

Machine Translation

CIS 526

Instructor: Chris Callison-Burch

TA: Jonny Weese

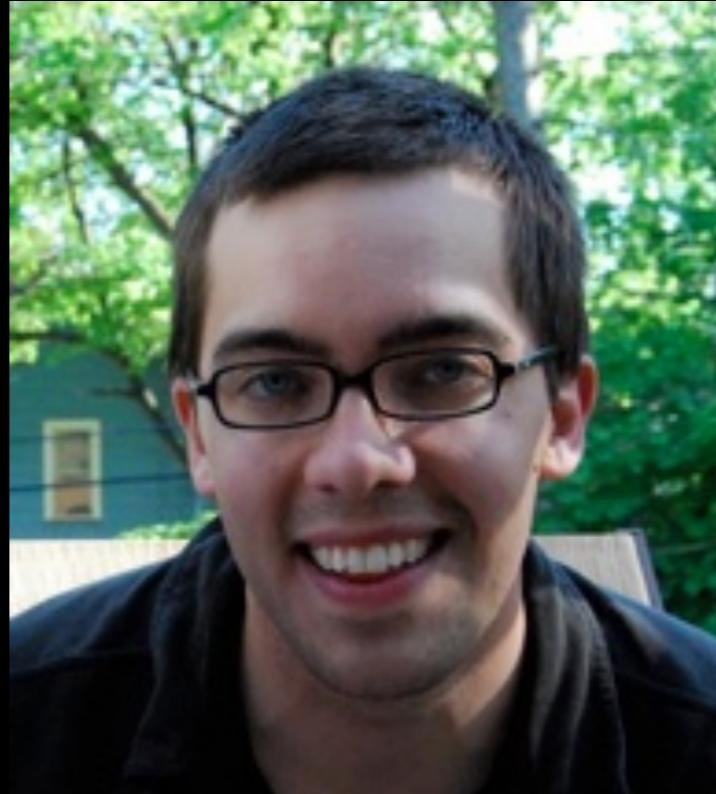
Course web site

mt-class.org/penn

Course materials developed with



Adam Lopez

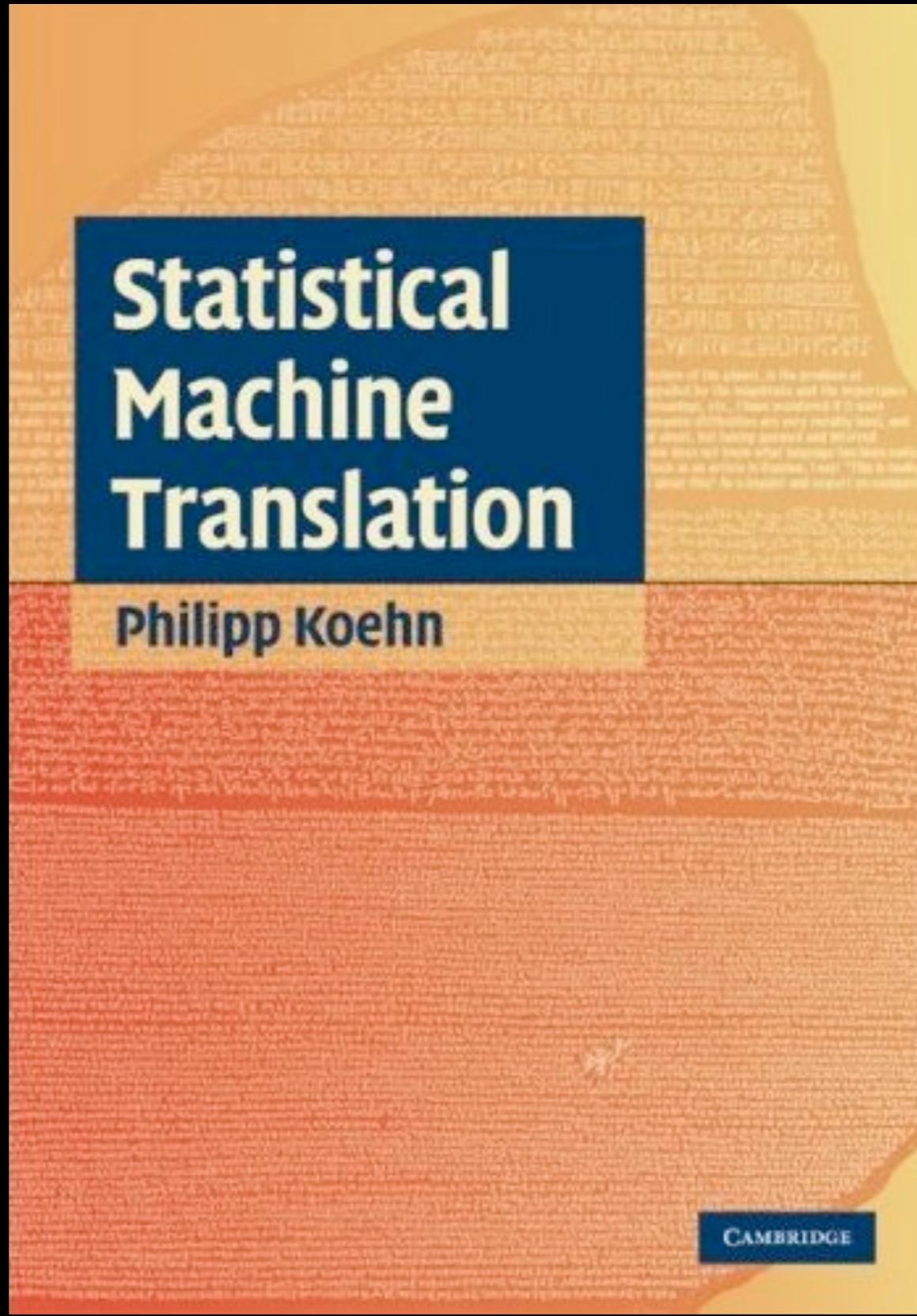


Matt Post



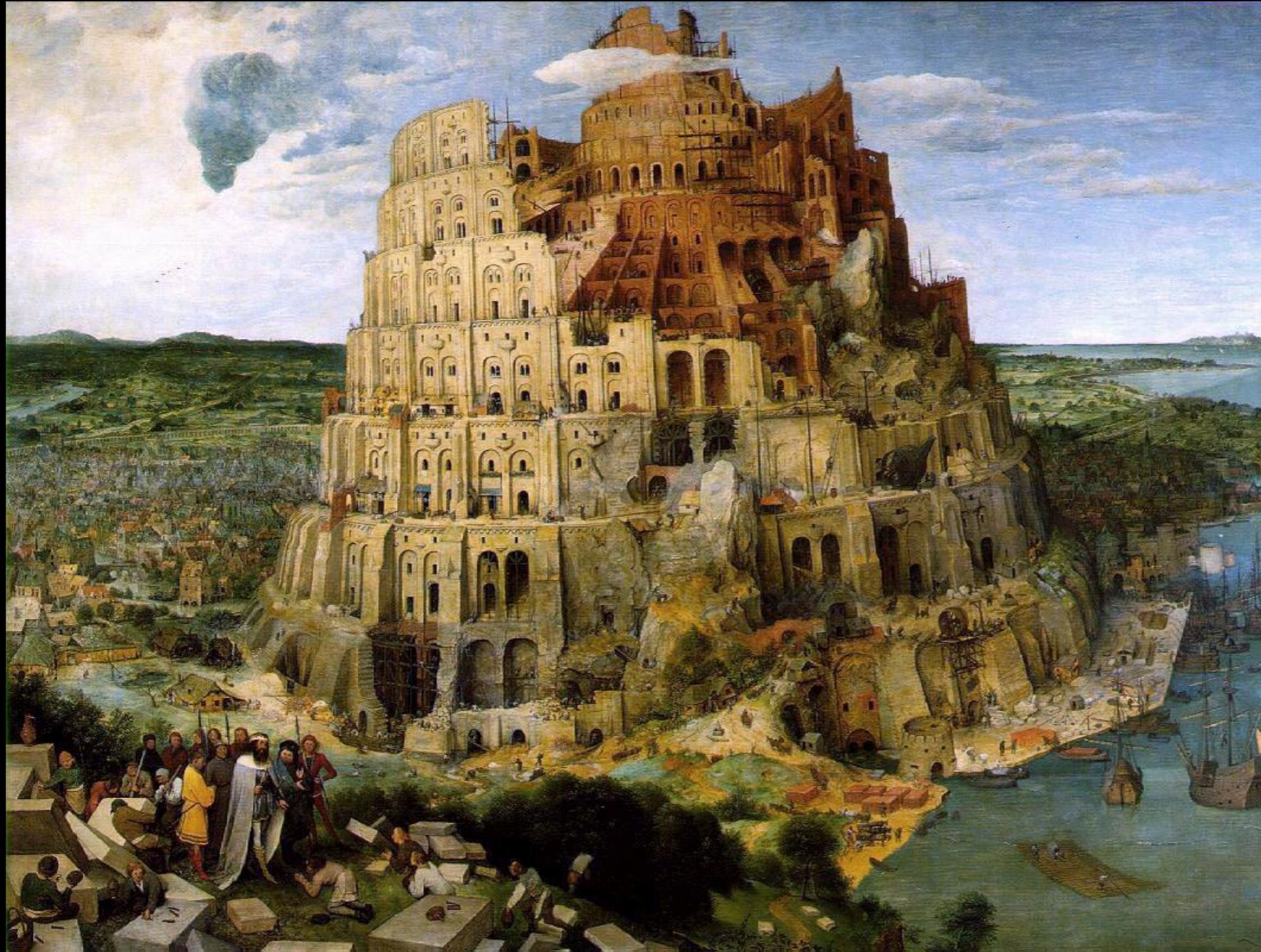
Chris Dyer

Textbook



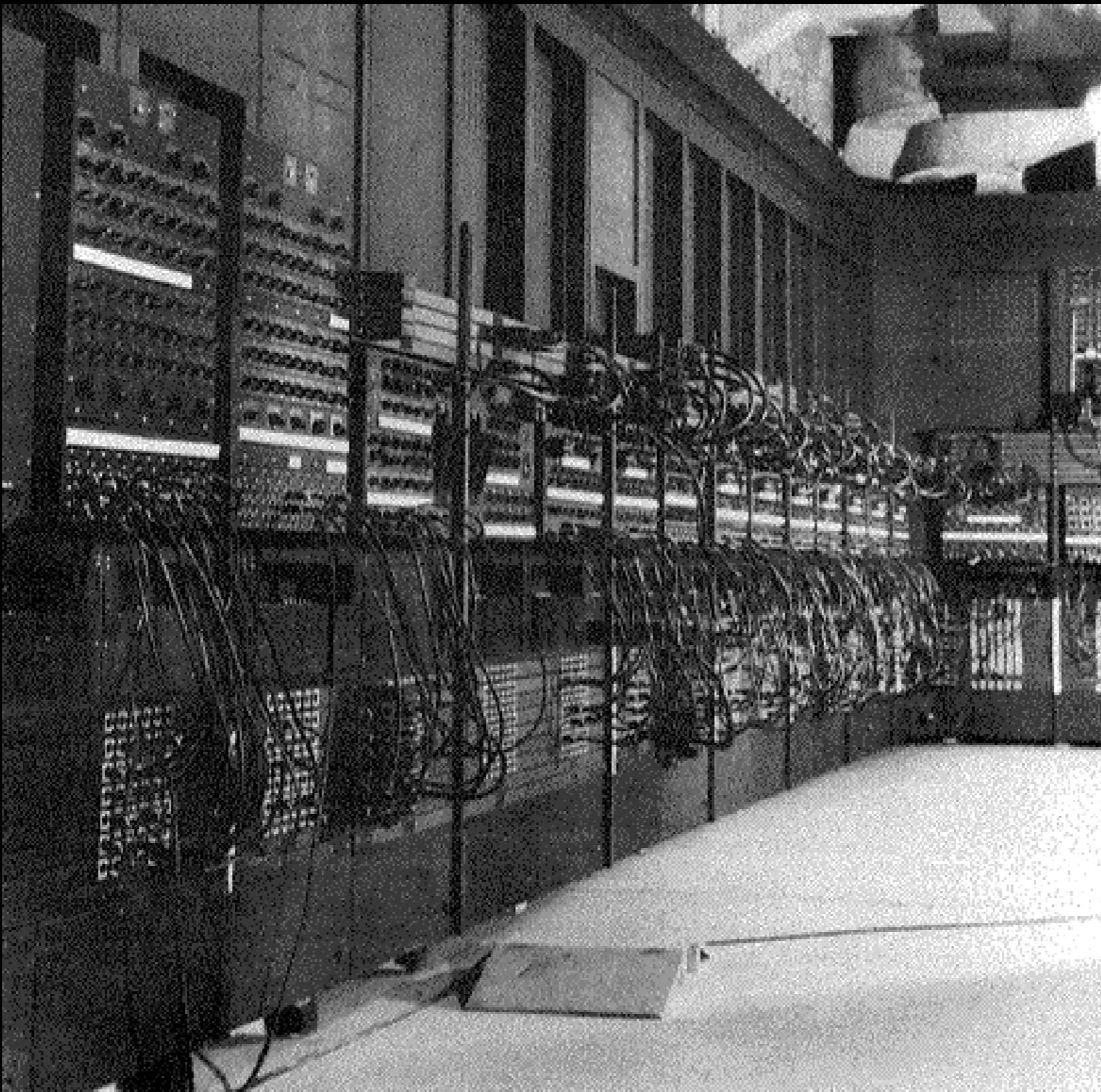
نائب امریکی صدر ڈک چینی کا کہنا ہے کہ میں اسامہ بن لادن کو زندہ یا مردہ دیکھنا چاہتا ہوں۔

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)

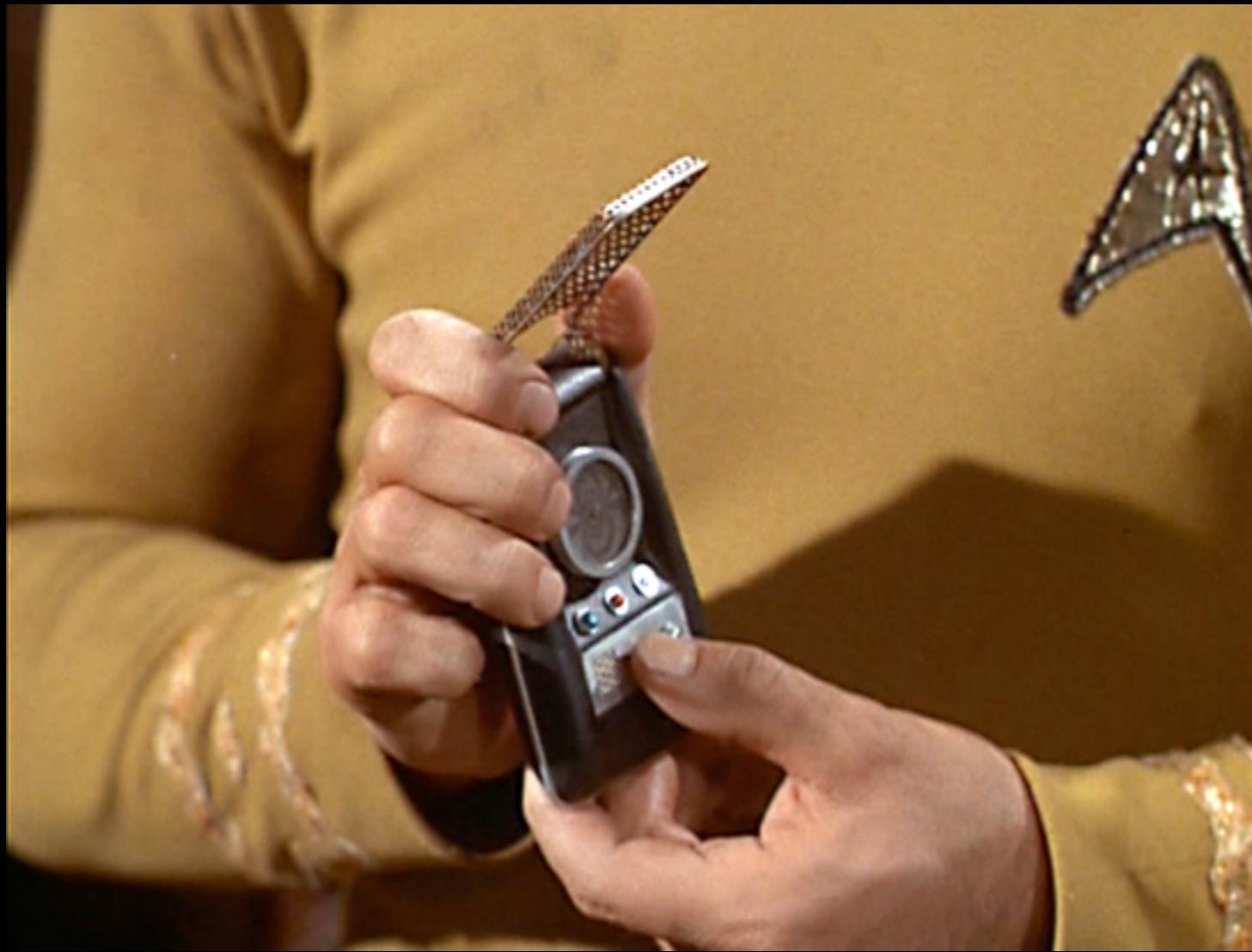


ENIAC (1946)

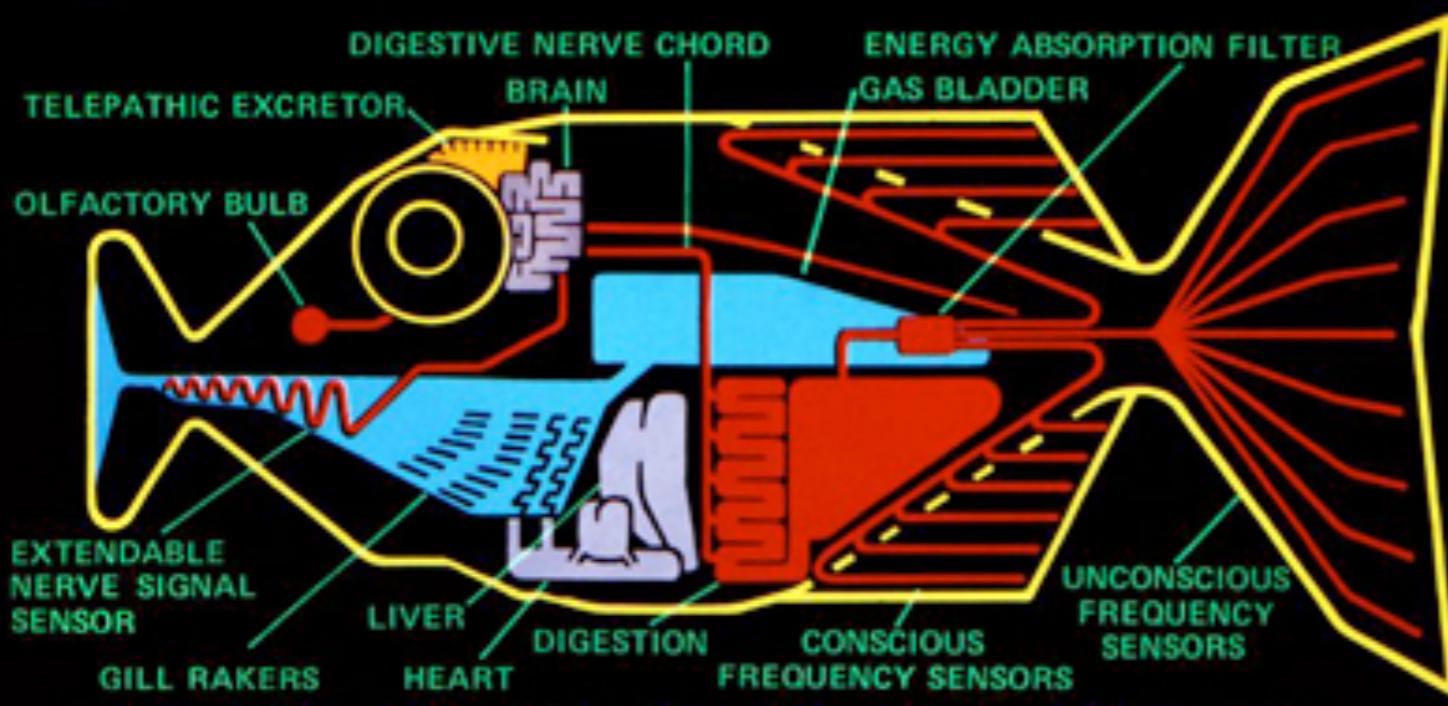


*When I look at an article
in Russian, I say: “This
is really written in
English, but it has been
coded in some strange
symbols. I will now
proceed to decode.”*

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!



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		



ΕΓΡΕΙ ΚΤΙΣΑΙΑ
ΟΔΑΤΕΙΣ ΕΙΔΑΙΛΛΟΙ ΠΕΡΕΙΣ ΠΛΥΝΕΣ ΙΔΑΠΑΝ ΤΗ ΛΑΝ
ΒΙΟΥ ΥΗΓΑΡΗ ΜΕΝΟΥ Υ ΠΟΤΟΥ Υ ΑΙΧΝΕΩΤΕΠΙΨΑ ΗΟΥΣΕ
ΠΟΒΙΟΧΗΤΑΠΗ ΜΕΝ· Ι ΠΟΤ· ΥΙ ΡΑΘΕΩΣ ΕΠΗΛΗΣΕΥΧ
ΗΗΕΑΙΤΟΥΒΑΙ ΙΑΕΙΔ ΝΤΛΕΣ ΕΝ ΟΥΛΠΑΝΤΑΙ ΣΥΠΛΙΧ
ΡΕΙΚΕΝΕΙΣΤΑΙ ΚΕΡΑΔΡΥΡΙΚΑΣ ΤΕΚΑΙΣΤΙΚΗΣ ΠΗΣ·
ΟΡΓΑΝΗΚΕ ΑΛΙΔΙΣΚΑΛΙ ΠΑΤΛΗΝΥ ΠΑΡΚ-ΥΙΛΗΝΕΙΔΙΓΥ Π
ΔΙΣΚΙΙΣΤΛΤΕΒΑΣΙΙΗΛΑΦΕΙΑ ΚΑΤΑ ΔΠΡΟΤΑΕ
ΝΤΛΣΕΚ ΠΟΔΛΟΥΧΡΟΝ ΟΥΔΑ ΠΕΛΥΙΚΤΛΗΝΗΚΕΑΝ
ΑΙΤΛΣΚΛΟΗΚΟΥΣΑΙΛΠΟΜ-ΠΛΙΣΤΙΟΙΣΟΕΙΣΑΠΟΥ
Ε ΚΛΙΠΕΡΙΤΛΝΙΕΡΕΛ Η ΠΟΛΕΜΗ ΕΜΠΑΙΟΝΙΟΙΔΛΙ
Α ΛΕΞΚΛΑΙΕΩΝ ΚΑΤΑ ΔΡΟΥ ΠΡΟΣ ΕΤΑΖΕΜΑΕΗΛΙΤ
ΑΤ ΕΒΓΑΛΕΔΕΙΜΜΗΝΔΠΛΗΤΛΕΝΤΟΙΣΠΡΤΒΡΗΙΧΙ
ΔΙΟΝΠΔΙΙΝ ΔΛΕΝΕ ΕΙ ΜΕΝΚΑΔΑ ΠΕΡΕΙΜΗΙΟ
ΦΑΡΗΗ ΚΑΙΡ ΟΙΚΑΤΕΛΘΡΩΝ ΤΑΙΜΕΝΕΙΝΕΠΙΤΛΗΝΙΔΙΩΝ
ΝΟΔΛΑΣΙΑ ΝΚΑΙ ΤΗ ΗΗΠΕΙΡΟΝ ΥΠΟΜΕΙΝΔΑ
ΗΤΔΙΒΟΧΕΙΡΙ ΤΗΗΗΗΝΚΑΤΕΙΛΙΝΜΕ ΝΗΚΑΙΛΧ
ΤΡΙΟΤΗΤΩΣ ΤΟΙΣΣΕ Π 12 (ΝΑΧΘΕΙΣΙΙΕΙΑ
ΣΙΚΑΙΤΕΙΧΕΣ ΗΙΑΥΤΗΝΛΞΙΟΔΟΓΡΙ ΣΙΕΡΙΕΙΛΒΕ
ΠΛΝΟΔΥΡΑΣΤΑΣΤΟΝ ΑΤΑΤΛΗΝ ΠΟ ΡΑΜΛΗ
ΤΟΛΙΝΚΑΤΛΚΡΑΤΟΣΕΙ ΛΕΝΙΚΑΙΤΟΣΙΕ ΝΑΥΤΗΛ
ΣΙΑΗΗΓΗΕΑΜΕΝΟΤΙ ΤΛΛΑΠΟΣΥΛΗΝΕΠΙΤ



www.un.org

http://www.un.org/english/

我们人民

National Bureau of Statistics

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)

Statistical Data

- Monthly
- Yearly
- Census
- Others

Related Links

- Chinese Version
- Others

Live and On-Demand Webcasts, 24 Hours a Day: Click on UN Webcast

联合国主页

http://www.un.org/chinese/

我们人民

中华人民共和国国家统计局

National Bureau of Statistics

105年9月18日 星期日

最新统计信息

- 2005年全年单位总产值比上年增长43.10% (09.16)
- 8月份“居民消费价格指数”为101.86 同比下降3.10点 (09.16)
- 1-6月固定资产投资同比增长17.64% 增幅回落3.41% (09.16)
- 1-6月甘肃固定资产投资增长17.64% 增幅回落3.41% (09.16)
- 经济全球化对江苏省国民经济产生重大影响 (09.16)
- 统计数据：8月份工业产品产量 各地区产品销售率 (09.15)
- 统计数据：8月份工业增加值 各地区工业增加值 (09.15)
- 1-6月份全国城镇固定资产投资同比增长27.4% (09.15)
- 加快云南人口城市化进程需解决四大关键问题 (09.15)
- 丹江口：通过教育系统改善“最初深入人心” (09.15)
- 1-8月浙江限额以上固定资产投资同比增长16.4% (09.15)
- 8月份广西消费品零售额与去年同期相比增长13.8% (09.15)
- 8月份我国工业增加值加薪59.68亿 同比增长16% (09.14)
- 调查显示：广东省企业流动资金短缺问题日益突出 (09.14)
- 实施品牌战略 推动吉林省经济社会可持续发展 (09.14)
- 无极：城乡居民收入剪刀差十年扩大0.46倍 (09.14)
- 8月份甘肃省工业总产值同比增长4.4% 增速略有回落 (09.14)

最新统计动态

- 河南省县强农促增千项统计知识培训班举行 (09.16)
- 白城市副市长火线启动统计工作者“春播” (09.16)
- 吉林磐石市统计机构建设加强 或立实体统计工作站 (09.16)
- 《新疆三十年》出版发行 自治区将为该书题词 (09.16)
- 国家统计局局长李德水深感心切：波澜壮阔话边疆 (09.15)
- 湖南省副省长蒋伟烈调研统计工作并要求提 (09.15)
- 山东省统计局实行行政许可办理窗口式办公 (09.15)

重要公告

- 公告
- 关于申报2005年度全国统计科研计划项目的通知
- 印发《关于统计上对公有和非公有控股经济的分类办法》的通知

统计机构

统计动态

支撑统计工作 公务软件下载

统计标准

统计制度

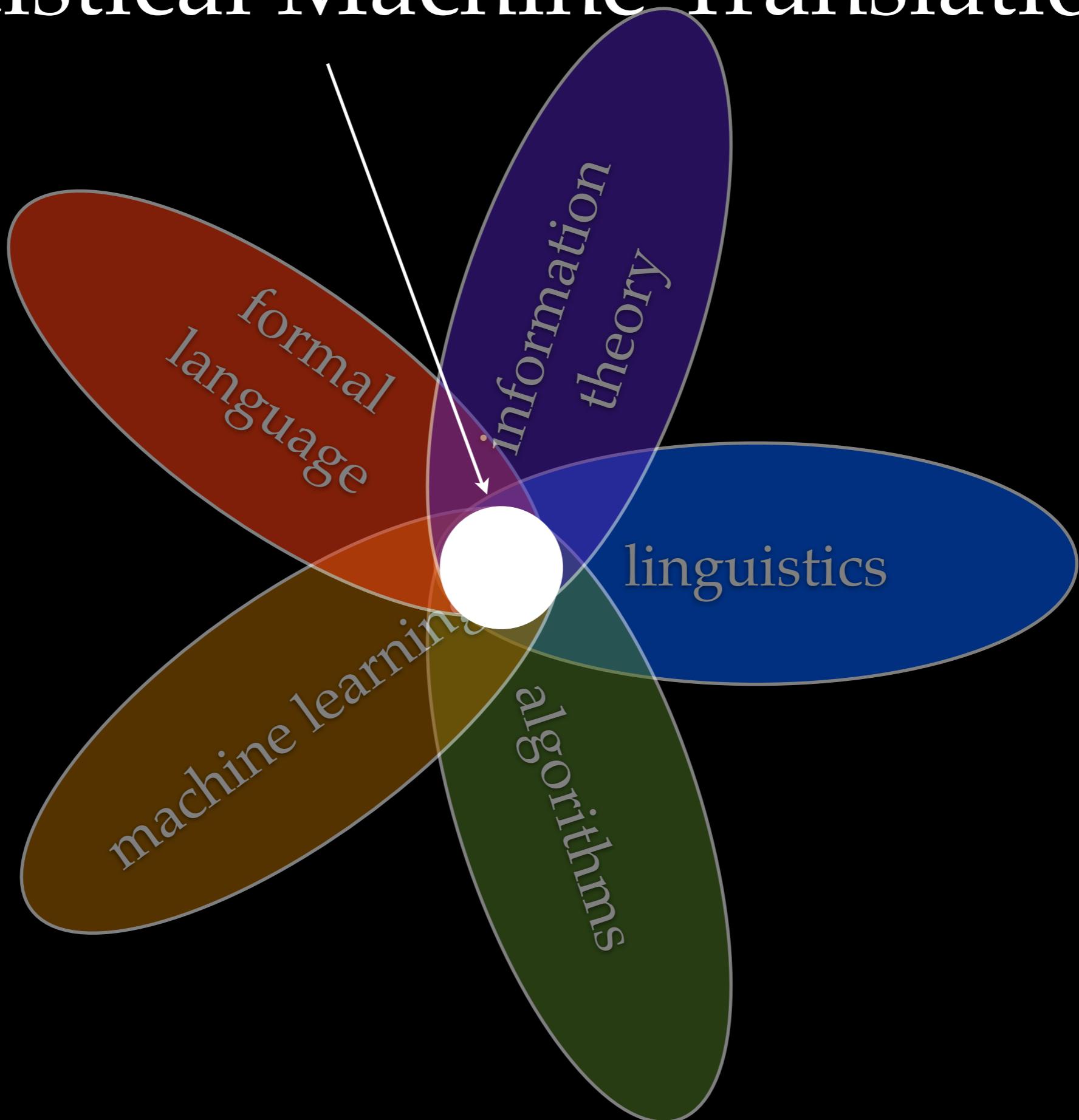
统计知识

联合国网络直播

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
$S \rightarrow$	$NP\textcircled{1} VP\textcircled{2}$	$NP\textcircled{1} VP\textcircled{2}$
$VP \rightarrow$	$PP\textcircled{1} VP\textcircled{2}$	$VP\textcircled{2} PP\textcircled{1}$
$VP \rightarrow$	$V\textcircled{1} AUX\textcircled{2}$	$AUX\textcircled{2} V\textcircled{1}$
$PP \rightarrow$	$NP\textcircled{1} P\textcircled{2}$	$P\textcircled{2} NP\textcircled{1}$
$NP \rightarrow$	<i>hamd ansary</i>	<i>Hamid Ansari</i>
$NP \rightarrow$	<i>na}b sdr</i>	<i>Vice President</i>
$V \rightarrow$	<i>namzd</i>	<i>nominated</i>
$P \rightarrow$	<i>kylye</i>	<i>for</i>
$AUX \rightarrow$	<i>taa</i>	<i>was</i>

NP1
hamd ansary

NP2
na}b sdr

P3
kylye

V4
namzd

AUX5
taa

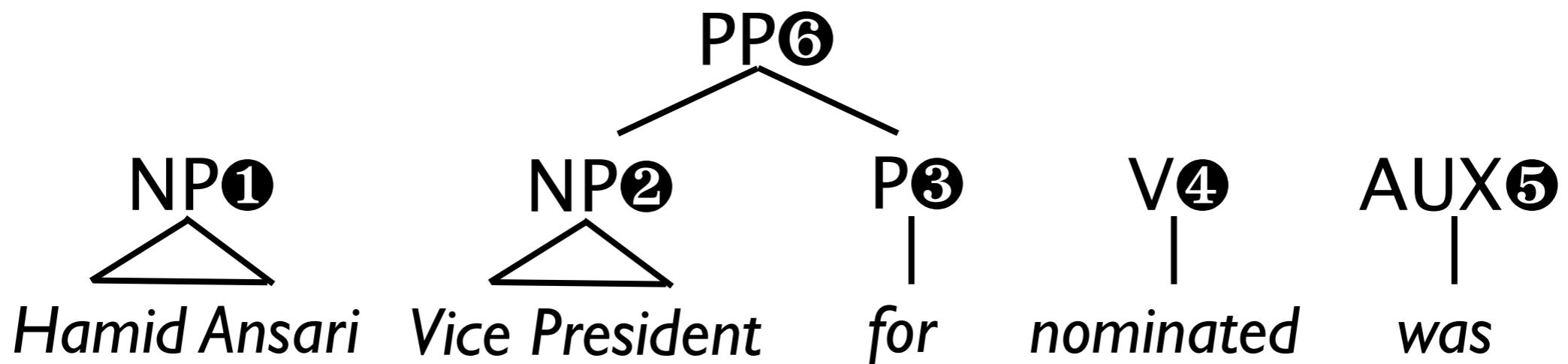
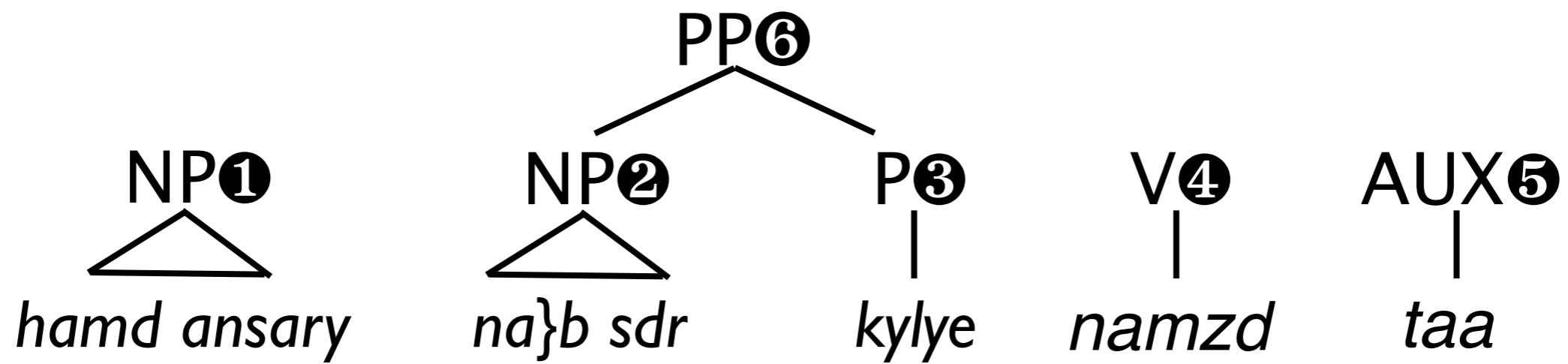
NP1
Hamid Ansari

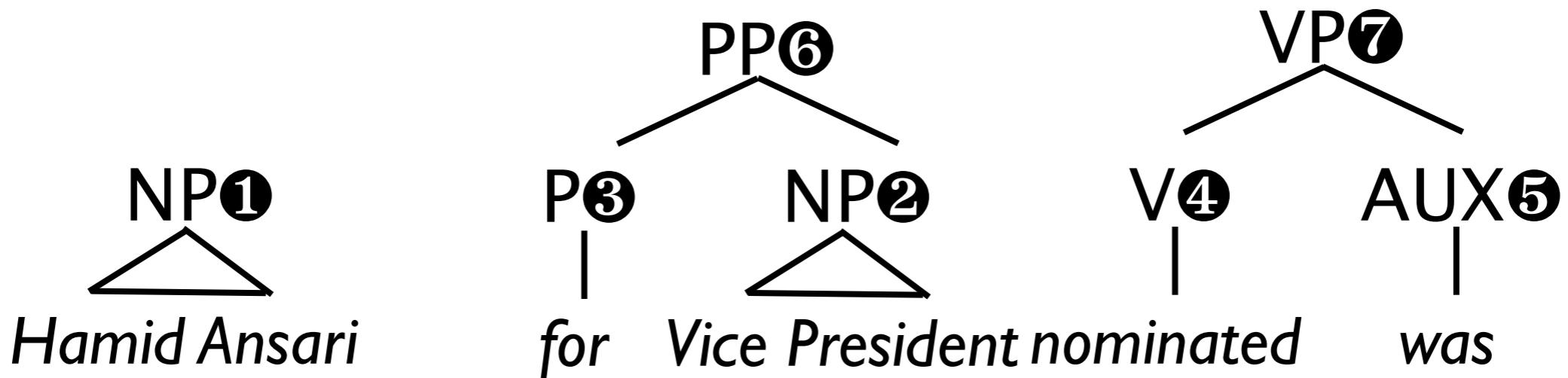
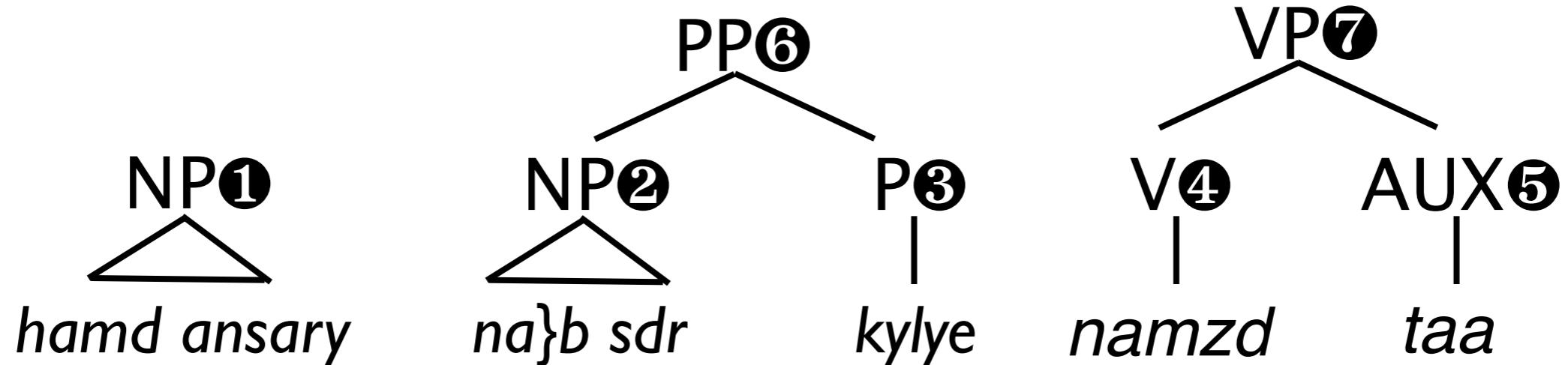
NP2
Vice President

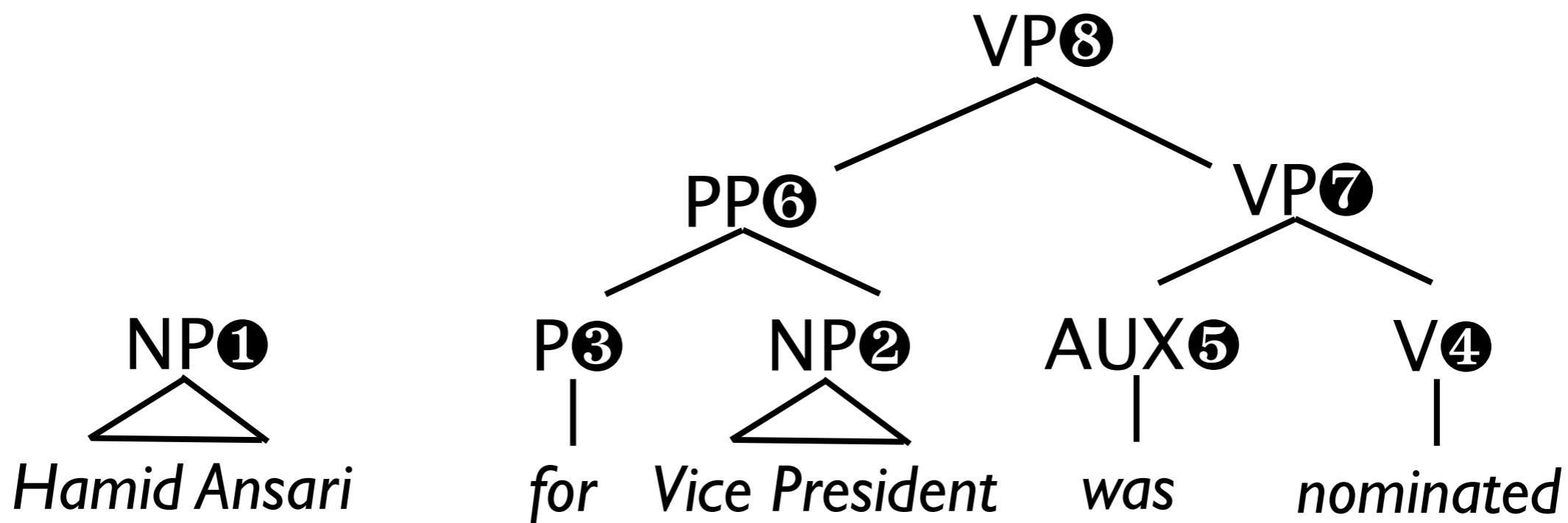
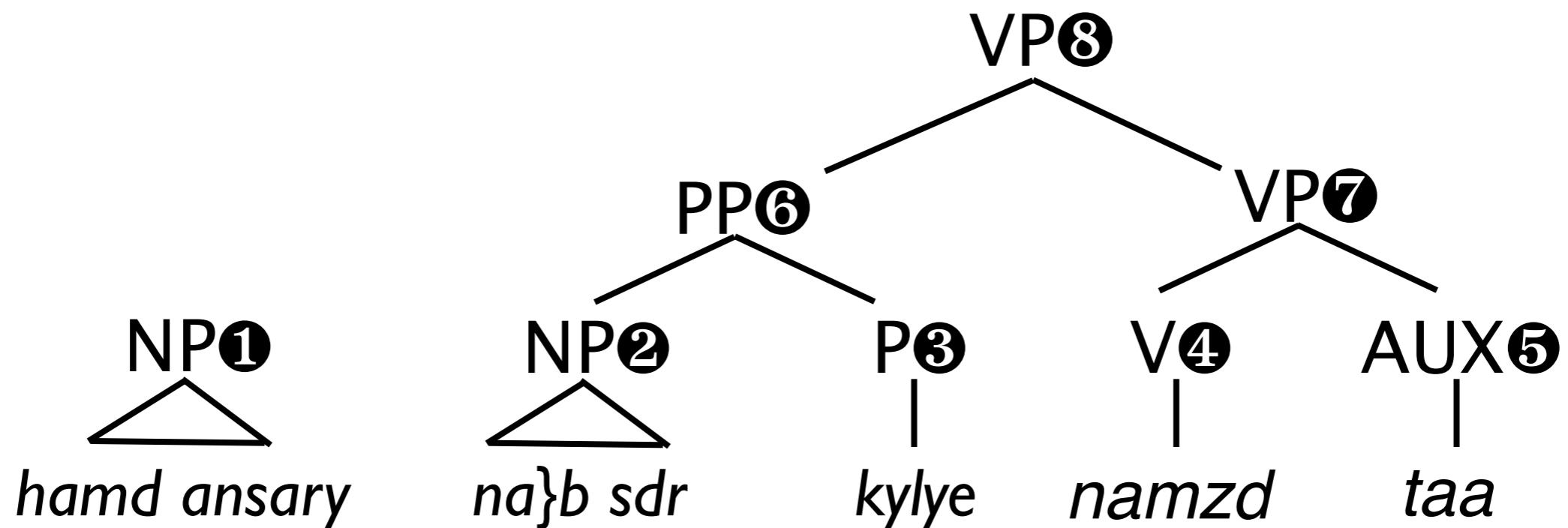
P3
for

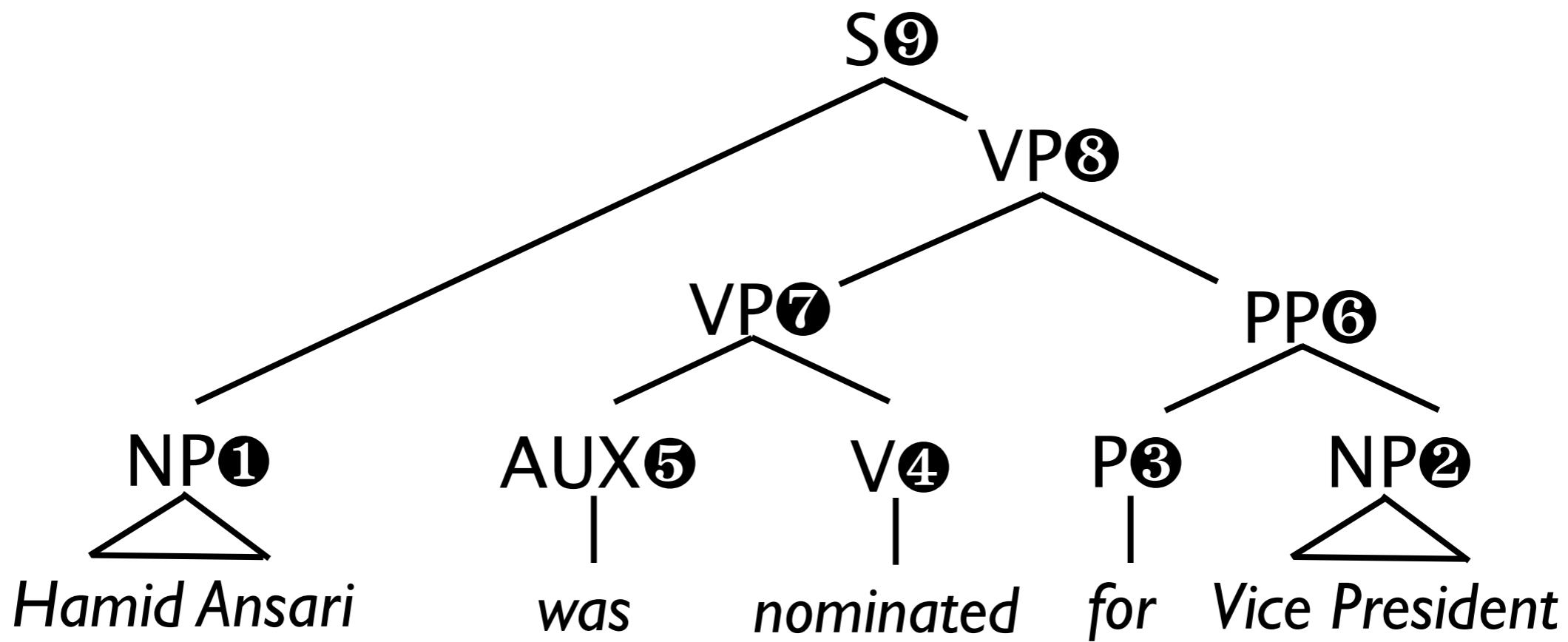
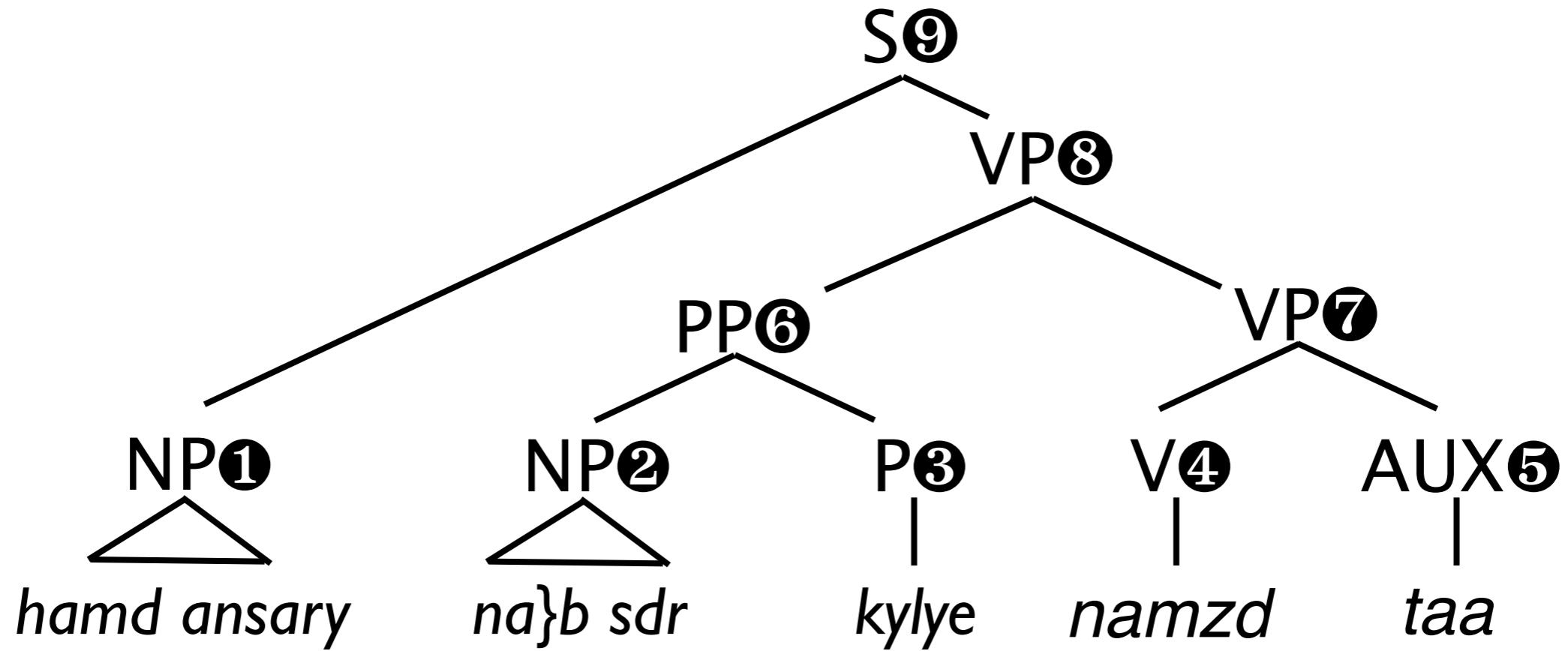
V4
nominated

AUX5
was









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 انجینئرنگس 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 چرائے 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 .

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 .

los asociados tambien estan enfadados .

Word aligner

Garcia and associates .

— 1 —

Garcia y asociados .

Carlos Garcia has three associates .

/ \ | | | /

Carlos Garcia tiene tres asociados .

his associates are not strong.

A horizontal row of five white line segments on a dark background. The first segment is a short vertical line on the left. The second segment is a diagonal line sloping down from left to right. The third segment is a long diagonal line sloping down from left to right, intersecting the second segment. The fourth segment is a short vertical line on the right. The fifth segment is a diagonal line sloping up from left to right.

sus asociados no son fuertes .

Garcia has a company also .

A horizontal black line with several white diagonal lines crossing it from the right side.

Garcia tambien tiene una empresa .

its clients are angry .

A horizontal row of four white L-shaped brackets on a black background. The first two brackets point upwards and to the left, while the last two point downwards and to the right.

sus clientes 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.

O I I I

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

The image consists of two sets of white lines on a dark gray background. The left set of lines forms an 'X' shape, with one line sloping upwards from left to right and the other sloping downwards. The right set of lines also forms an 'X' shape, but it is rotated 45 degrees clockwise relative to the left one. Both sets of lines are thin and white.

los grupos modernos venden medicinas fuertes

the groups do not sell zanzanine .

| / /

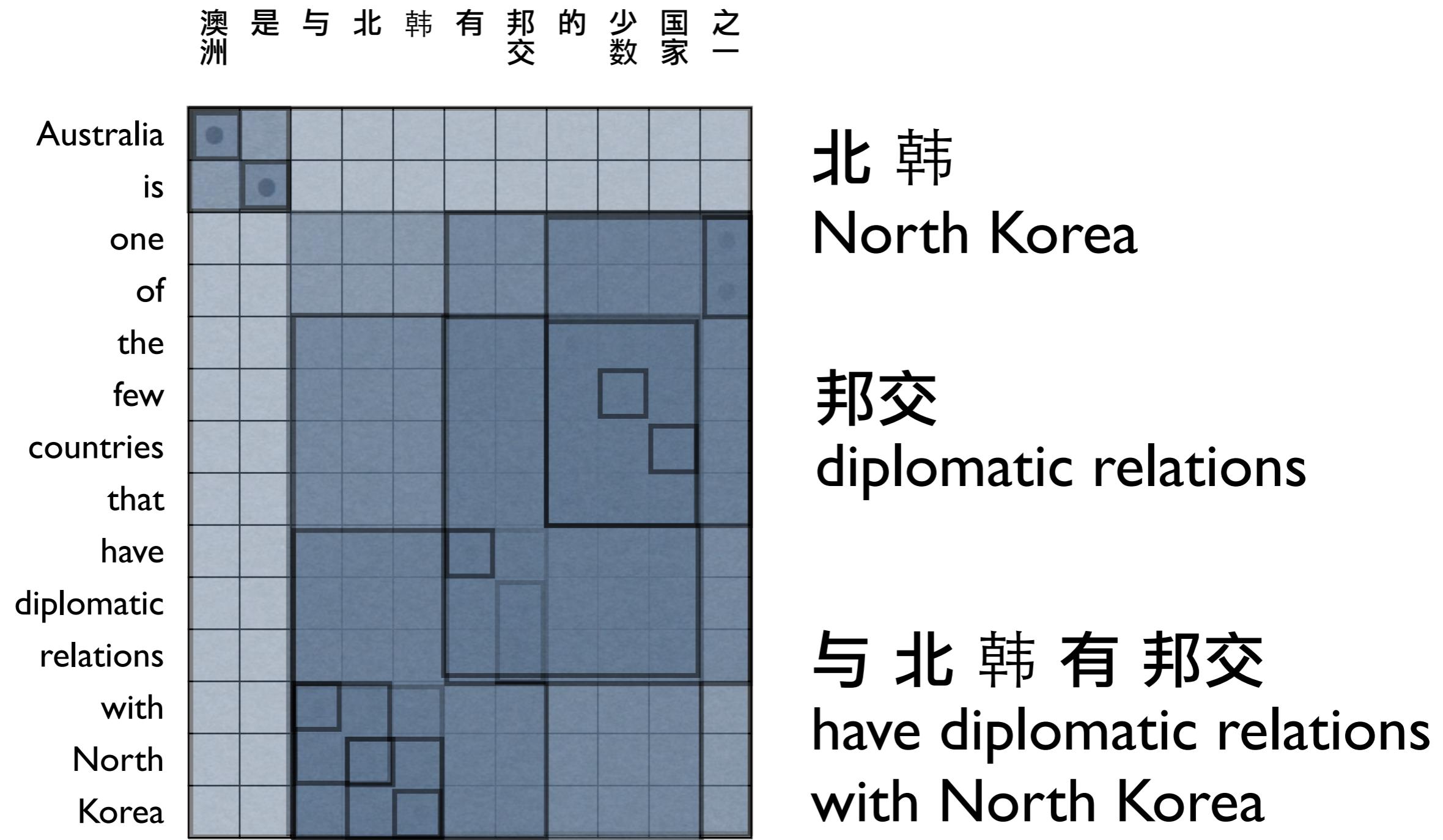
los grupos no venden zan zanina .

the small groups are not modern .

/ X X \

los grupos pequeños no son modernos .

Phrase Extractor



Phrase-based Decoder

er

he
it
, it
, he

it is
he will be
it goes
he goes

geht

is
are
goes
go

is
are
goes
go

ja

yes
is
, of course

not
is not
does not
do not

nicht

not
do not
does not
is not

to
following
not after
not to

nach

after
to
according to
in

home
under house
return home
do not

hause

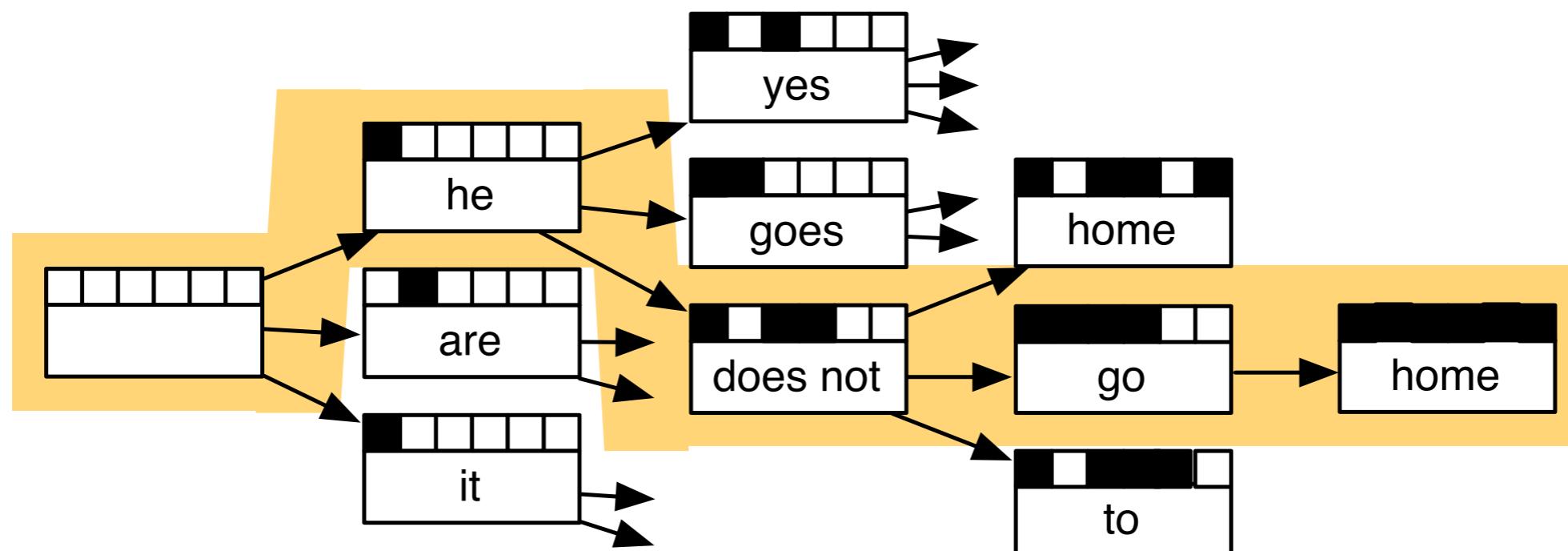
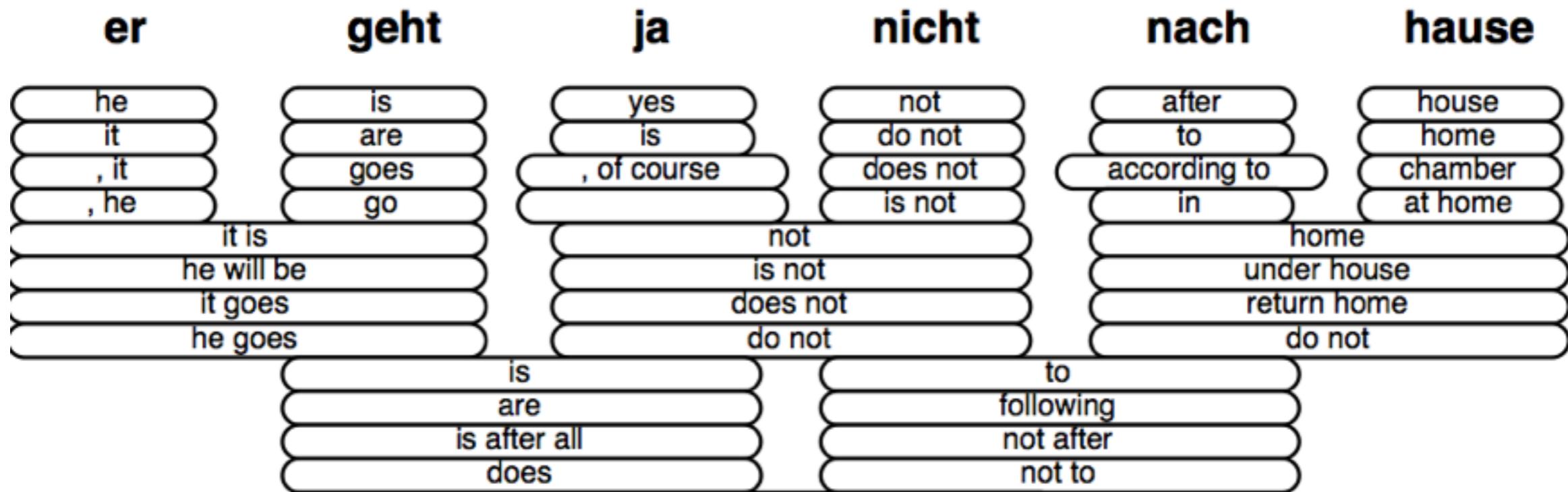
house
home
chamber
at home

home
under house
return home
do not

is
are
is after all
does

not
is not
are not
is not a

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, one assignment not yet past its due date is downloaded according the base URL and assignment number you turn in. The output is rescored if it changed.

Rank	Handle	Assignments				
		#0	#1	#2	#3	#4
AER	model score	Spearman's	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 assignment contained invalid content.

The light coral line is the default line. You run the code that was provided with pale golden rod is the baseline. Typically, this line depicts the effort of the project. Additional points will be given by a significant amount, and the reward for placing at or near the top of the ranking.

The oracle system is a metric-aware system from each student submission. It is slow but it gives you an idea of how much work your 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