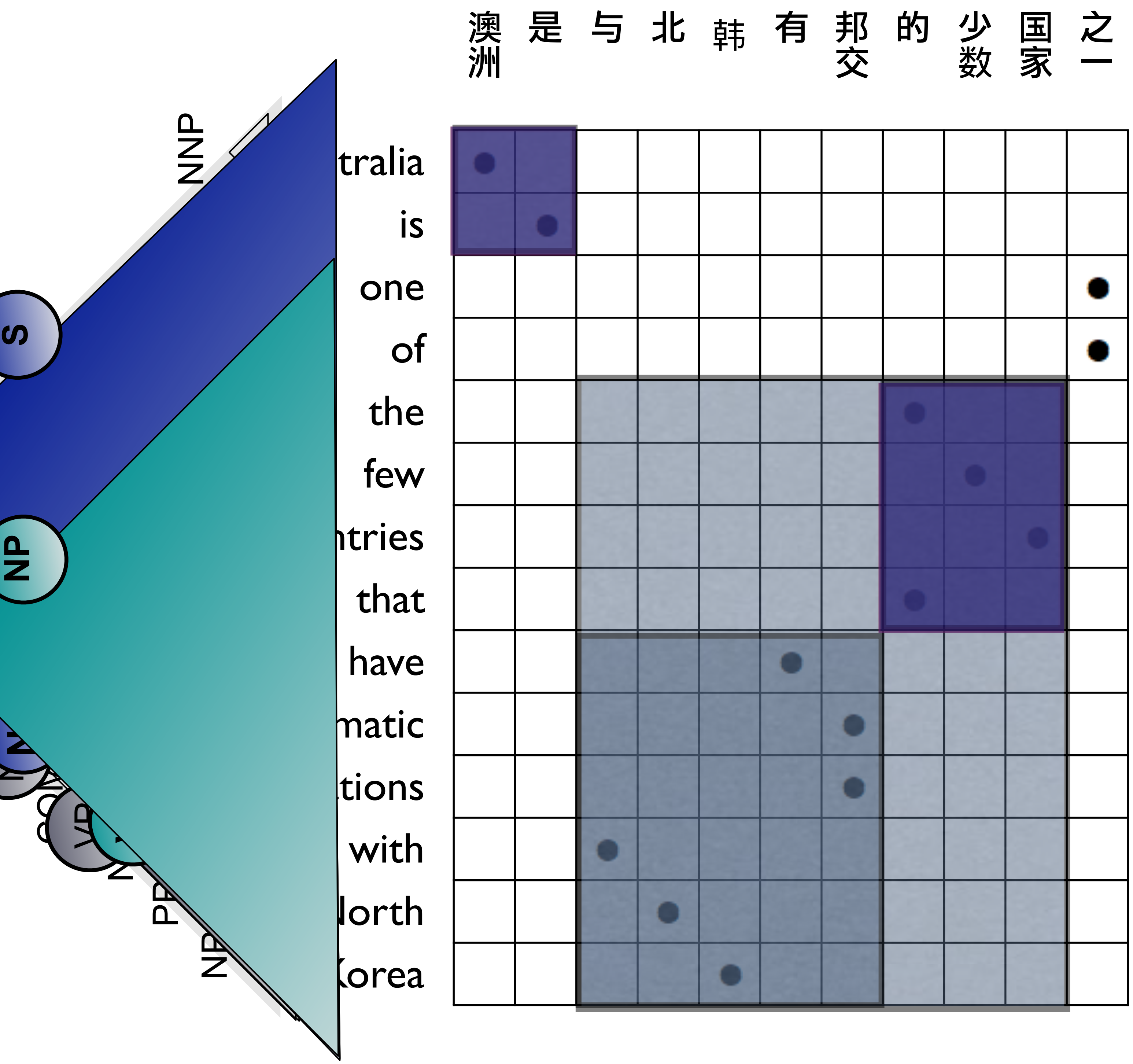
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

NP → 与 北 韩 有 邦交 的 少数 国家, the few countries that have diplomatic relations with North Korea

??? → 的 少数 国家, the few countries that

??? → 澳洲 是, Australia is

Extracting Syntactic Rules



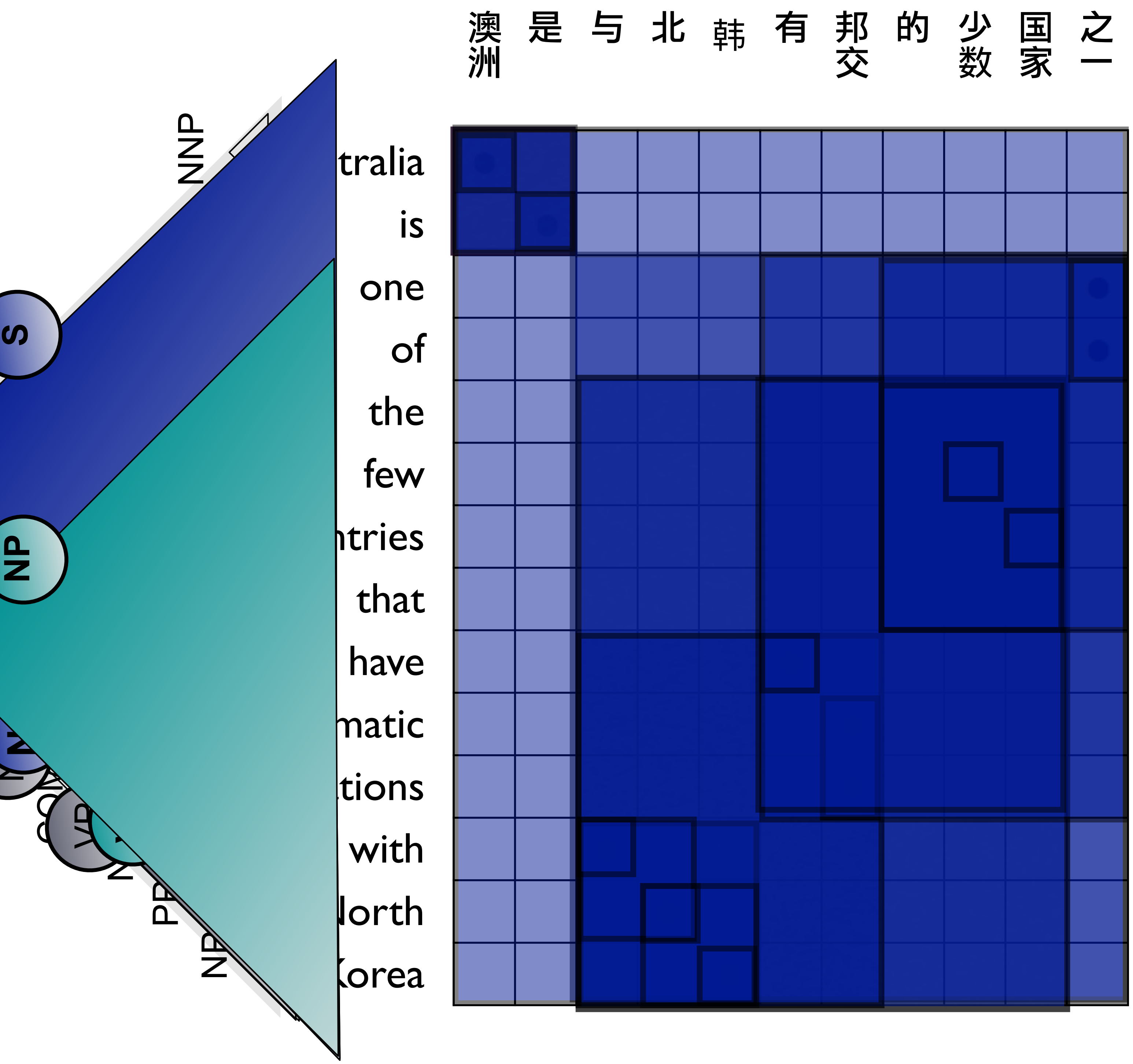
VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

NP → 与 北 韩 有 邦交 的 少数 国家, the few countries that have diplomatic relations with North Korea

NP/ VP → 的 少数 国家, the few countries that

S/ NP → 澳洲 是, Australia is

Extracting Syntactic Rules



VP → 与 北 韩 有 邦交, have diplomatic relations with North Korea

NP → 与 北 韩 有 邦交 的 少数 国家, the few countries that have diplomatic relations with North Korea

NP/ VP → 的 少数 国家, the few countries that

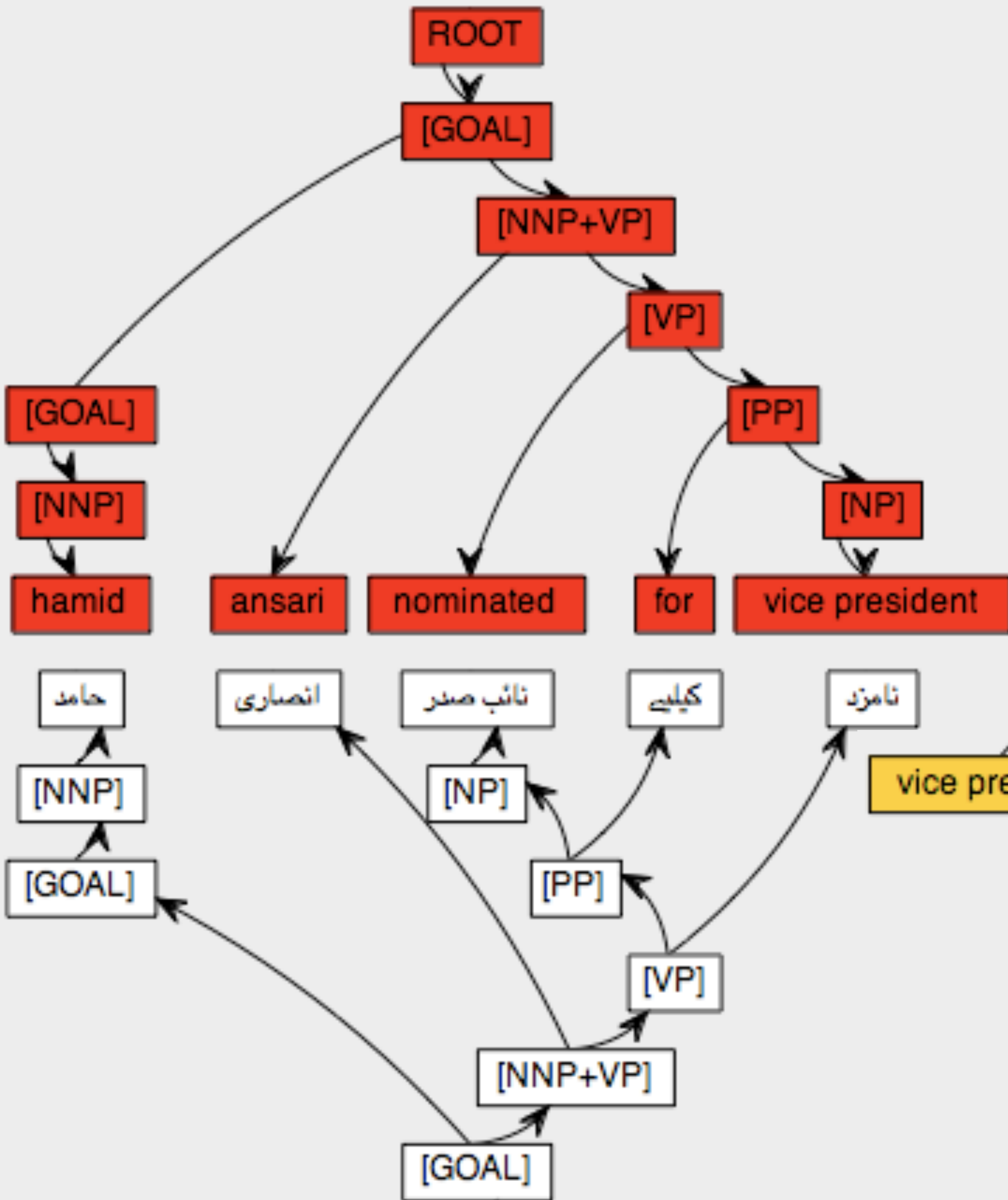
S/ NP → 澳洲 是, Australia is

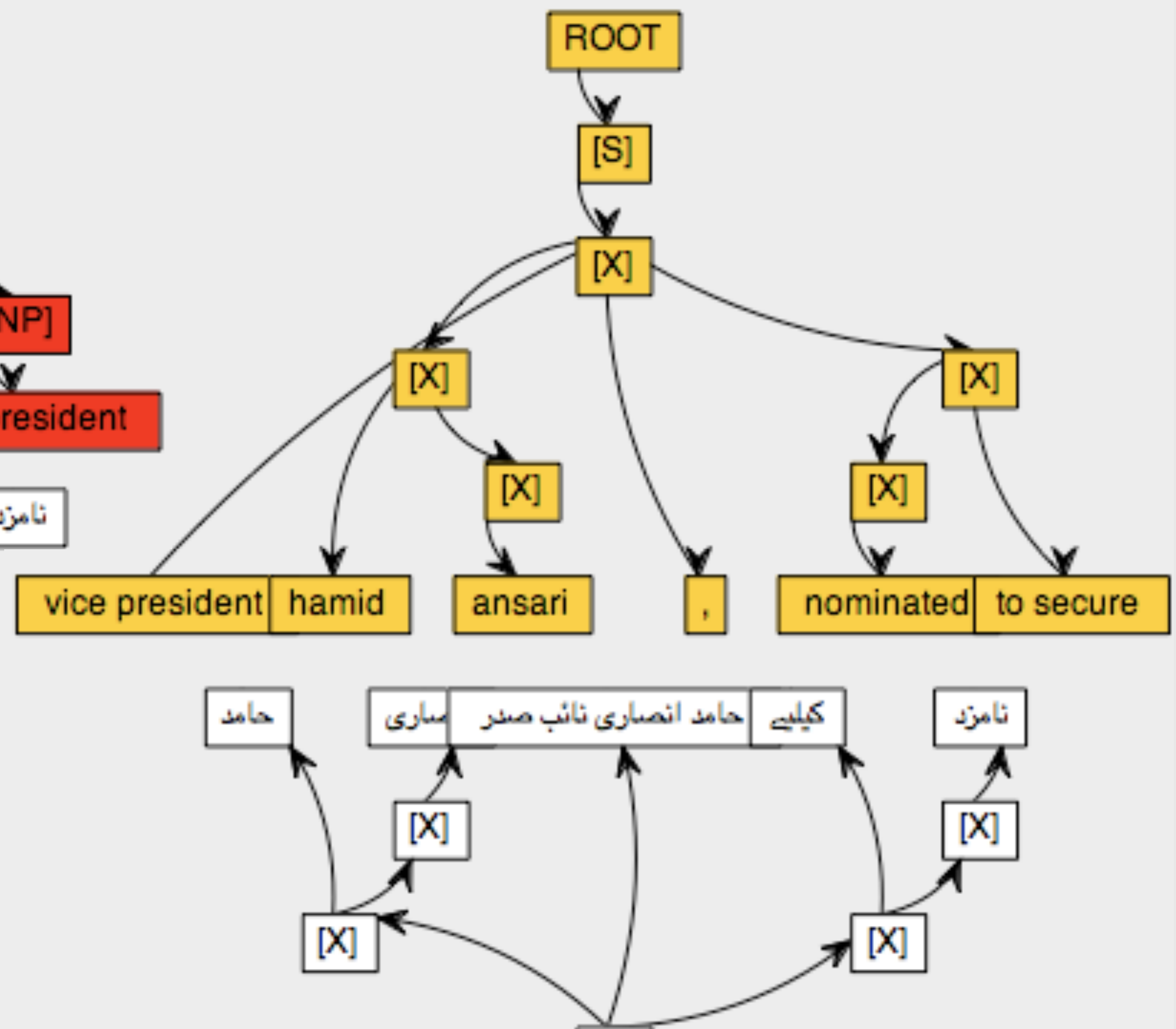
New training paradigm

- Training data: word-aligned bilingual parallel corpus, with parse trees
  - No need to parse the Urdu, just parse the English
  - Method is therefore transferable to other resource poor languages
- Extract SCFG rules with syntactic nonterminals
- For non-constituent phrases use CCG-style nonterminals
- Same coverage as Hiero model

Syntax captures Urdu reordering

More details in Baker, Bloodgood, Callison-Burch, Dorr, Fliardo, Levin, Miller and Piatko (AMTA 2010 and Computational Linguistics 2012)





Translation improvements

'first nuclear experiment in 1990 was'
Thomas red Unilever National Laboratory of the United States
in ویپن designer, are already working on the book of Los ایلموس National Laboratory ڈینی, former director of the technical انٹیلجنس written with the cooperation of سٹلمین.

This book 'nuclear express: political history and the expansion of bomb' has been written, and the two writers have also claimed that the country has made nuclear bomb is he or any other country's nuclear secrets to
چرائیے or that of any other nuclear power cooperation is achieved.

First nuclear test conducted in 1990
Thomas Reed, who has worked as a weapons designer at Livermore National Laboratory in the United States, has written a book in collaboration with Danny Stillman, former director of the technical intelligence division at Los Alamos National Laboratory.

In their book, 'The Nuclear Express: A Political History of the Bomb and its Proliferation,' Reed and Stillman claim that every country that has ever produced a nuclear bomb has been able to do so because it stole the nuclear secrets of another country or enjoyed the cooperation of another nuclear power.

Who did what to whom?

Baseline

He said that China, North Korea, Iran, Syria, Pakistan, through Egypt, Libya and Yemen is to provide nuclear technology.

Thomas was red when this question why China has provided the nuclear technology to Pakistan, In response, He said as China and India was joint enemy of Pakistan.

Syntactic final system

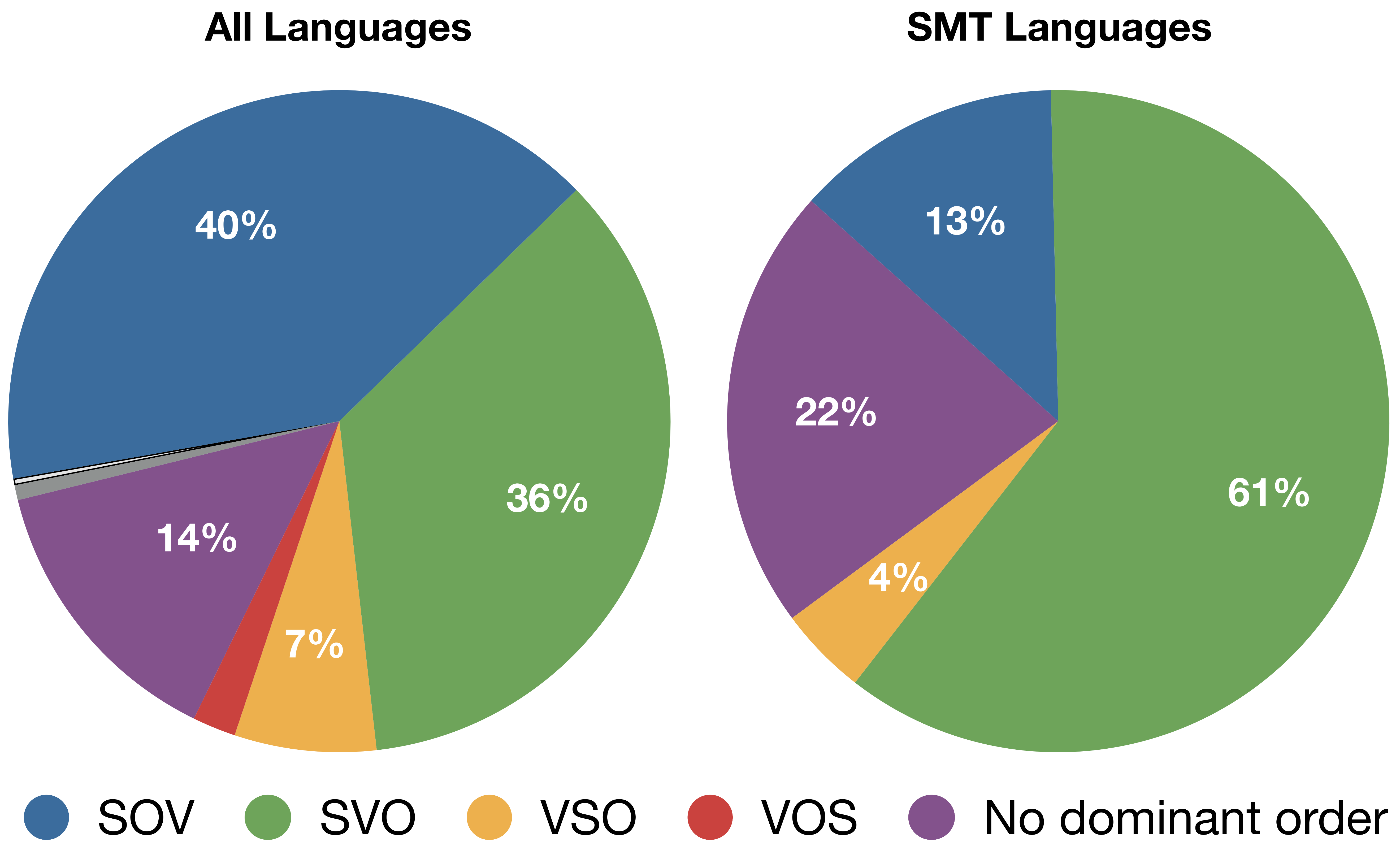
He said that China would provide nuclear technology to North Korea, Iran, Syria, Pakistan, Egypt, Libya and Yemen.

Thomas red when was this question why China has provided to Pakistan nuclear technology, he said in response to China, Pakistan and India as a common enemy.

Why did this work?

- Using syntax-based translation models resulted in huge improvements in quality
- Previous work on syntax did not shown significant gains, so why did it work here?
- Urdu is an ideal language to show off the advantages of syntax
  - Very small amount of training data
  - Very different word order than English
- Can't simply memorize translations of phrases
- Must generalize

Distribution of Word Orders



Available training data