Machine Translation

CIS 526

Instructor: Chris Callison-Burch
TA: Jonny Weese

Course web site

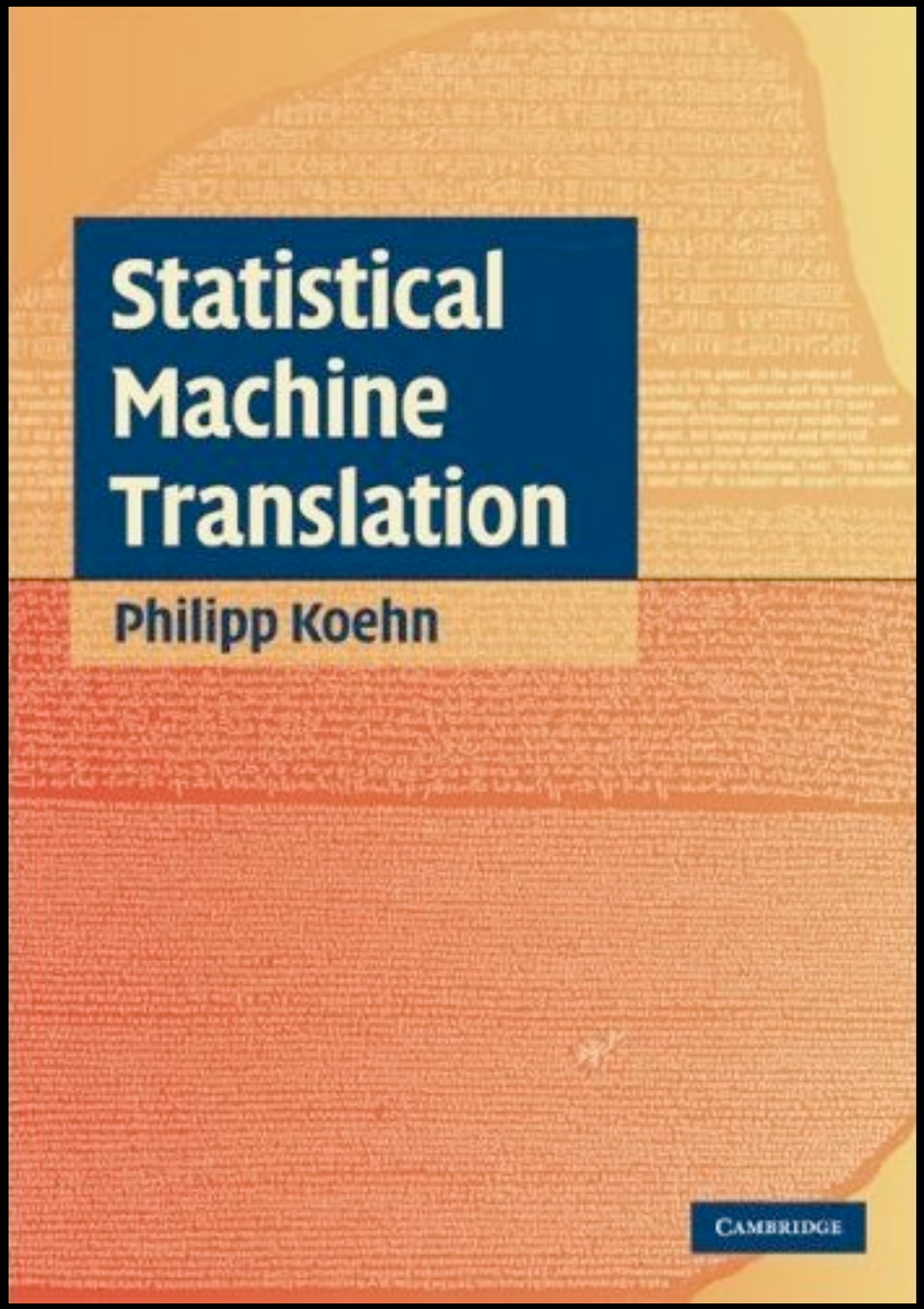
mt-class.org/penn

Course materials developed with

Adam Lopez

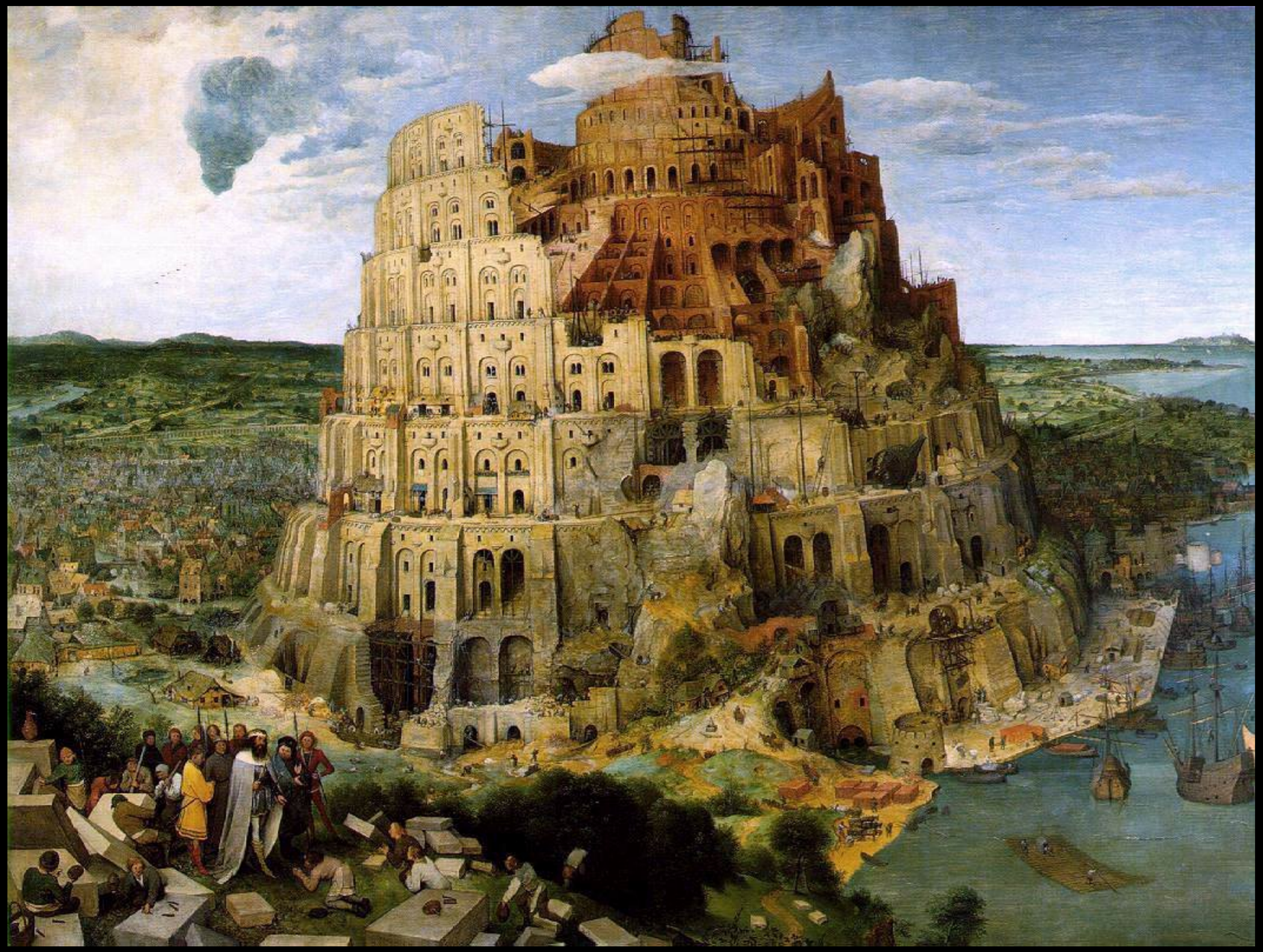
Matt Post

Textbook


Statistical
Machine
Translation
Philipp Koehn
CAMBRIDGE

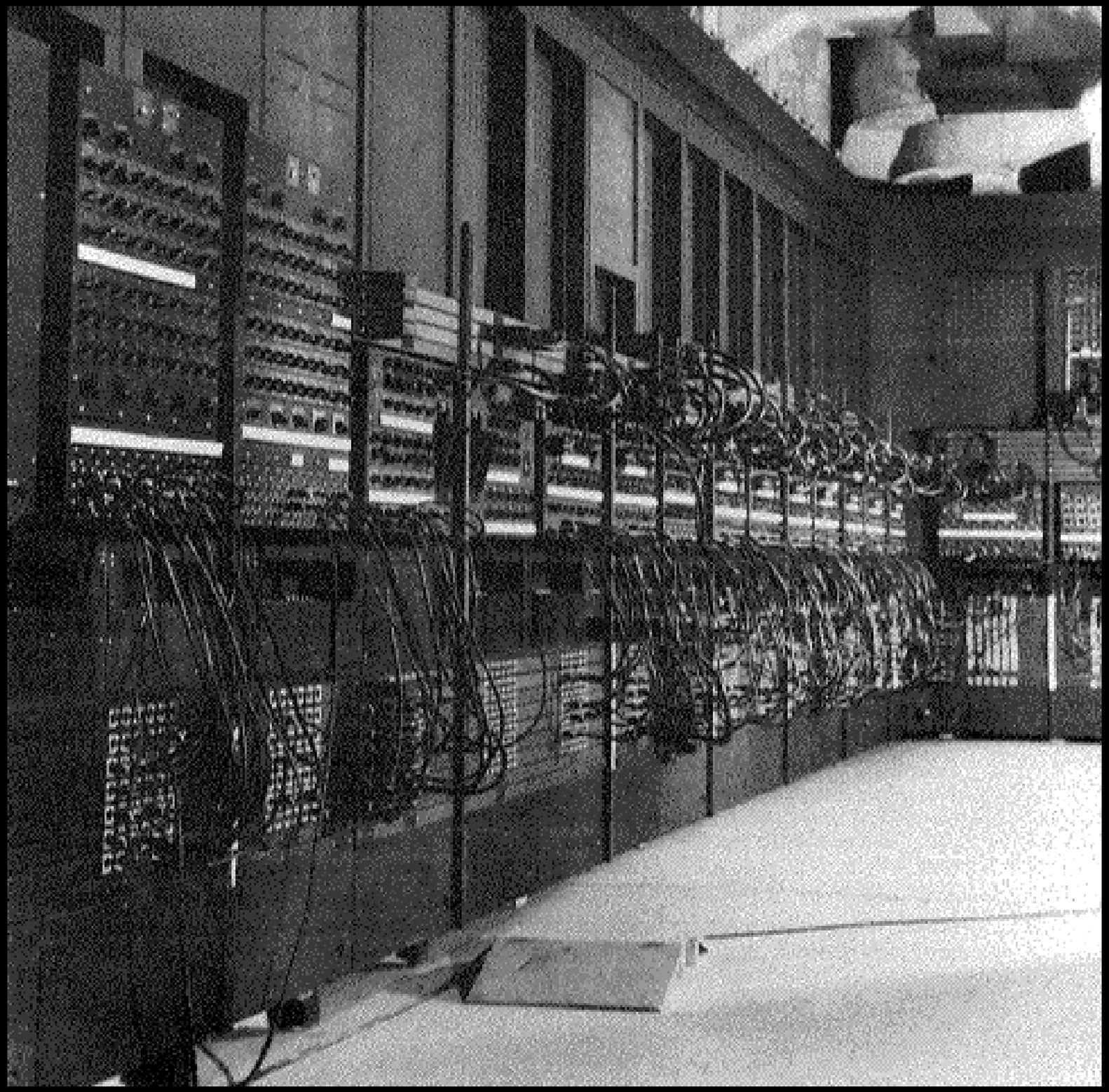
نائب امریکی صدر ڈک چینی کا کہنا ہے کہ میں اسامہ بن لادن
کو زندہ یا مردہ دیکھنا چاہتا ہوں۔

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



The Tower of Babel

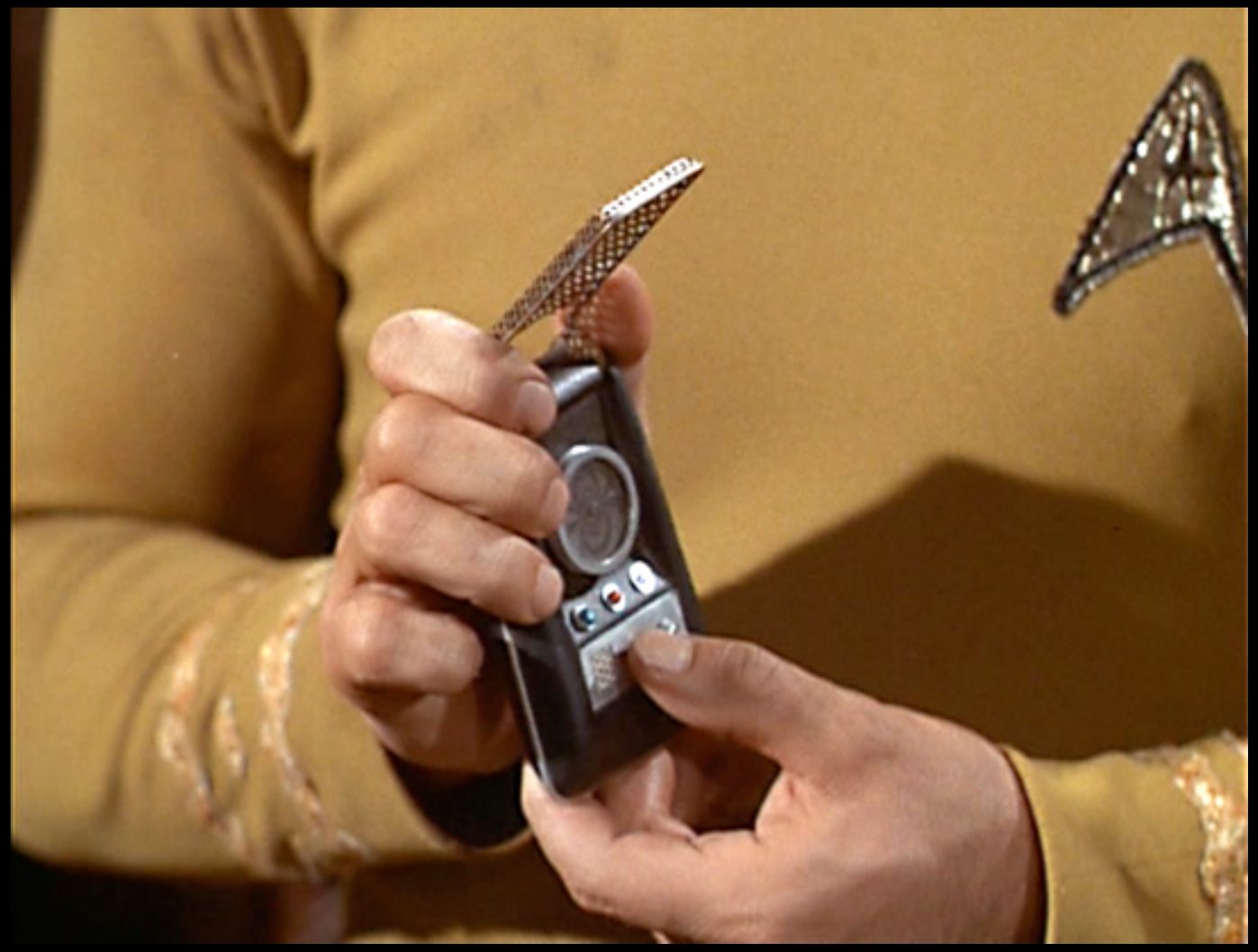
Pieter Brueghel the Elder (1563)



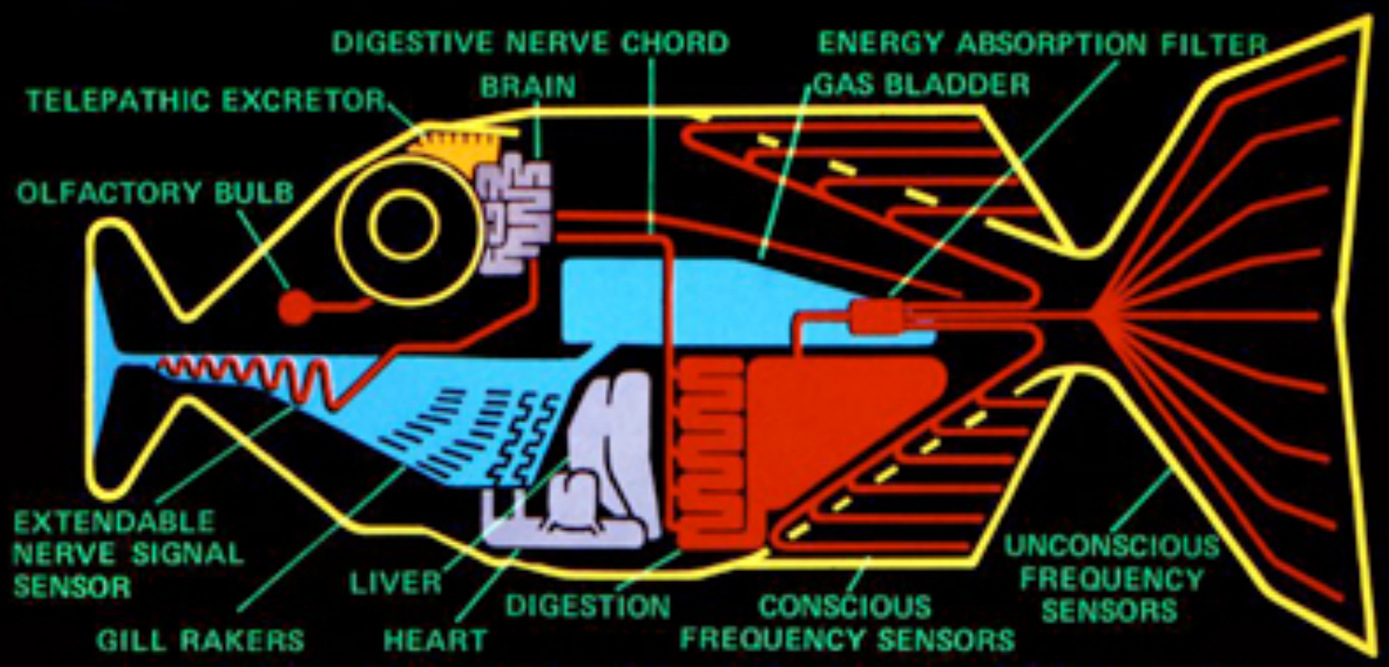
ENIAC (1946)

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)



Star Trek



Hitchhiker's Guide to the Galaxy

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!

Google

Translate

English | Spanish | French | Detect language ▾ ⇄ English | Spanish | Arabic ▾

Afrikaans	Cebuano	Finnish	Hungarian	Latin	Romanian	Turkish
Albanian	Chinese (Simplified)	French	Icelandic	Latvian	Russian	Ukrainian
Arabic	Chinese (Traditional)	Galician	Indonesian	Lithuanian	Serbian	Urdu
Armenian	Croatian	Georgian	Irish	Macedonian	Slovak	Vietnamese
Azerbaijani	Czech	German	Italian	Malay	Slovenian	Welsh
Basque	Danish	Greek	Japanese	Maltese	Spanish	Yiddish
Belarusian	Dutch	Gujarati	Javanese	Marathi	Swahili	
Bengali	**English**	Haitian Creole	Kannada	Norwegian	Swedish	
Bosnian	Esperanto	Hebrew	Khmer	Persian	Tamil	
Bulgarian	Estonian	Hindi	Korean	Polish	Telugu	
Catalan	Filipino	Hmong	Lao	Portuguese	Thai	

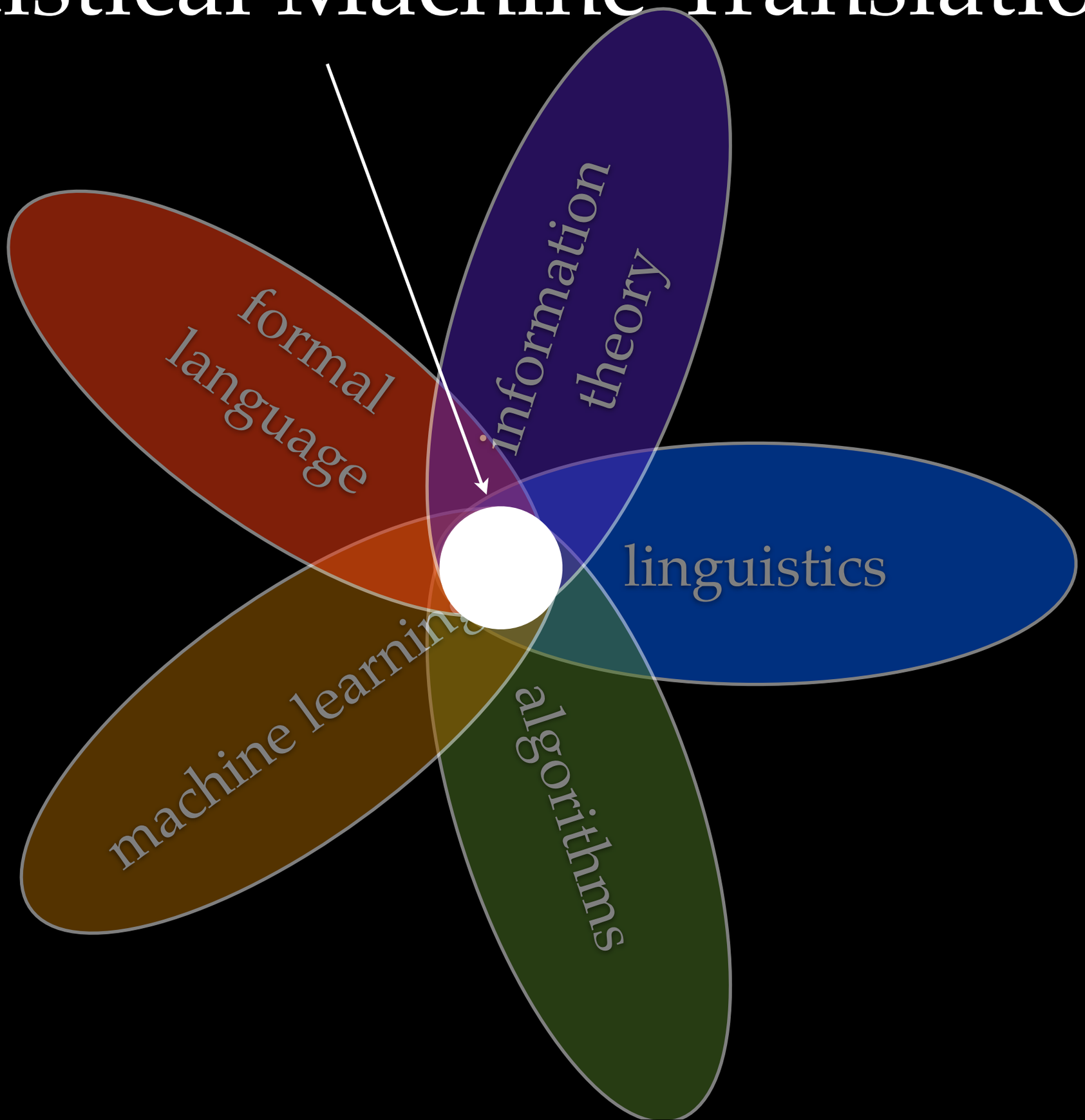

www.un.org
http://www.un.org/english/
Google
National Bureau
http://www.stats.gov.cn/engl
NBS
中文简体
www.stats.gov.cn
News and Coming
Memorial Ceremony for Late Deputy Commissioner Zhu Xiangdong Held in Beijing(09.16)
The Urban Investment in Fixed Assets Continued Increasing in August(09.16)
German Delegation Visited the National Bureau of Statistics of China(09.15)
The Value-added of Industry up by 16 Percent in August(09.15)
The Total Retail Sale of Consumer Goods Increased in August(09.14)
The Consumer Price Index (CPI) Increased in August(09.13)
The producers' Price Index (PPI) For Manufactured Goods Kept Advancing in August(09.12)
Global Manager of ICP of World Bank Visited Beijing(09.08)
What's New
Monthly Data Updated(09.15)
Statistical Data: Women and Men in China----Facts and Figures 2004(09.08)
Monthly Data Updated(09.07)
Monthly Data Updated(08.29)
Monthly Data Updated(08.23)
Agency Information
Laws & Regulations
Survey Program s
Statistical Standards
Statistical Indicators
New Releases
Int. Cooperation
NBS/IMF Meeting
Int. Conference On Ag Census
Statistical Data
Monthly
Yearly
Census
Others
Related links
Chinese Version
Others
Search
Live and On-Demand Webcasts, 24 Hours a Day: Click on UN Webcas


联合国主页
http://www.un.org/chinese/
Google
http://www.stats.gov.cn/
中华人民共和国国家
National Bureau of Statistics
中文繁体
ENGLISH
统计公报
统计数据
统计分析
统计出版
统计法规
统计管理
数据直报
统计链接
最新统计动态
重要公告
统计机构
统计动态
统计标准
统计制度
统计知识
联合国网播直播

Statistical Machine Translation

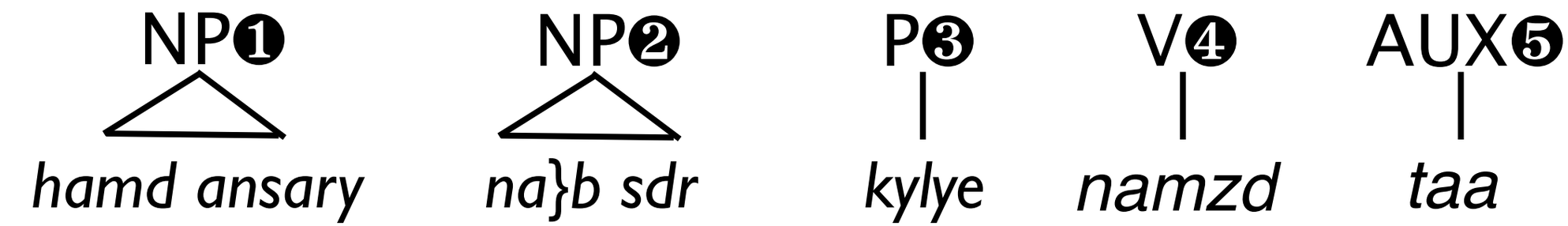
Develop a statistical ***model*** of translation that can be ***learned*** from ***data*** and used to ***predict*** the correct English translation of new Chinese sentences.

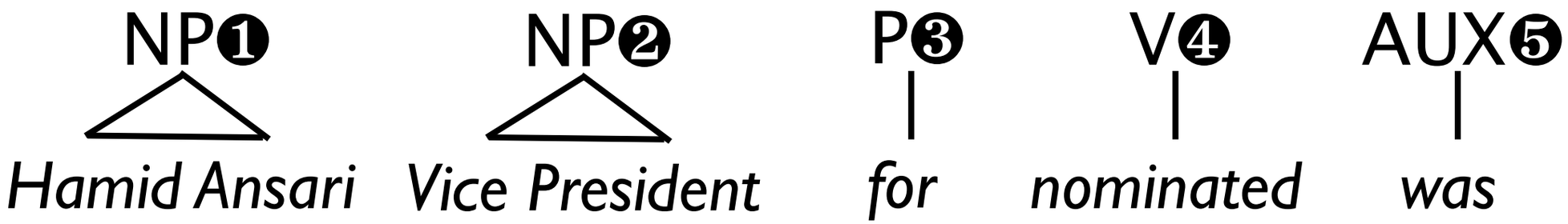
Statistical Machine Translation

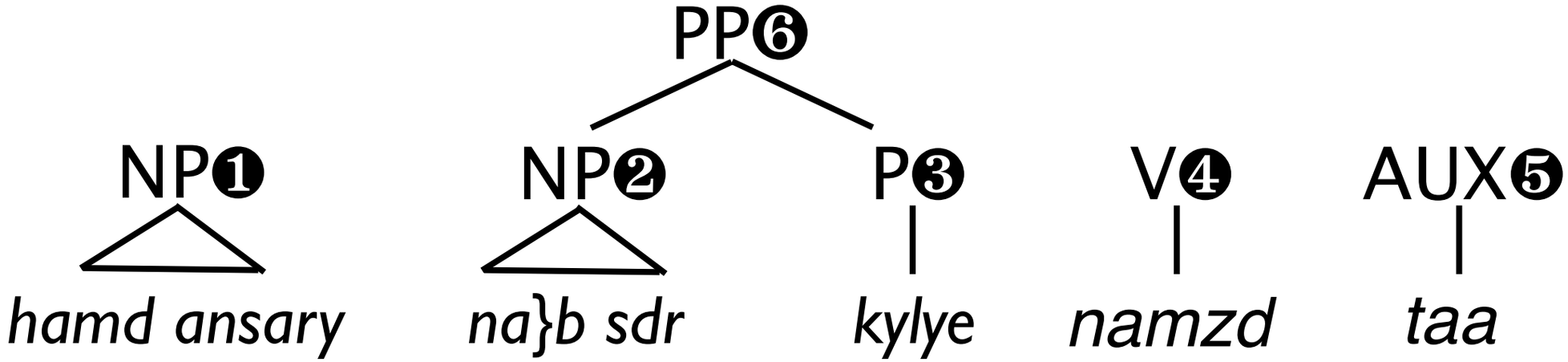


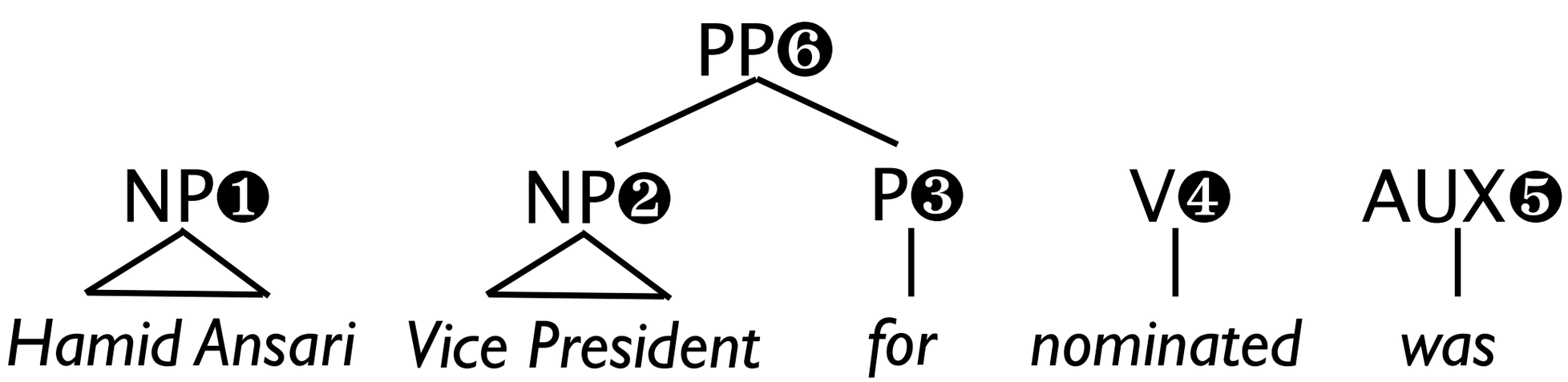
Synchronous Context Free Grammar

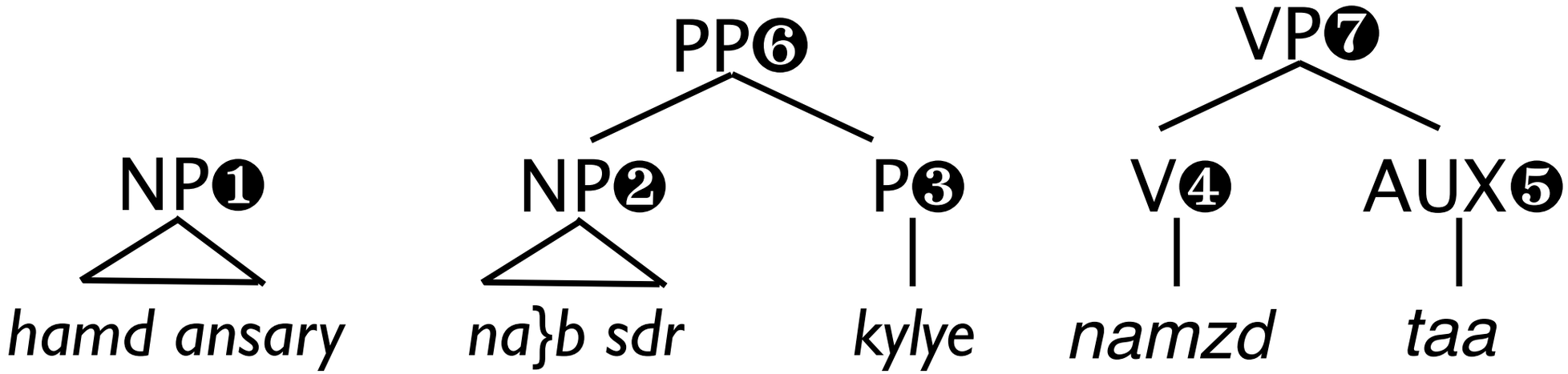
	Urdu	English
S →	NP① VP②	NP① VP②
VP→	PP① VP②	VP② PP①
VP→	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	*hamd ansary*	*Hamid Ansari*
NP →	*na}b sdr*	*Vice President*
V →	*namzd*	*nominated*
P →	*kylye*	*for*
AUX →	*taa*	*was*

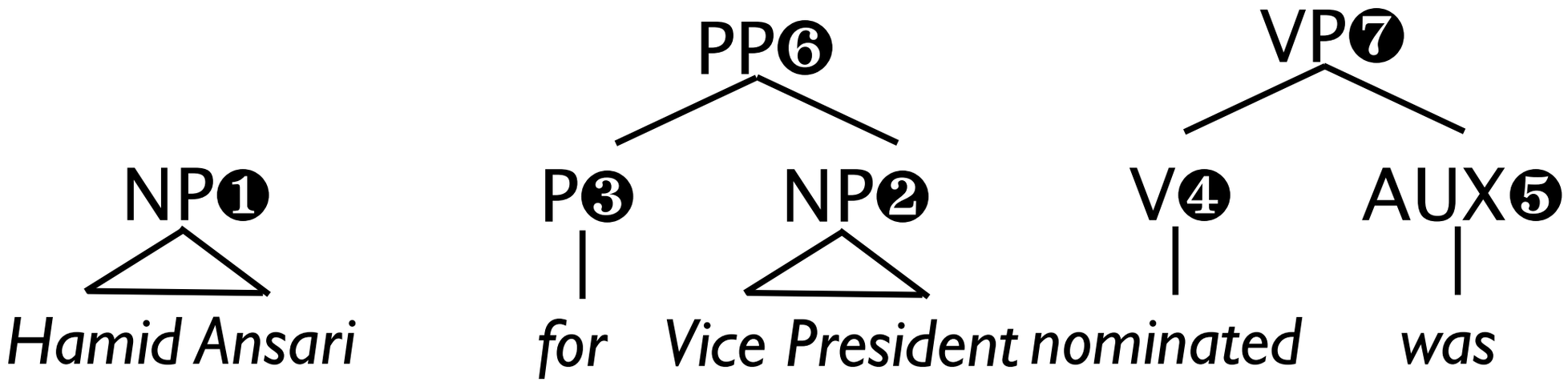

NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

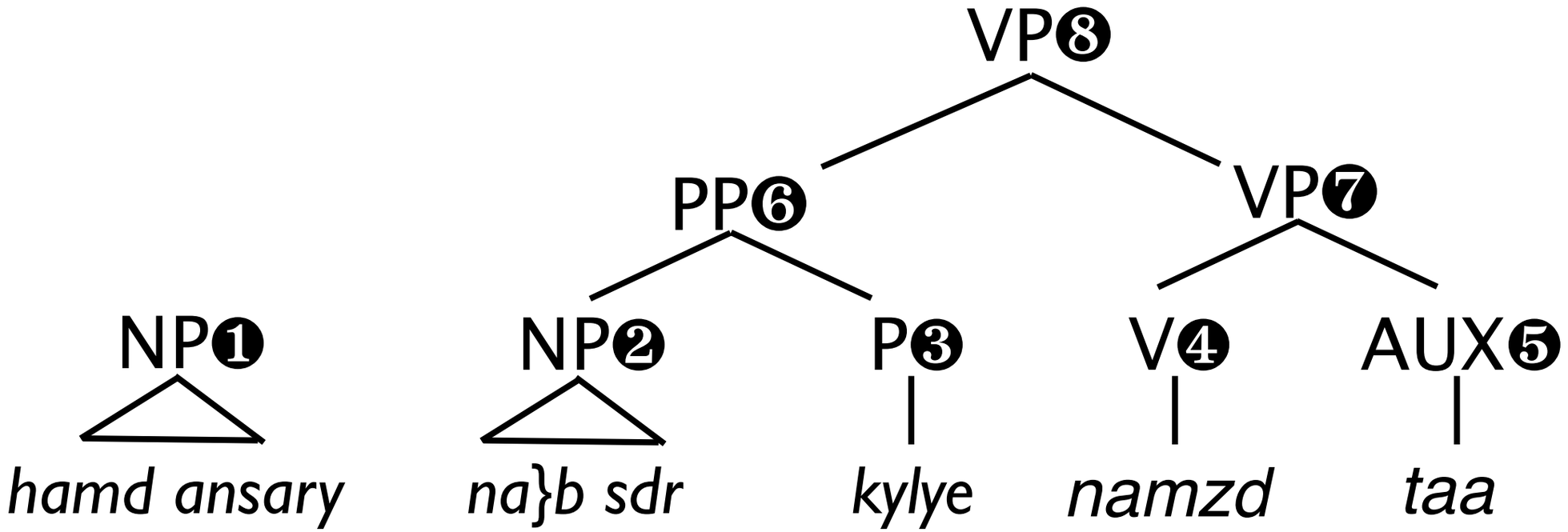

NP❶
NP❷
P❸
V❹
AUX❺
Hamid Ansari
Vice President
for
nominated
was

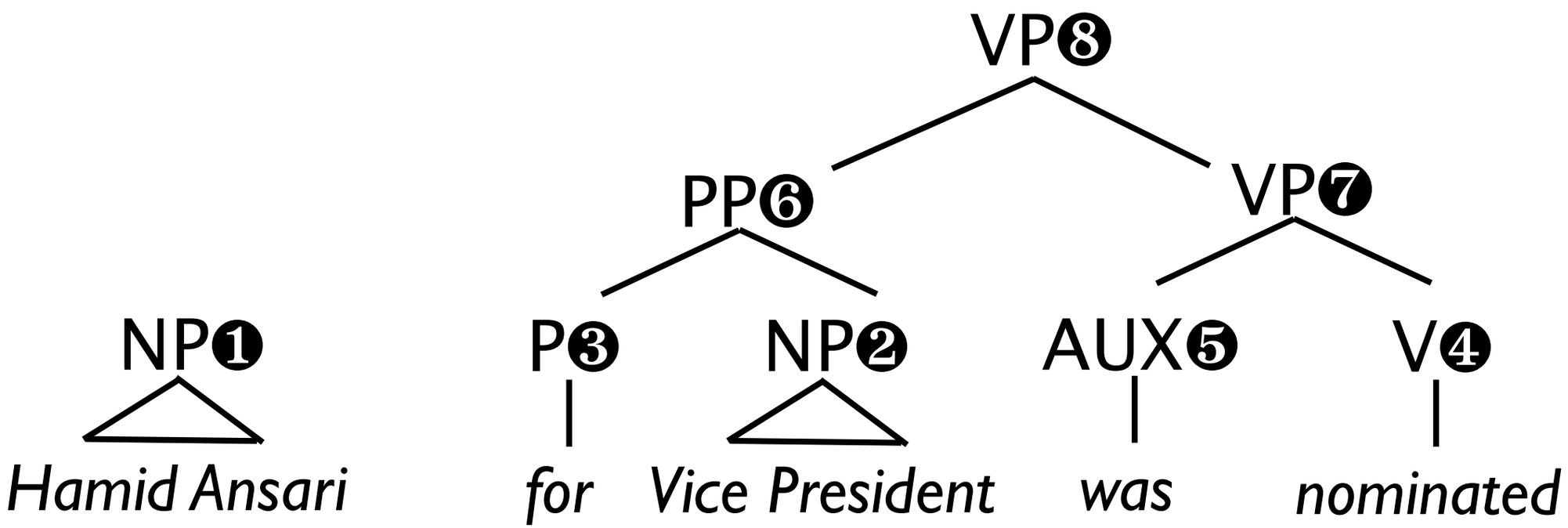

PP❻
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa

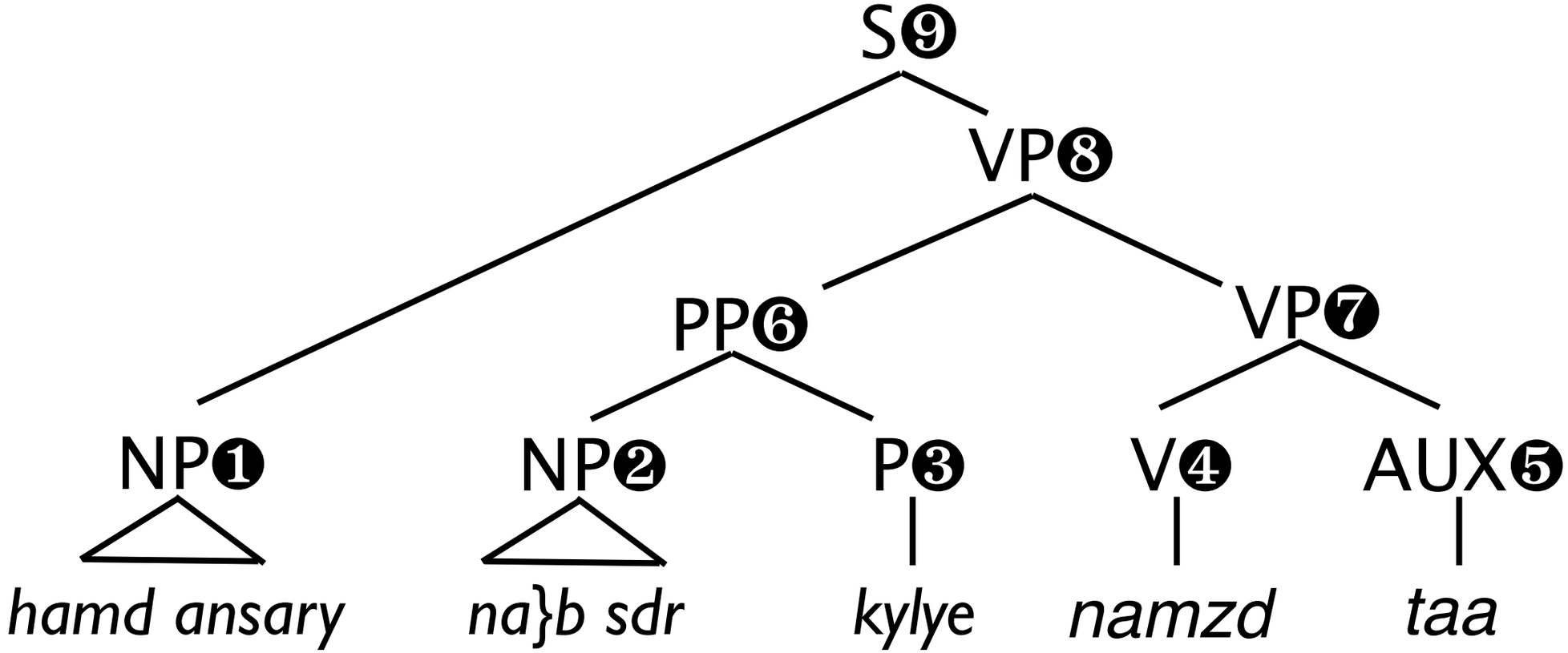

PP❻
NP❶
NP❷
P❸
V❹
AUX❺
Hamid Ansari
Vice President
for
nominated
was

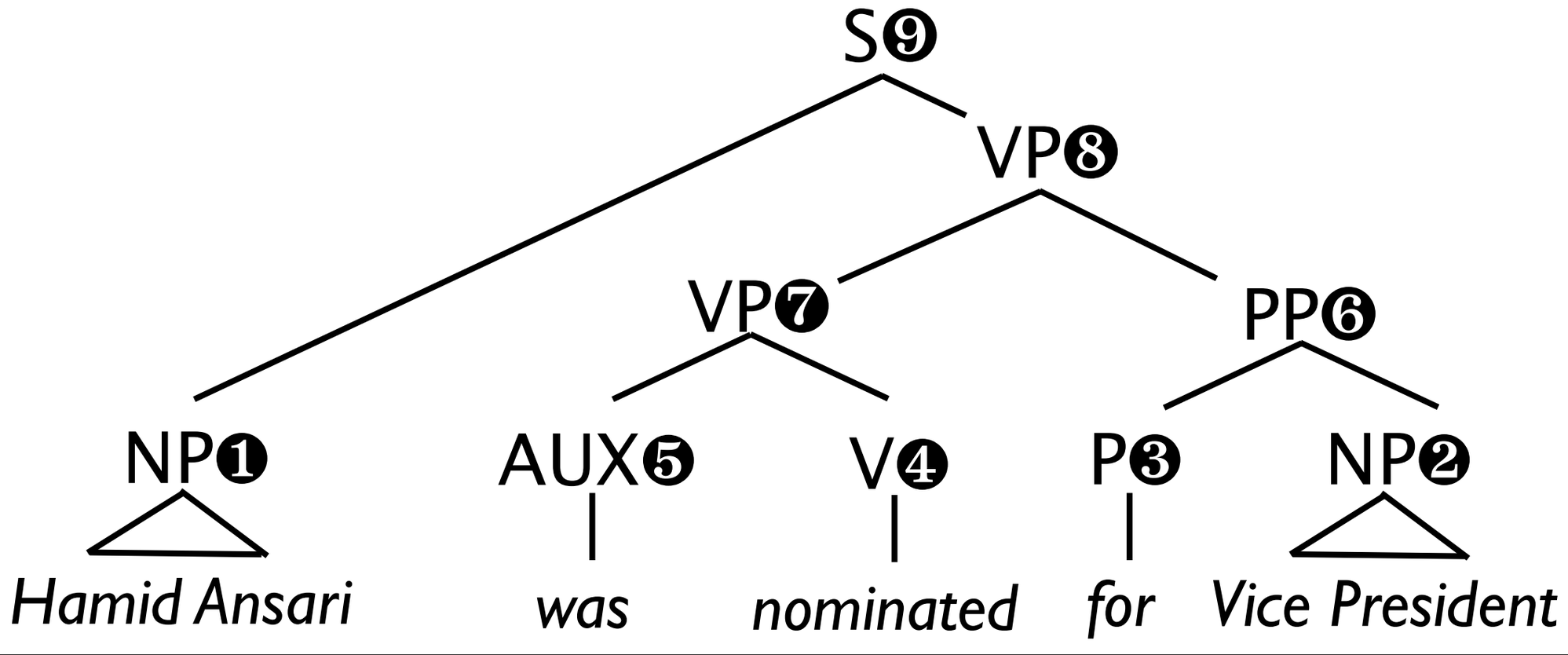

PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


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

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.

Example Research questions

- How can we learn translation rules automatically from data?
- When one word has several different possible translations, which one should we choose?
- When languages have different word orders, how do we properly re-reorder the words from one language into the other language?
- What is the most effective way of searching over all permutations and combinations of words?

Topics in CIS 526

- Probability and Language Models
- Word Alignment and Translation Models
- Searching for the most probable translation
- Phrase-based models and discriminative training
- Evaluating translation quality
- Syntax-based models of translation
- Collecting training data through crowdsourcing and web crawling

Who should take this class?

- Anyone who is interested in machine translation, natural language processing, or AI
- I don't presume any background in linguistics or statistics
- The only prerequisite is good programming skills
- All the code for the class is written in Python

Assignments

- Programming assignments are designed to teach you the fundamental algorithms in SMT and illustrate the research challenges

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 .

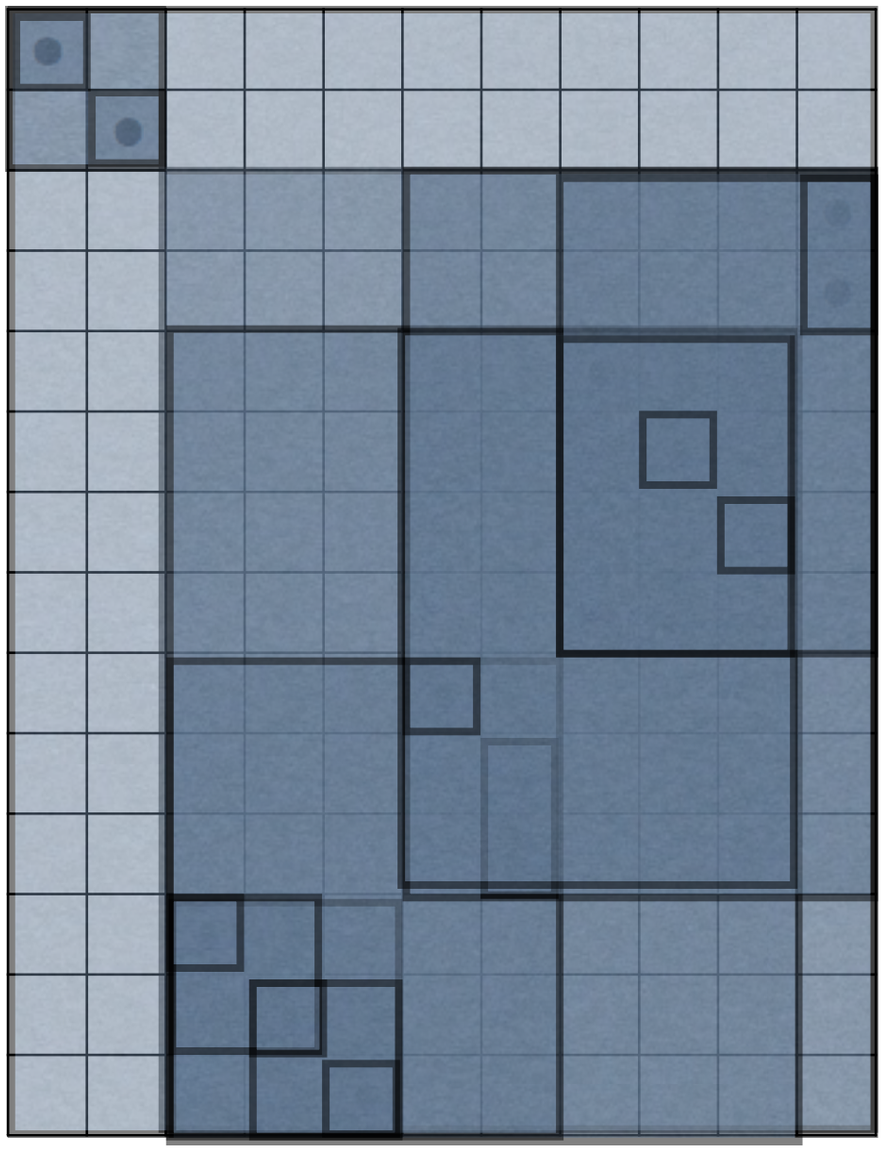
the small groups are not modern .

los grupos pequenos no son modernos .

Word aligner

Phrase Extractor

澳洲 是 与 北 韩 有 邦交 的 少数 国家 之一



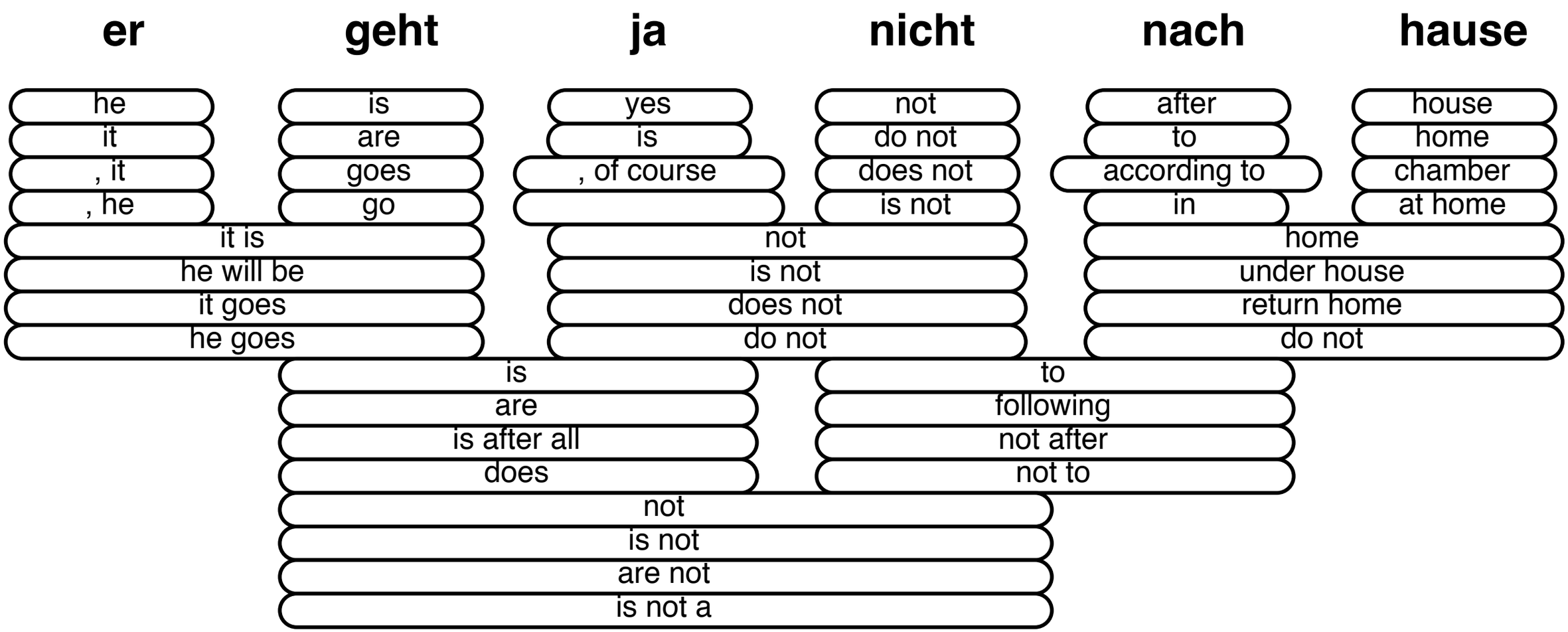
Australia is one of the few countries that have diplomatic relations with North Korea

北 韩
North Korea

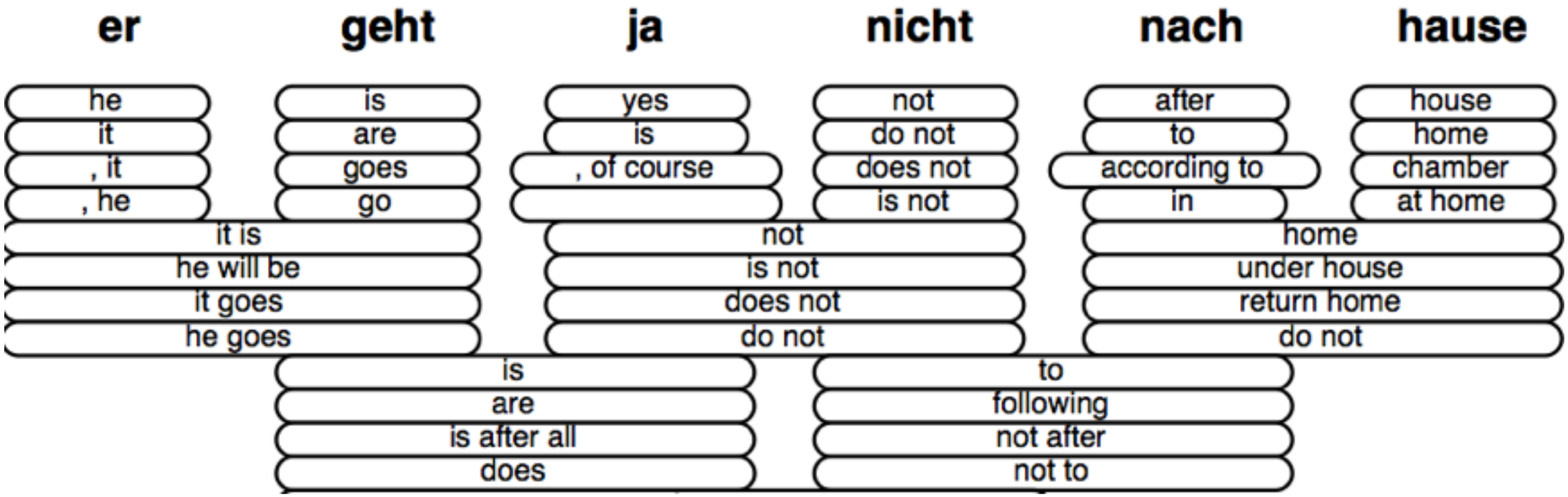
邦交
diplomatic relations

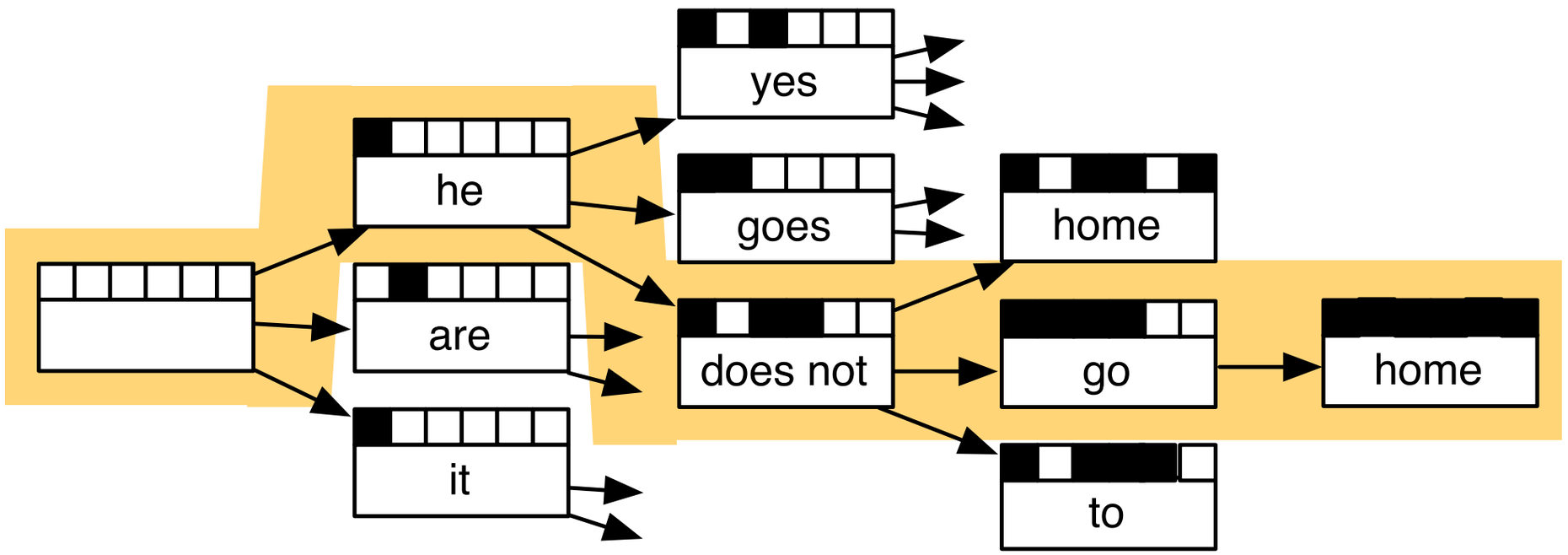
与 北 韩 有 邦交
have diplomatic relations with North Korea

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

Assignments

- All assignments have the following properties:
  - Clearly defined baseline systems that you can reimplement
  - Open-ended research problems with no "correct" solutions (lots of room for creativity)
  - Objective measures of how accurate a solution is

Leaderboard

This page contains the assignment leaderboard. The leaderboard is updated as follows. Every five minutes, o
assignment not yet past its due date is downloaded according the base URL and assignment number you turn
output is rescored if it changed.

			Assignments				
Rank		Handle	#0	#1 AER	#2 model score	#3 Spearman's	#4 BLEU
1	⬆	oracle	0	0	-1209.469789	0	100.00
2	⬇	SI	7	24.62	-1322.762025	0	27.39
	⬇	TangDou	5	30.40	-1219.224764	80.15	27.39
	⬇	subzero	100	28.76	-1238.241705	69.47	27.39
	⬇	madmaze	100	27.28	-1309.715268	52.63	27.39
	⬇	been	0	28.48	-1261.561138	61.20	27.39
	⬇	obzk	99	23.83	-1442.727775	0	27.39
8	⬇	1010	7	18.41	-1248.399735	83.46	27.08
9	⬇	YSL	0	30.46	-1	42.71	25.33
	⬇	far	79	30.46	-1442.727775	42.71	25.33
11	⬇	Lakie	99	21.94	-1266.738137	78.05	24.97
	⬇	Nihilist	3	21.94	-1317.024564	78.05	24.97
13	⬇	Shibboleth	4	29.89	-1287.430711	46.02	24.33
14	⬇	tgrhp	100	18.05	-1231.275486	81.80	24.14
15	⬇	default	0	65.77	-1442.727775	33.98	24.02
16	⬇	NathanStark	100	31.08	-1442.727775	67.82	-1

Legend

A value of -1 indicates that the as
contained invalid content.

The light coral line is the default
you run the code that was provide
with pale golden rod is the baseli
Typically, this line depicts the eff
project. Additional points will be
by a significant amount, and the r
placing at or near the top of the ra

The oracle system is a metric-awa
from each student submission. It
but it gives you an idea of how m
system to find.

Language in 10 minutes

- In-class presentation about a language
- What properties does it have?
- What makes it different than English?
- What are the challenges for machine translation?

- Jonny will give an example presentation

Straw Poll

- Last time I ran this class, the grading was:
  - 4 homework assignments (10 points each = 40 points)
  - Language in 10 minutes (10 points)
  - Quizzes about the reading (10 points total)
  - Self-designed final project (40 points)

- Poll question: Who would prefer more homework assignments instead of a final project?

By next week, please

- Buy the textbook (Kindle version is $35)
- Sign up on piazza.com/upenn/spring2014/cis526
- Fill out the piazza poll about when would be best for us to hold office hours.
- Do assignment 0, the setup assignment. I'll post a link to the assignment on piazza tomorrow.

Questions?

ccb@cis.upenn.edu