

# Machine Translation I

Wei Xu

(many slides from Greg Durrett)

# Semester So-far

---

- ▶ Machine Learning Models
  - Linear models: Naive Bayes, Logistic Regression, SVM, Perceptron
  - Neural models: FeedForward Neural Networks, Back-prop, ...
- ▶ Sequence Models (NER, POS tagging, etc)
  - Hidden Markov Model, Viterbi Algorithm, Conditional Random Fields
- ▶ Word Embeddings
- ▶ Recurrent NN, Convolutional NN, Neural CRF

# Rest of the Semester

---

- ▶ Applications in Natural Language Processing
  - Machine Translation (2 weeks)
  - Information Extraction
  - Reading Comprehension
  - Automatic Summarization (if time)
  - Dialog System
  - Contextual Word Embeddings
  - etc.

# This Lecture

---

- ▶ MT and evaluation
- ▶ Word alignment
- ▶ Language models
- ▶ Phrase-based decoders
- ▶ Syntax-based decoders

# MT Basics

# MT Basics



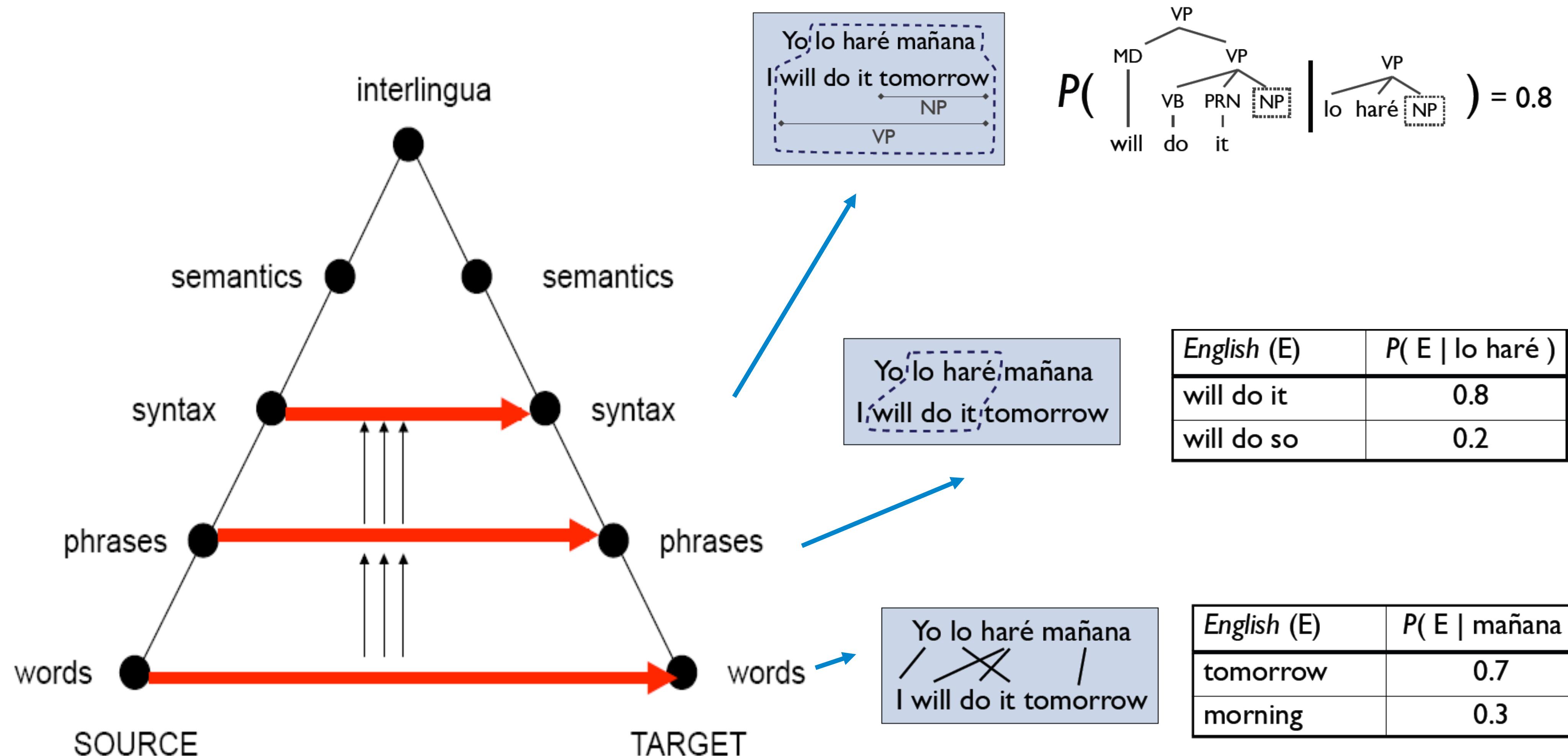
Trump Pope family watch a hundred years a year in the White House balcony

# MT Ideally

---

- ▶ I have a friend =>  $\exists x \text{ friend}(x, \text{self})$  => J'ai un ami  
J'ai une amie
- ▶ May need information you didn't think about in your representation
- ▶ Hard for semantic representations to cover everything
- ▶ Everyone has a friend =>  $\exists x \forall y \text{ friend}(x, y)$  => Tous a un ami  
 $\forall x \exists y \text{ friend}(x, y)$
- ▶ Can often get away without doing all disambiguation — same ambiguities may exist in both languages

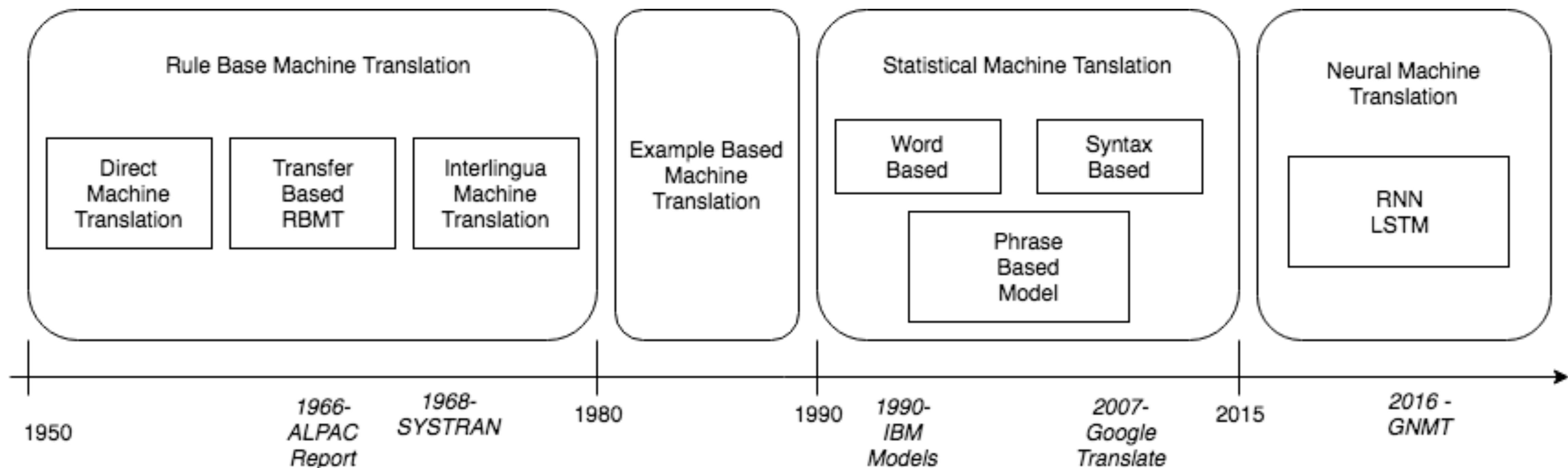
# Levels of Transfer: Vauquois Triangle



- Today: mostly phrase-based, some syntax

Slide credit: Dan Klein

# History of MT



# Parallel Training Corpus

	facing with the swelling flow of through traffic zooming past their doors .		reclama de inconvenientes que mas y mas gente tiene que soportar por el tráfico que pasa por delante_de sus casas , que aumenta a_diario .
5	#77501757 Weekend traffic bans and traffic <b>jams</b> are a curse to road transport .	#74765580	Las prohibiciones de conducir los fines de semana y los <b>embotellamientos</b> asolan el transporte por carretera .
6	#79500725 Some people also want to recoup the cost of traffic <b>jams</b> from those who get stuck in them , according to the ' polluter pays ' principle .	#76764676	Algunos son partidarios de que incluso los costes ocasionados por los <b>atascos</b> se carguen a el ciudadano que se encuentra atrapado en ellos , de conformidad con el principio de que " quien contamina paga " .
7	#79500765 I think this is an excellent principle and I would like to see it applied in full , but not to traffic <b>jams</b> .	#76764713	Me parece un principio acertado y estoy dispuesta a aplicarlo íntegramente , pero no sobre los <b>atascos</b> , ya_que éstos son un claro indicio de el fracaso de la política gubernamental en_materia_de infraestructuras .
8	#79500768 Traffic <b>jams</b> are indicative of failed government policy on the infrastructure front , which is why the government itself , certainly in the Netherlands , must be regarded as the polluter .	#76764747	Por eso es preciso subrayar que en estos casos quien contamina es el propio Gobierno , a el menos en los Países_Bajos .
9	#81309716 This would increase traffic <b>jams</b> , weaken road safety and increase costs .	#78586130	Esto aumentaría los <b>atascos</b> , mermaría la seguridad vial e incrementaría los costes .
10	#81997391 In the previous legislature , Parliament gave its opinion on the Commission ' s proposals on the simplification of vertical directives on sugar , honey , fruit juices , milk and <b>jams</b> .	#79281114	En efecto , durante la precedente legislatura , el Parlamento se manifestó sobre las propuestas de la Comisión relativas a la simplificación de directivas verticales sobre el azúcar , la miel , los <b>zumos de frutas</b> , la leche y las <b>confituras</b> .
11	#81998167 For <b>jams</b> , I personally reintroduced an amendment that was not accepted by the Committee on the Environment , Public Health and Consumer Policy , but which I hold to .	#79281936	Para las <b>confituras</b> , yo personalmente volví a introducir una enmienda que no fue aceptada por la Comisión_de_Medio_Ambiente , Salud_Pública y Política_de_el_Consumidor , pero que es importante para mí .
12	#81998209 It concerns not accepting the general use of a chemical flavouring in <b>jams</b> and marmalades , that is vanillin .	#79281966	Se trata de no aceptar la utilización generalizada de un aroma químico en las <b>confituras</b> y " marmalades " , a saber , la vainillina .
13	#82800065 This is highlighted particularly in towns where it is necessary to find ways of solving environmental problems and the difficulties caused by traffic <b>jams</b> .	#80085988	Esto se pone_de_relieve aún más en las ciudades , en las que hay que encontrar medios para eliminar los inconvenientes derivados de los problemas medioambientales y de la congestión de el tráfico .

# Phrase-Based MT

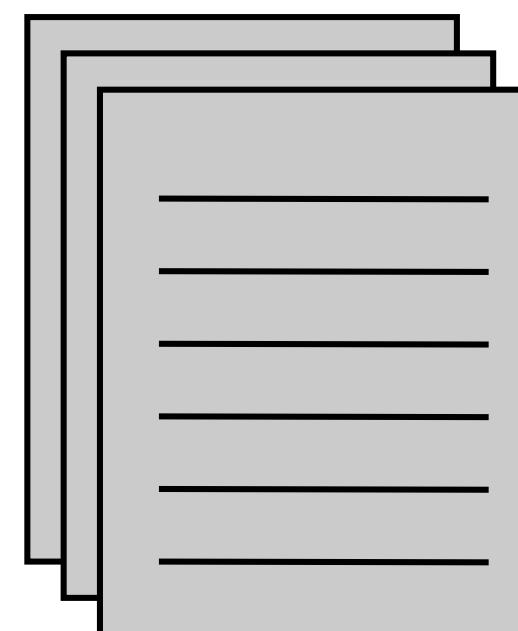
---

- ▶ Key idea: translation works better the bigger chunks you use
- ▶ Remember phrases from training data, translate piece-by-piece and stitch those pieces together to translate
  - ▶ How to identify phrases? Word alignment over source-target bitext
  - ▶ How to stitch together? Language model over target language
- ▶ Decoder takes phrases and a language model and searches over possible translations
- ▶ NOT like standard discriminative models (take a bunch of translation pairs, learn a ton of parameters in an end-to-end way)

# Phrase-Based MT

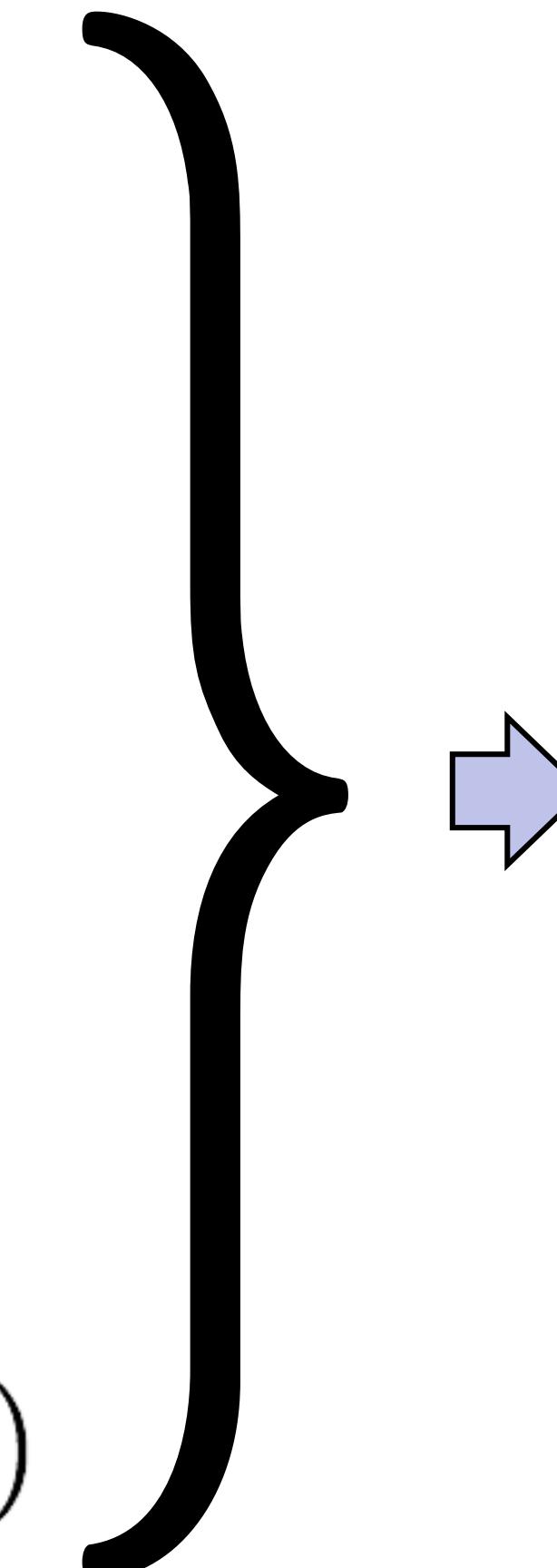
cat     chat     0.9
the cat     le chat     0.8
dog     chien     0.8
house     maison     0.6
my house     ma maison     0.9
language     langue     0.9
...

Phrase table  $P(f|e)$



Unlabeled English data

Language model  $P(e)$



$$P(e|f) \propto P(f|e)P(e)$$

Noisy channel model:  
combine scores from  
translation model +  
language model to  
translate foreign to  
English

“Translate faithfully but make fluent English”

# Evaluating MT

- ▶ Fluency: does it sound good in the target language?
- ▶ Fidelity/adequacy: does it capture the meaning of the original?
- ▶ BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right)$$

	<b>hypothesis 1</b>	<b>hypothesis 2</b>	<b>hypothesis 3</b>	<b>reference 1</b>	<b>reference 2</b>	1-gram	2-gram	3-gram
	I am exhausted	Tired is I	I I I	I am tired	I am ready to sleep now and so exhausted	3/3	1/2	0/1
						1/3	0/2	0/1
						1/3	0/2	0/1

# Evaluating MT

---

- ▶ Fluency: does it sound good in the target language?
- ▶ Fidelity/adequacy: does it capture the meaning of the original?
- ▶ BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad \blacktriangleright \text{Typically } n = 4, w_i = 1/4$$

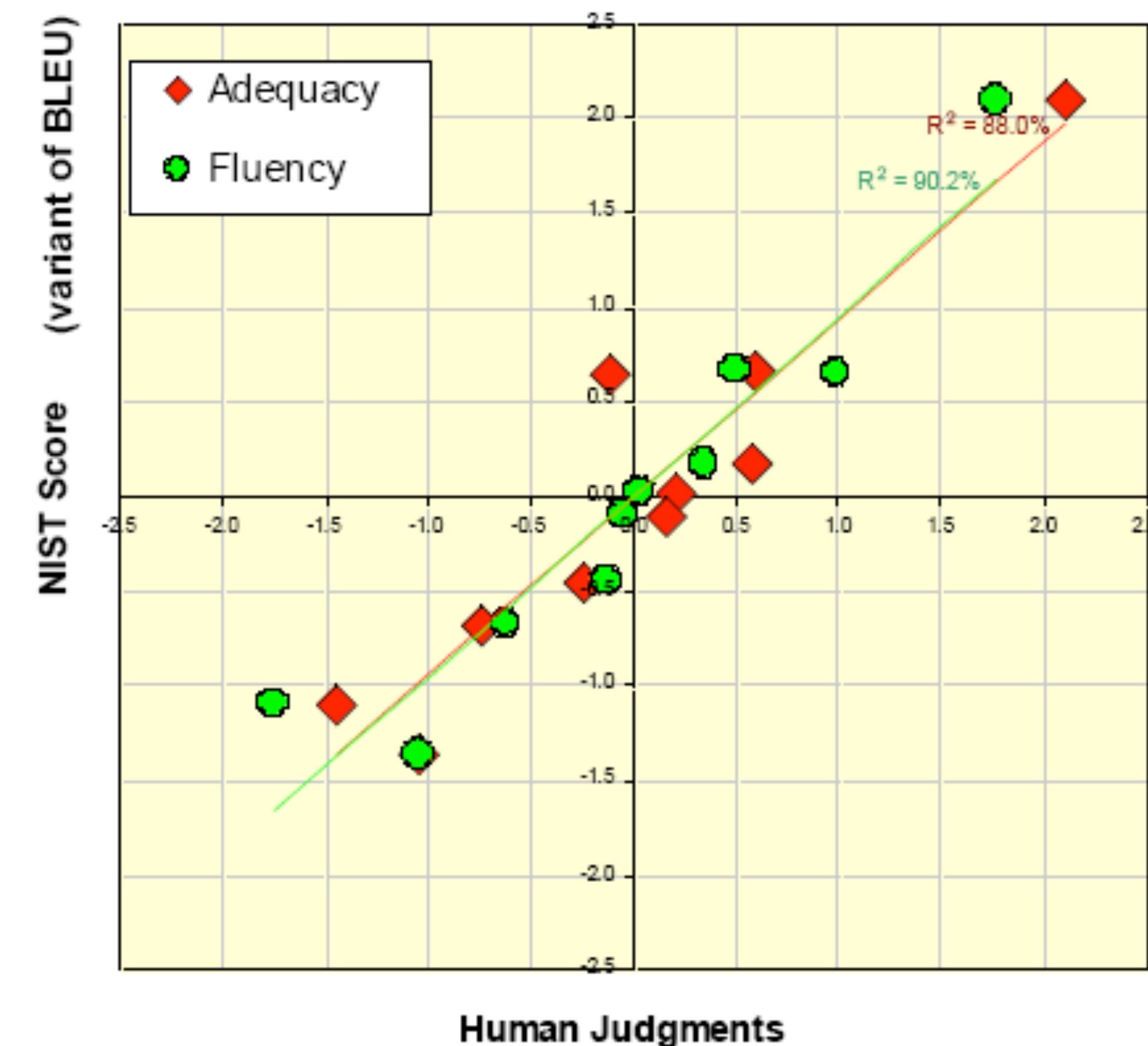
$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad \blacktriangleright r = \text{length of reference} \\ c = \text{length of system output}$$

- ▶ Does this capture fluency and adequacy?

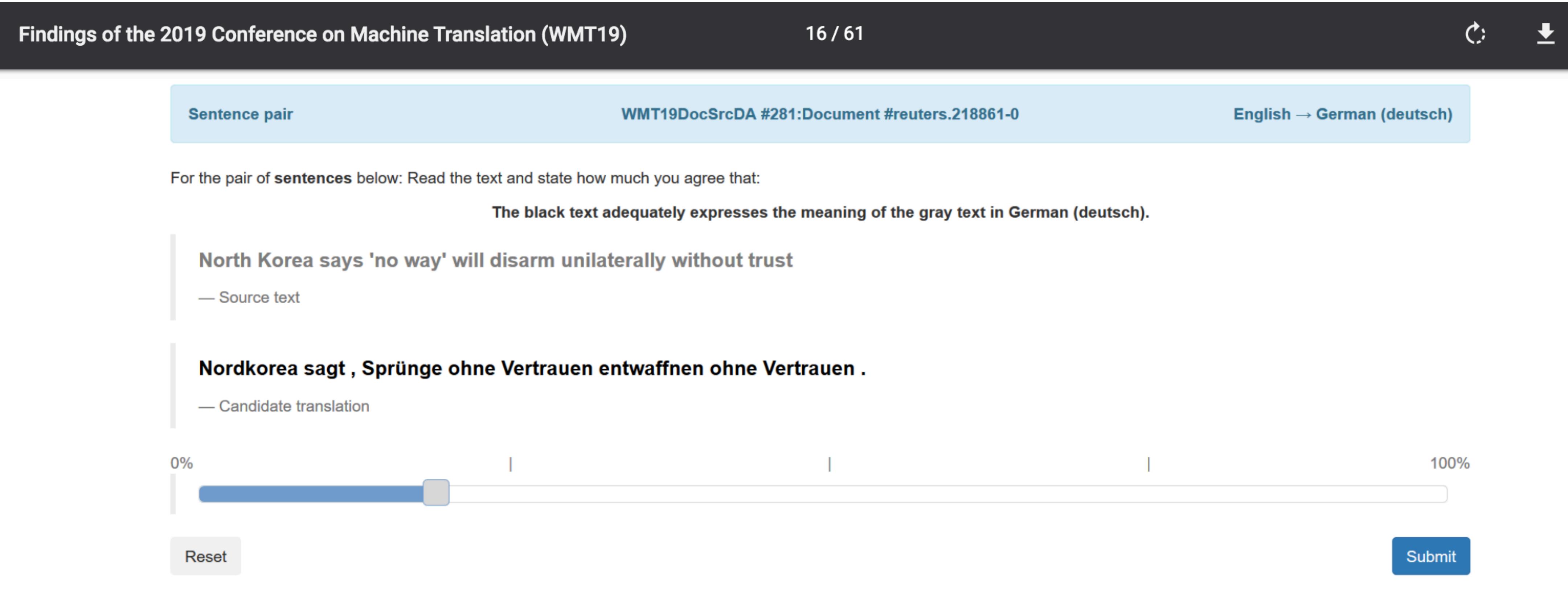
# BLEU Score

---

- ▶ Better methods with human-in-the-loop
- ▶ HTER: human-assisted translation error rate
- ▶ If you're building real MT systems, you do user studies.  
In academia, you mostly use BLEU.



# Appraise - Human Evaluation Interface



**Figure 3:** Screen shot of segment-rating portion of document-level direct assessment in the Appraise interface for an example English to German assessment from the human evaluation campaign. The annotator is presented with the machine translation output segment randomly selected from competing systems (anonymized) and is asked to rate the translation on a sliding scale.

# Word Alignment

# Word Alignment

---

- ▶ Input: a bitext, pairs of translated sentences

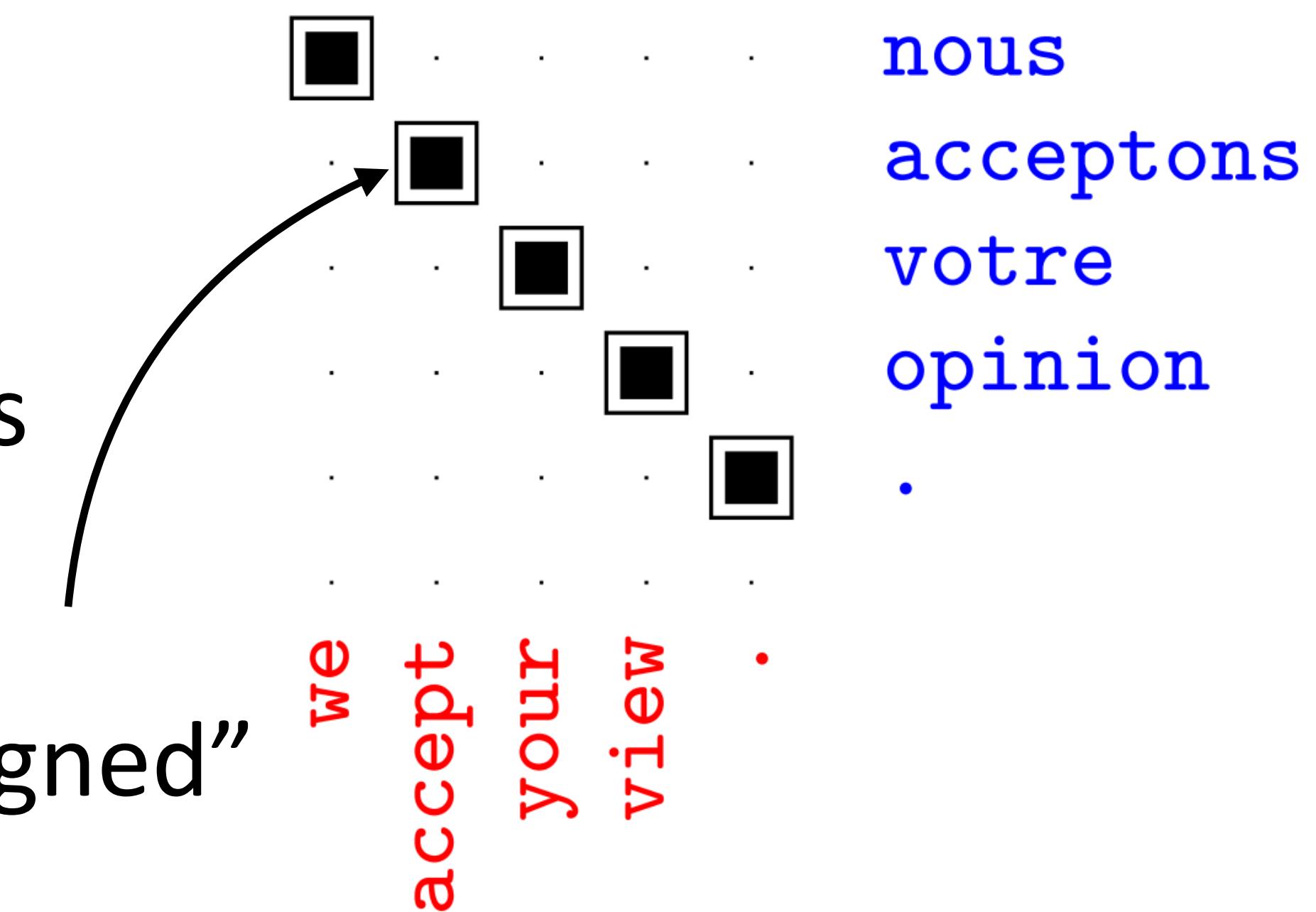
nous acceptons votre opinion . ||| we accept your view

nous allons changer d'avis ||| we are going to change our minds

- ▶ Output: alignments between words in each sentence

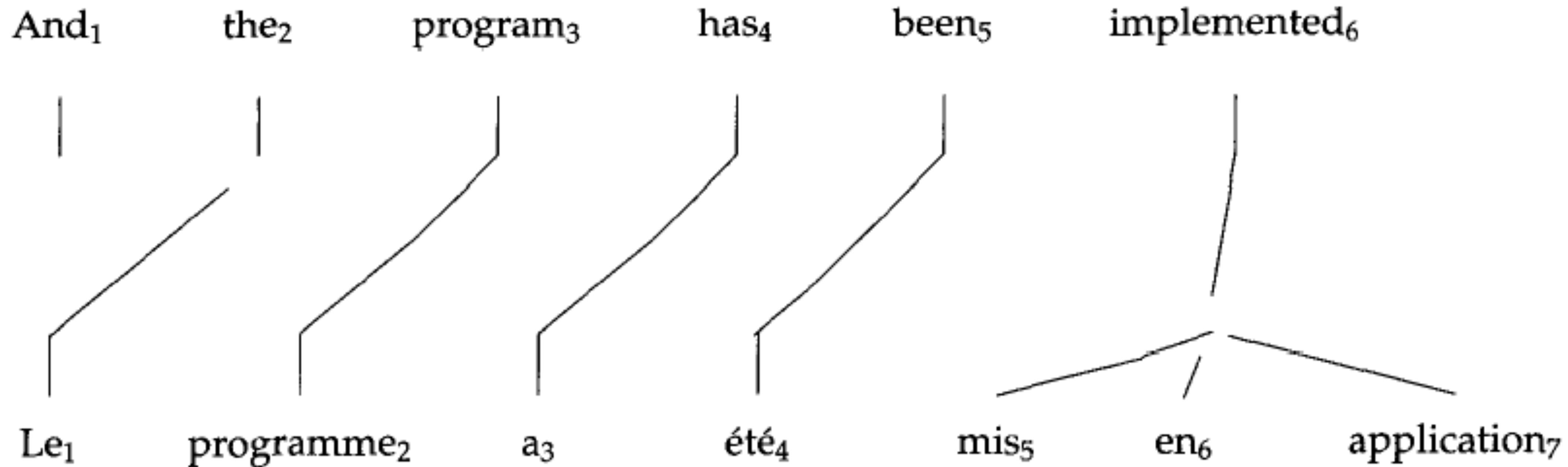
- ▶ We will see how to turn these into phrases

“accept and acceptons are aligned”



# 1-to-Many Alignments

---



# Word Alignment

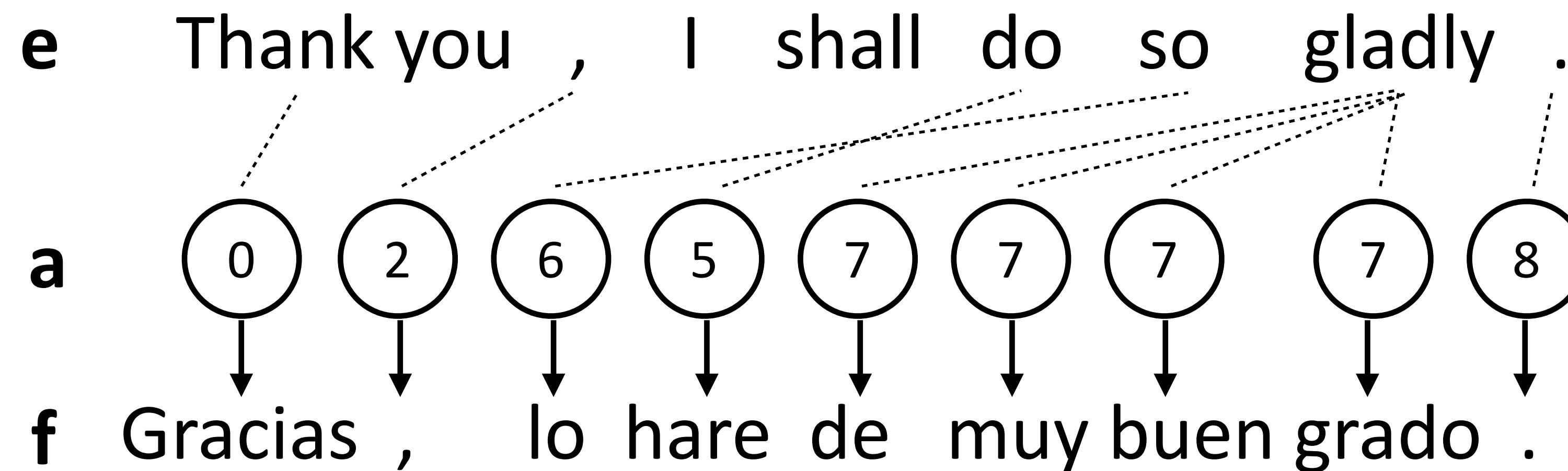
---

- ▶ Models  $P(f|e)$ : probability of “French” sentence being generated from “English” sentence according to a model
- ▶ Latent variable model: 
$$P(f|e) = \sum_a P(f, a|e) = \sum_a P(f|a, e)P(a)$$
- ▶ Correct alignments should lead to higher-likelihood generations, so by optimizing this objective we will learn correct alignments

# IBM Model 1

- ▶ Each French word is aligned to *at most* one English word

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{i=1}^n P(f_i|e_{a_i})P(a_i)$$

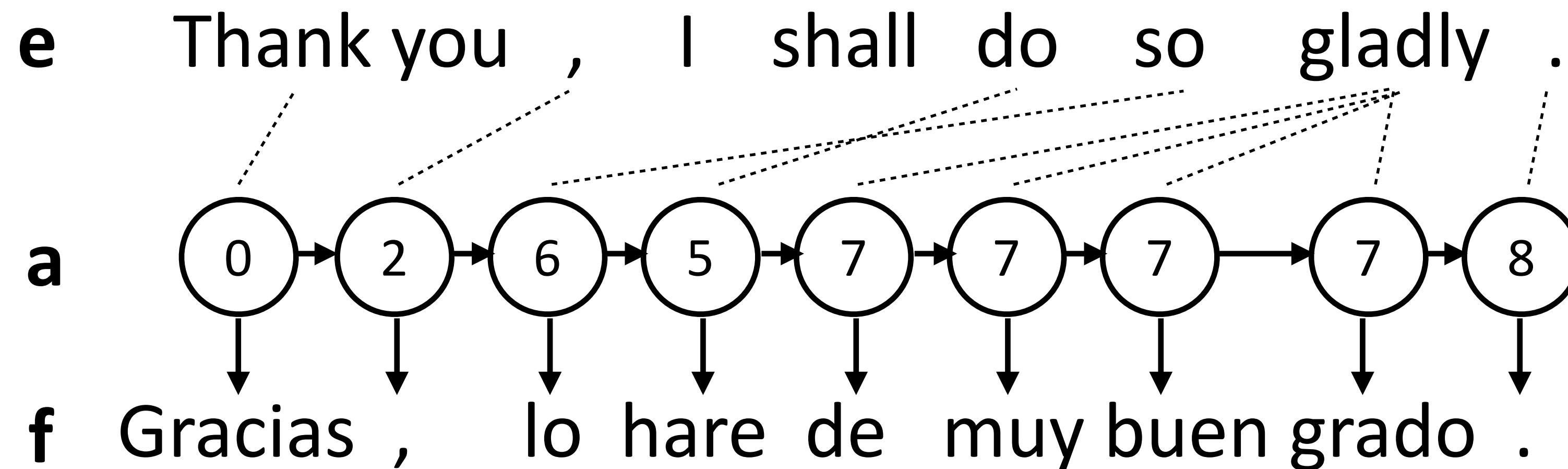


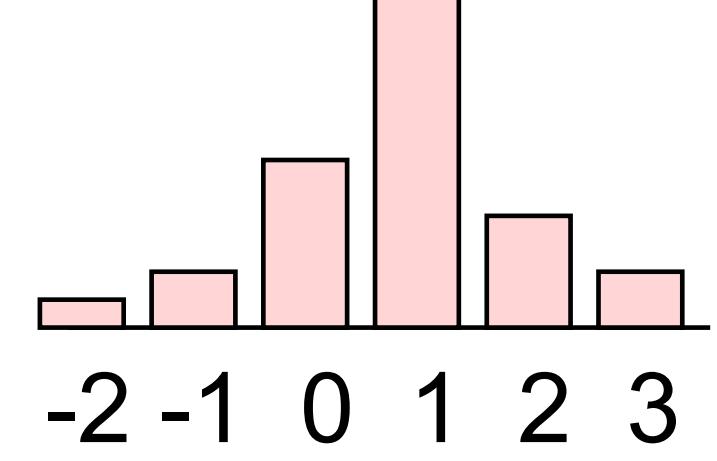
- ▶ Set  $P(a)$  uniformly (no prior over good alignments)
- ▶  $P(f_i|e_{a_i})$ : word translation probability table

# HMM for Alignment

- ▶ Sequential dependence between a's to capture monotonicity

$$P(\mathbf{f}, \mathbf{a} | \mathbf{e}) = \prod_{i=1}^n P(f_i | e_{a_i}) P(a_i | a_{i-1})$$

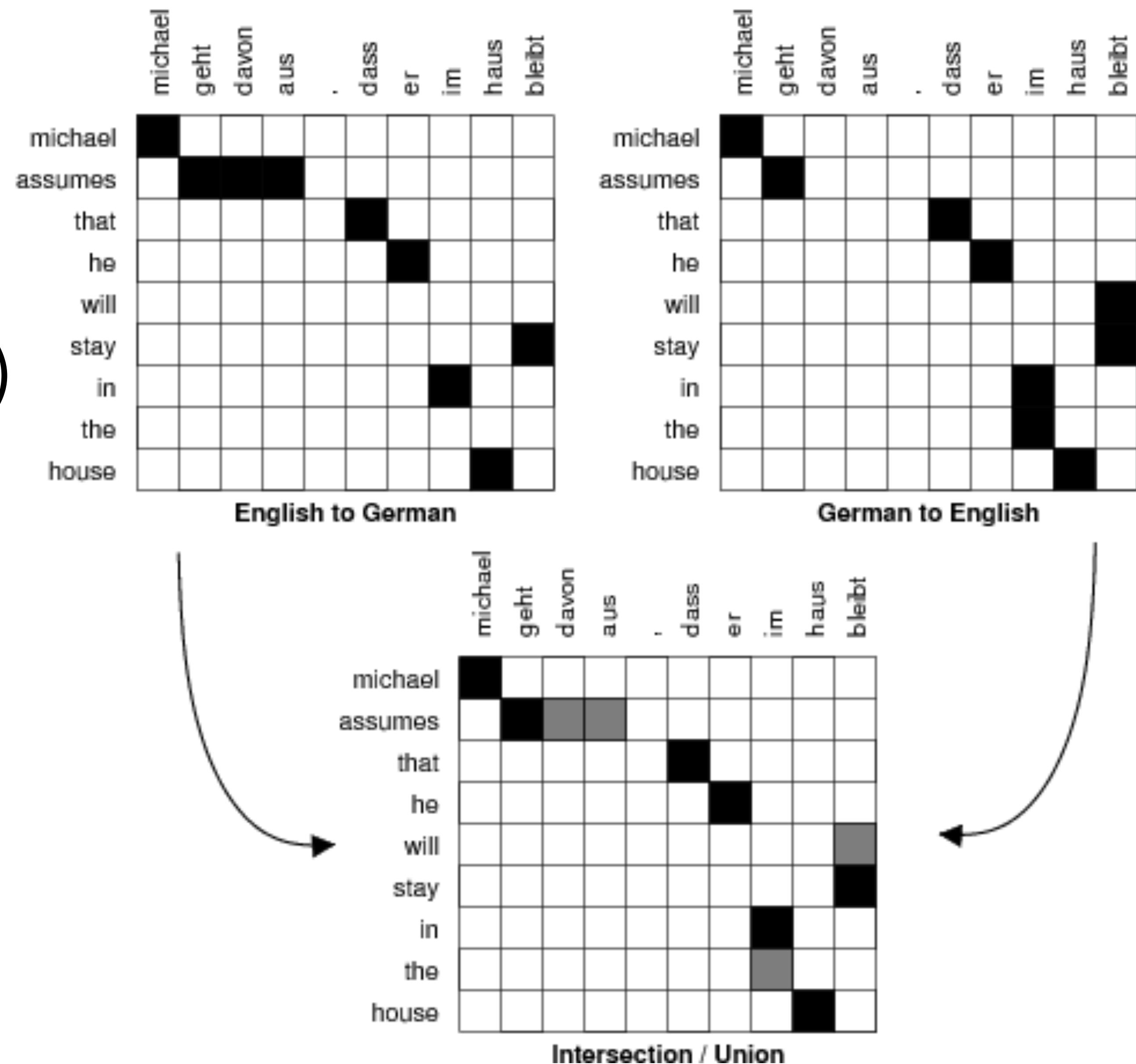


- ▶ Alignment dist parameterized by jump size:  $P(a_j - a_{j-1})$  
- ▶  $P(f_i | e_{a_i})$ : same as before

Brown et al. (1993)

# HMM Model

- ▶ Which direction is this?
- ▶ Alignments are generally monotonic (along diagonal)
- ▶ Some mistakes, especially when you have rare words (*garbage collection*)



# Evaluating Word Alignment

- ▶ “Alignment error rate”: use labeled alignments on small corpus

Model	AER
Model 1 INT	19.5
HMM E→F	11.4
HMM F→E	10.8
HMM AND	7.1
HMM INT	4.7
GIZA M4 AND	6.9

- ▶ Run Model 1 in both directions and intersect “intelligently”
- ▶ Run HMM model in both directions and intersect “intelligently”

# Phrase Extraction

- ▶ Find contiguous sets of aligned words in the two languages that don't have alignments to other words

d'assister à la reunion et ||| to attend the meeting and

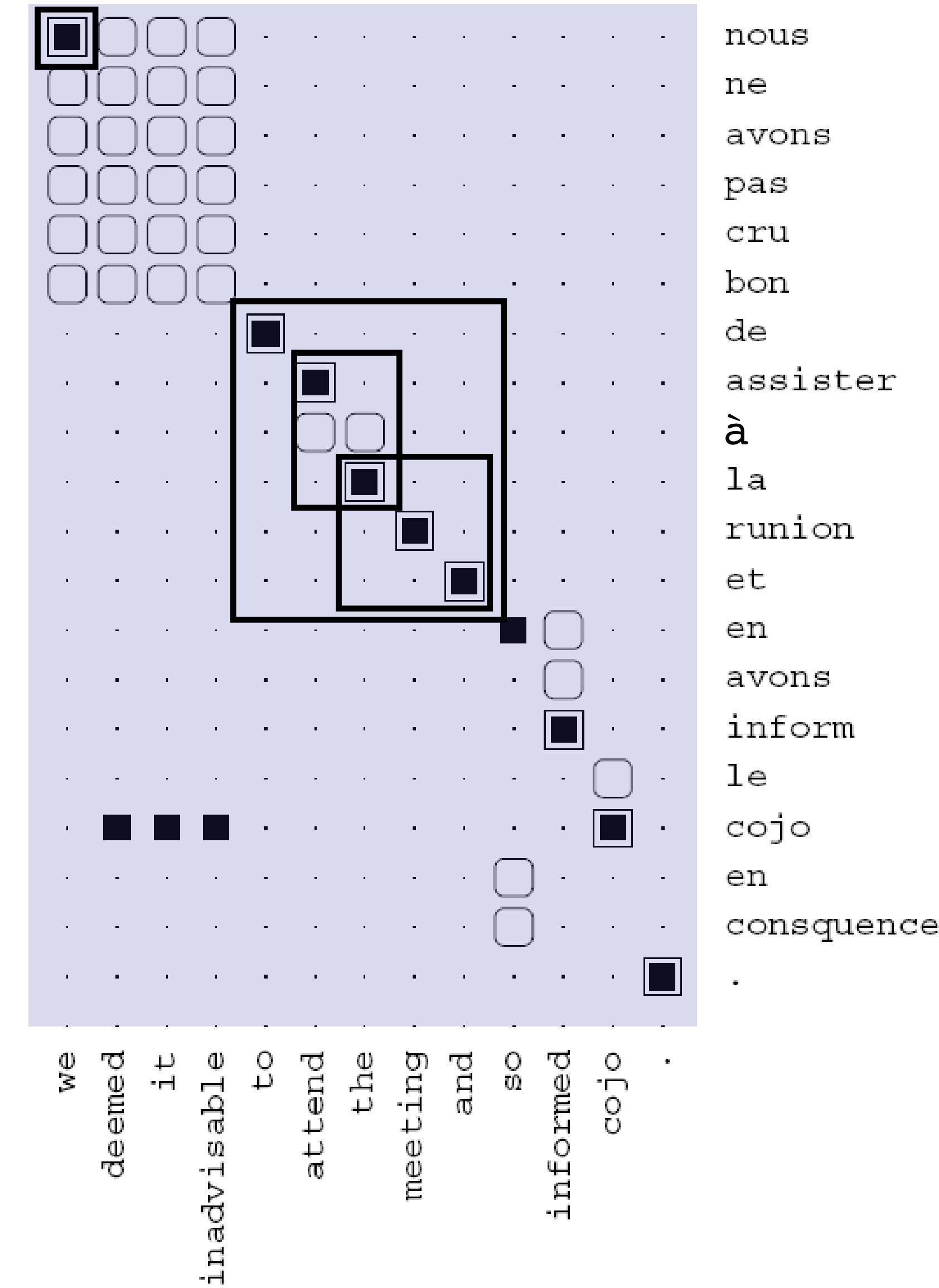
assister à la reunion ||| attend the meeting

la reunion and ||| the meeting and

nous ||| we

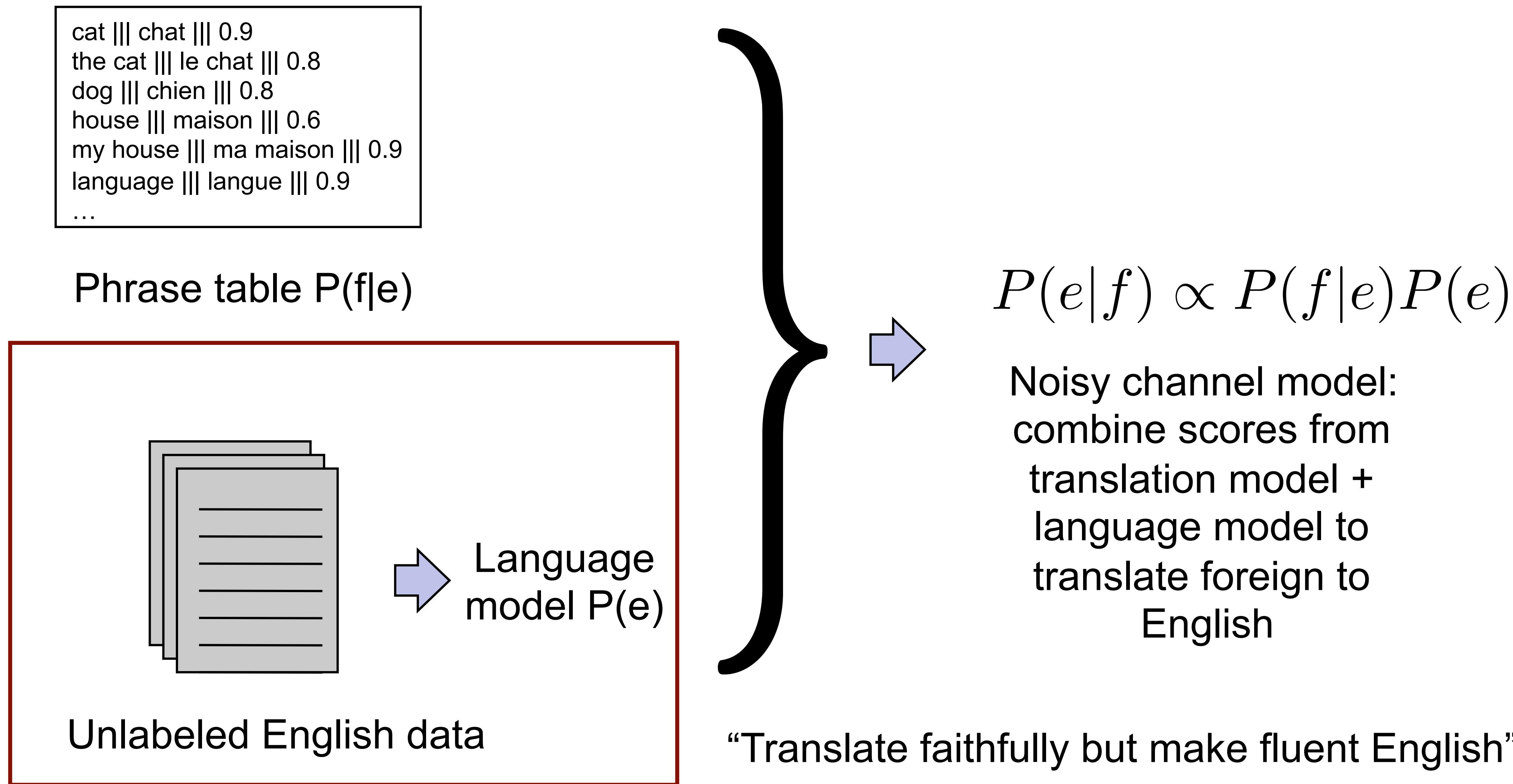
...

- ▶ Lots of phrases possible, count across all sentences and score by frequency



# Language Modeling

# Phrase-Based MT



# N-gram Language Models

---

I visited San \_\_\_\_\_ put a distribution over the next word

- ▶ Simple generative model: distribution of next word is a multinomial distribution conditioned on previous n-1 words

$$P(x|\text{visited San}) = \frac{\text{count}(\text{visited San}, x)}{\text{count}(\text{visited San})}$$

Maximum likelihood estimate of this probability from a corpus

- ▶ Just relies on counts, even in 2008 could scale up to 1.3M word types, 4B n-grams (all 5-grams occurring >40 times on the Web)

# Smoothing N-gram Language Models

---

I visited San \_\_\_\_\_ put a distribution over the next word!

- ▶ Smoothing is very important, particularly when using 4+ gram models

$$P(x|\text{visited San}) = (1 - \lambda) \frac{\text{count}(\text{visited San}, x)}{\text{count}(\text{visited San})} + \lambda \frac{\text{count}(\text{San}, x)}{\text{count}(\text{San})}$$

smooth  
this  
too!

- ▶ One technique is “absolute discounting:” subtract off constant  $k$  from numerator, set lambda to make this normalize ( $k=1$  is like leave-one-out)

$$P(x|\text{visited San}) = \frac{\text{count}(\text{visited San}, x) - k}{\text{count}(\text{visited San})} + \lambda \frac{\text{count}(\text{San}, x)}{\text{count}(\text{San})}$$

- ▶ Kneser-Ney smoothing: this trick, plus low-order distributions modified to capture fertilities (how many distinct words appear in a context)

# Engineering N-gram Models

- ▶ For 5+-gram models, need to store between 100M and 10B context-word-count triples

(a) Context-Encoding

$w$	$c$	$val$
1933	15176585	3
1933	15176587	2
1933	15176593	1
1933	15176613	8
1933	15179801	1
1935	15176585	298
1935	15176589	1

(b) Context Deltas

$\Delta w$	$\Delta c$	$val$
1933	15176585	3
+0	+2	1
+0	+5	1
+0	+40	8
+0	+188	1
+2	15176585	298
+0	+4	1

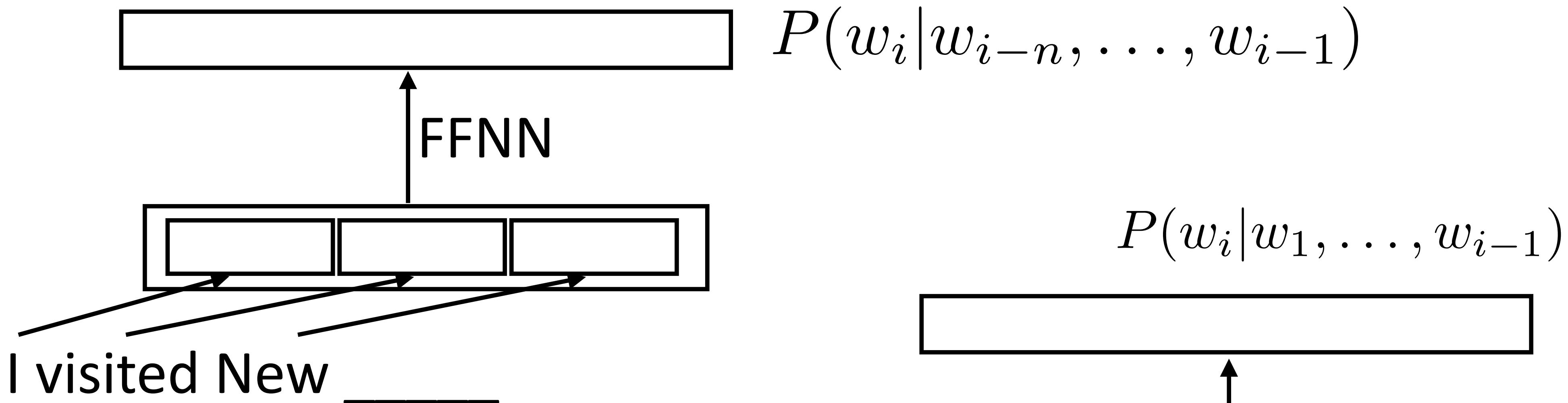
(c) Bits Required

$ \Delta w $	$ \Delta c $	$ val $
24	40	3
2	3	3
2	3	3
2	9	6
2	12	3
4	36	15
2	6	3

- ▶ Make it fit in memory by *delta encoding* scheme: store deltas instead of values and use variable-length encoding

# Neural Language Models

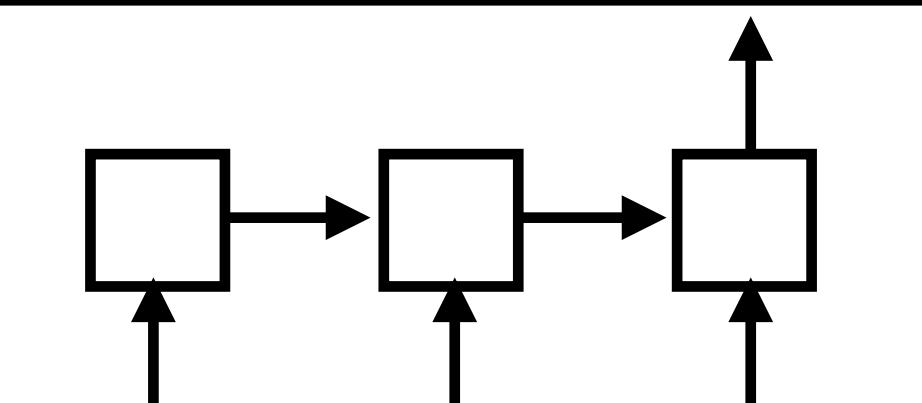
- ▶ Early work: feedforward neural networks looking at context



- ▶ Variable length context with RNNs:

- ▶ Works like a decoder with no encoder

- ▶ Slow to train over lots of data!



I visited New

Mnih and Hinton (2003)

# Evaluation

---

- ▶ (One sentence) negative log likelihood:  $\sum_{i=1}^n \log p(x_i | x_1, \dots, x_{i-1})$
- ▶ Perplexity:  $2^{-\frac{1}{n} \sum_{i=1}^n \log_2 p(x_i | x_1, \dots, x_{i-1})}$ 
  - ▶ NLL (base 2) averaged over the sentence, exponentiated
  - ▶ NLL = -2  $\rightarrow$  on average, correct thing has prob 1/4  $\rightarrow$  PPL = 4. PPL is sort of like branching factor

# Results

---

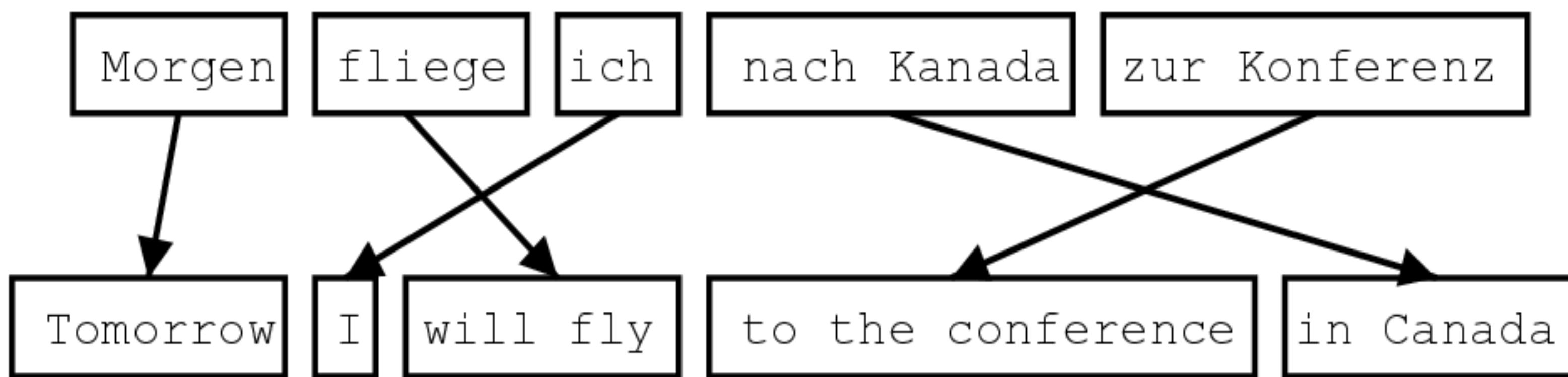
- ▶ Evaluate on Penn Treebank: small dataset (1M words) compared to what's used in MT, but common benchmark
- ▶ Kneser-Ney 5-gram model with cache: PPL = 125.7
- ▶ LSTM: PPL ~ 60-80 (depending on how much you optimize it)
- ▶ Melis et al.: many neural LM improvements from 2014-2017 are subsumed by just using the right regularization (right dropout settings). So LSTMs are pretty good

# Decoding

# Phrase-Based Decoding

---

- ▶ Inputs:
  - ▶ Language model that scores  $P(e_i|e_1, \dots, e_{i-1}) \approx P(e_i|e_{i-n-1}, \dots, e_{i-1})$
  - ▶ Phrase table: set of phrase pairs  $(e, f)$  with probabilities  $P(f|e)$
- ▶ What we want to find:  $e$  produced by a series of phrase-by-phrase translations from an input  $f$ , possibly with reordering:



# Phrase lattices are big!

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

the	7 people	including	by some	and	the russian	the	the astronauts	,
it	7 people included	by france		and the	the russian		international astronautical	of rapporteur .
this	7 out	including the	from	the french	and the russian	the fifth		.
these	7 among	including from		the french and	of the russian	of	space	members .
that	7 persons	including from the		of france	and to	russian	of the	aerospace members .
	7 include	from the	of france and		russian		astronauts	. the
	7 numbers include	from france		and russian		of astronauts who		."
	7 populations include	those from france		and russian		astronauts .		
	7 deportees included	come from	france	and russia	in	astronautical	personnel	;
7 philtrum	including those from		france and	russia	a space		member	
	including representatives from		france and the	russia	astronaut			
	include	came from	france and russia		by cosmonauts			
	include representatives from	french	and russia		cosmonauts			
	include	came from france	and russia 's		cosmonauts .			
	includes	coming from	french and	russia 's	cosmonaut			
			french and russian	's	astronavigation	member .		
			french	and russia	astronauts			
				and russia 's			special rapporteur	
				, and russia			rapporteur	
				, and russia			rapporteur .	
				, and russia				
				or	russia 's			

# Phrase-Based Decoding

- ▶ Input

lo haré | rápidamente |.

- ▶ Translations

I'll do it | quickly |.

quickly | I'll do it |.

- ▶ Decoding  
objective (for  
3-gram LM)

$$\arg \max_{\mathbf{e}} [P(\mathbf{f}|\mathbf{e}) \cdot P(\mathbf{e})]$$

$$\arg \max_{\mathbf{e}} \left[ \prod_{\langle \bar{e}, f \rangle} P(\bar{f}|\bar{e}) \cdot \prod_{i=1}^{|\mathbf{e}|} P(e_i|e_{i-1}, e_{i-2}) \right]$$

*The decoder...*

*tries different segmentations,*

*translates phrase by phrase,*

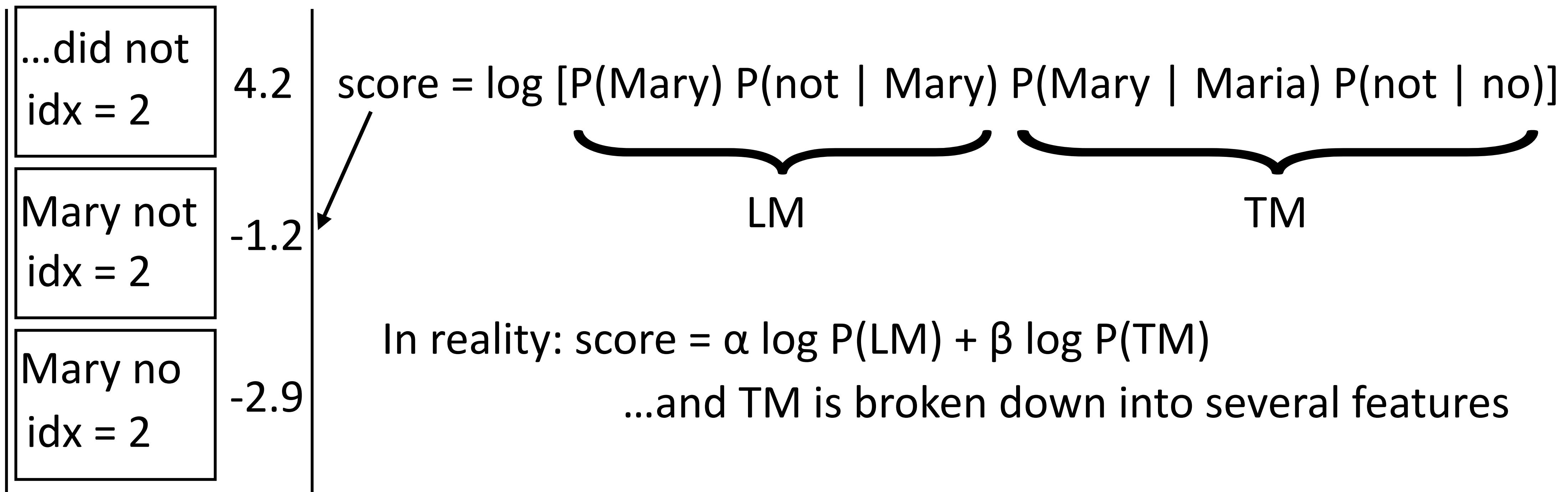
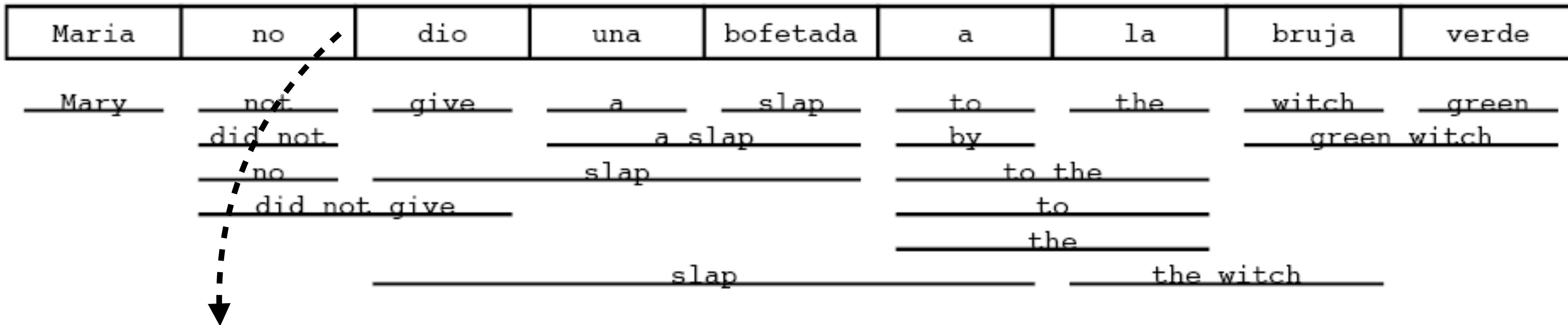
*and considers reorderings.*

# Monotonic Translation

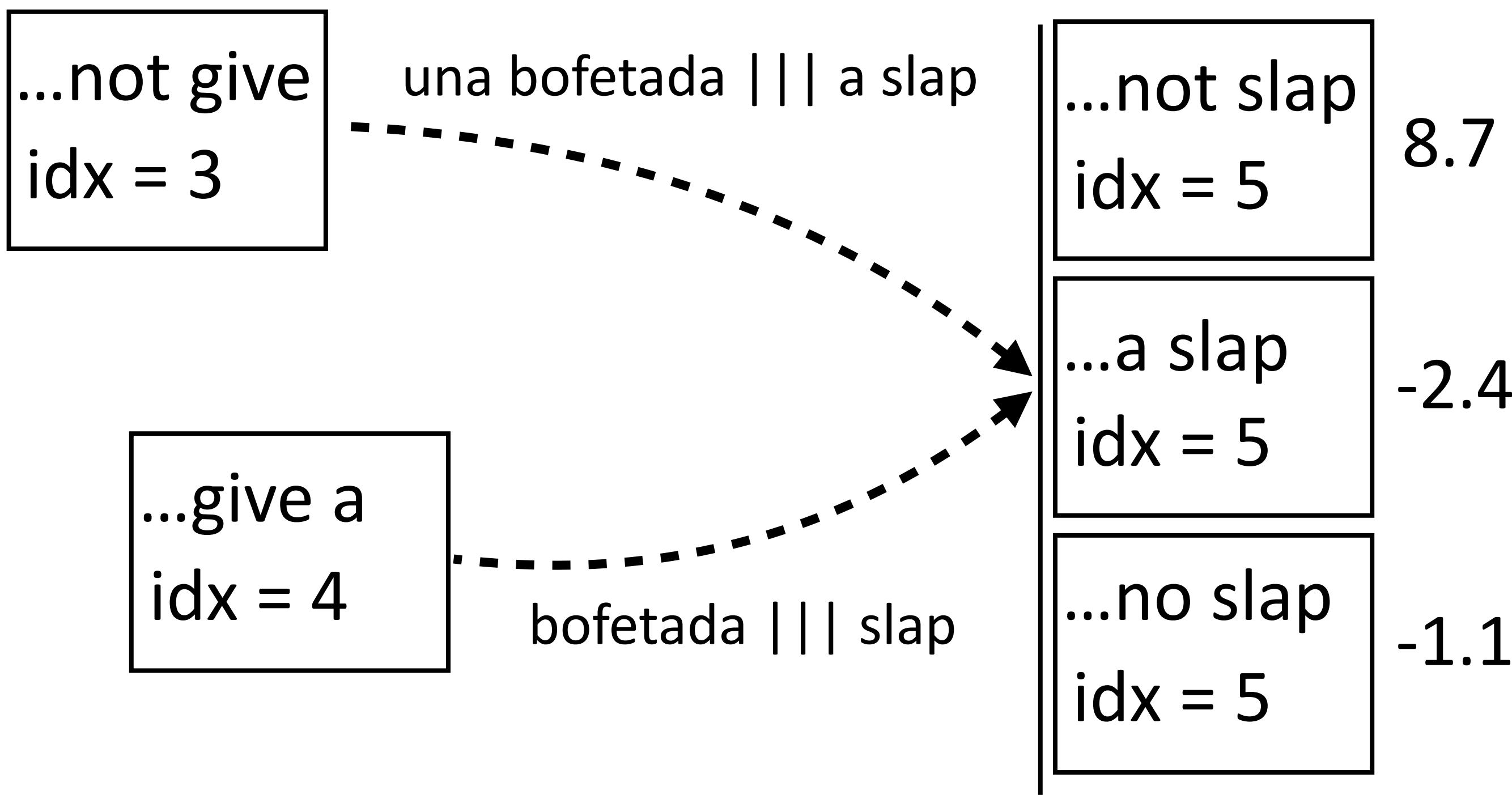
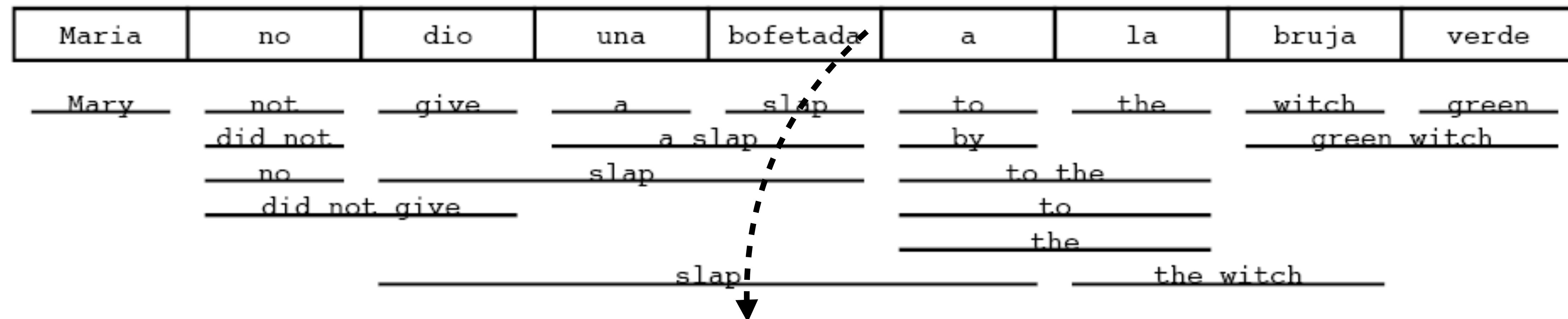
Maria	no	dio	una	bofetada	a	la	bruja	verde
<u>Mary</u>	<u>not</u>	<u>give</u>	<u>a</u>	<u>slap</u>	<u>to</u>	<u>the</u>	<u>witch</u>	<u>green</u>
	<u>did not</u>			<u>a slap</u>		<u>by</u>		<u>green witch</u>
	<u>no</u>		<u>slap</u>			<u>to the</u>		
		<u>did not give</u>				<u>to</u>		
						<u>the</u>		
				<u>slap</u>			<u>the witch</u>	

- ▶ If we translate with beam search, what state do we need to keep in the beam?
- ▶ What have we translated so far?  $\arg \max_{\mathbf{e}} \left[ \prod_{\langle \bar{e}, \bar{f} \rangle} P(\bar{f} | \bar{e}) \cdot \prod_{i=1}^{|\mathbf{e}|} P(e_i | e_{i-1}, e_{i-2}) \right]$
- ▶ What words have we produced so far?
- ▶ When using a 3-gram LM, only need to remember the last 2 words!

# Monotonic Translation



# Monotonic Translation

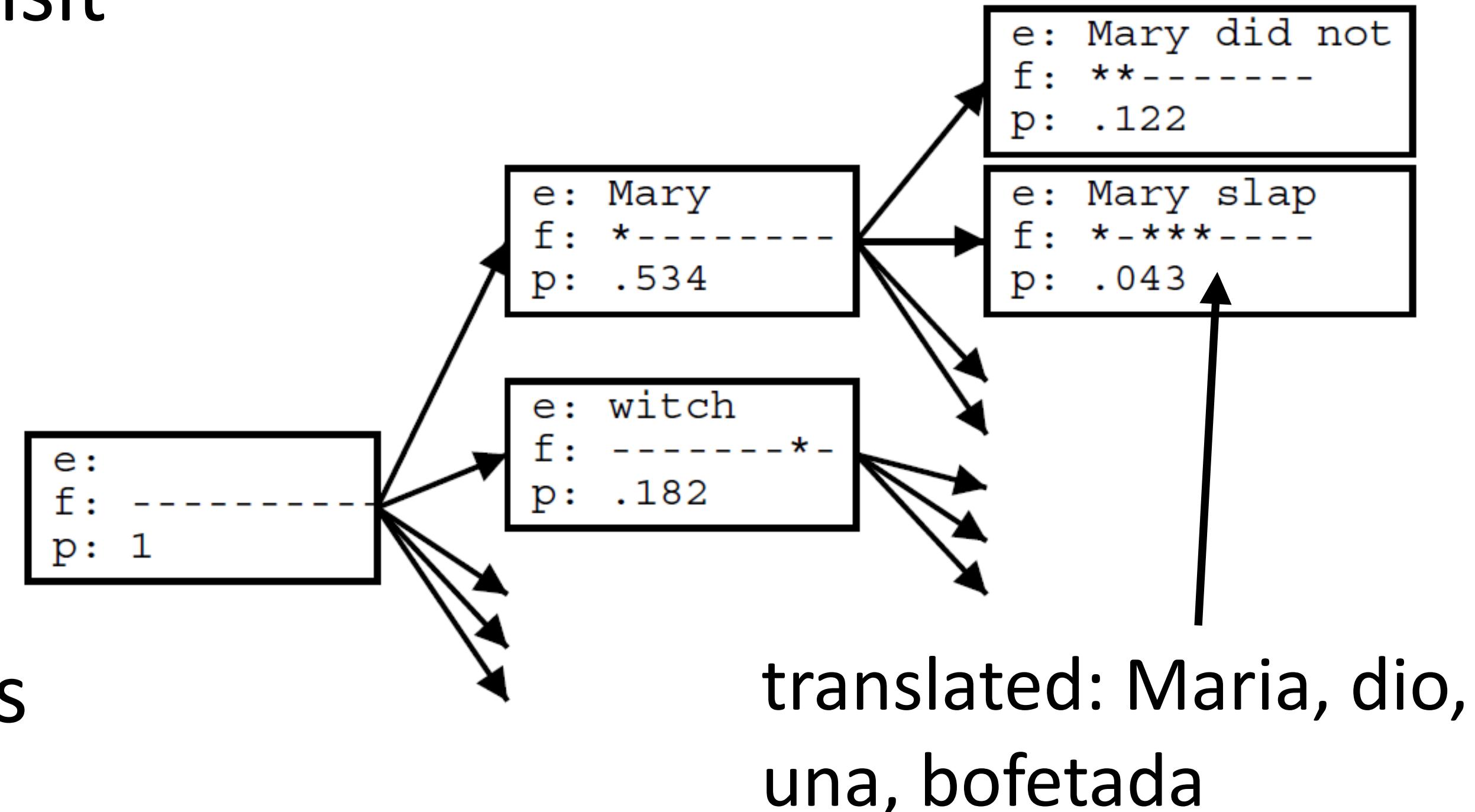


- ▶ Several paths can get us to this state, max over them (like Viterbi)
- ▶ Variable-length translation pieces = semi-HMM

# Non-Monotonic Translation

Maria	no	dio	una	bofetada	a	la	bruja	verde
<u>Mary</u>	<u>not</u>	<u>give</u>	<u>a</u>	<u>slap</u>	<u>to</u>	<u>the</u>	<u>witch</u>	<u>green</u>
	<u>did not</u>			<u>a slap</u>		<u>by</u>		<u>green witch</u>
	<u>no</u>			<u>slap</u>		<u>to the</u>		
						<u>to</u>		
						<u>the</u>		
							<u>the witch</u>	
				<u>slap</u>				

- ▶ Non-monotonic translation: can visit source sentence “out of order”
- ▶ State needs to describe which words have been translated and which haven’t
- ▶ Big enough phrases already capture lots of reorderings, so this isn’t as important as you think

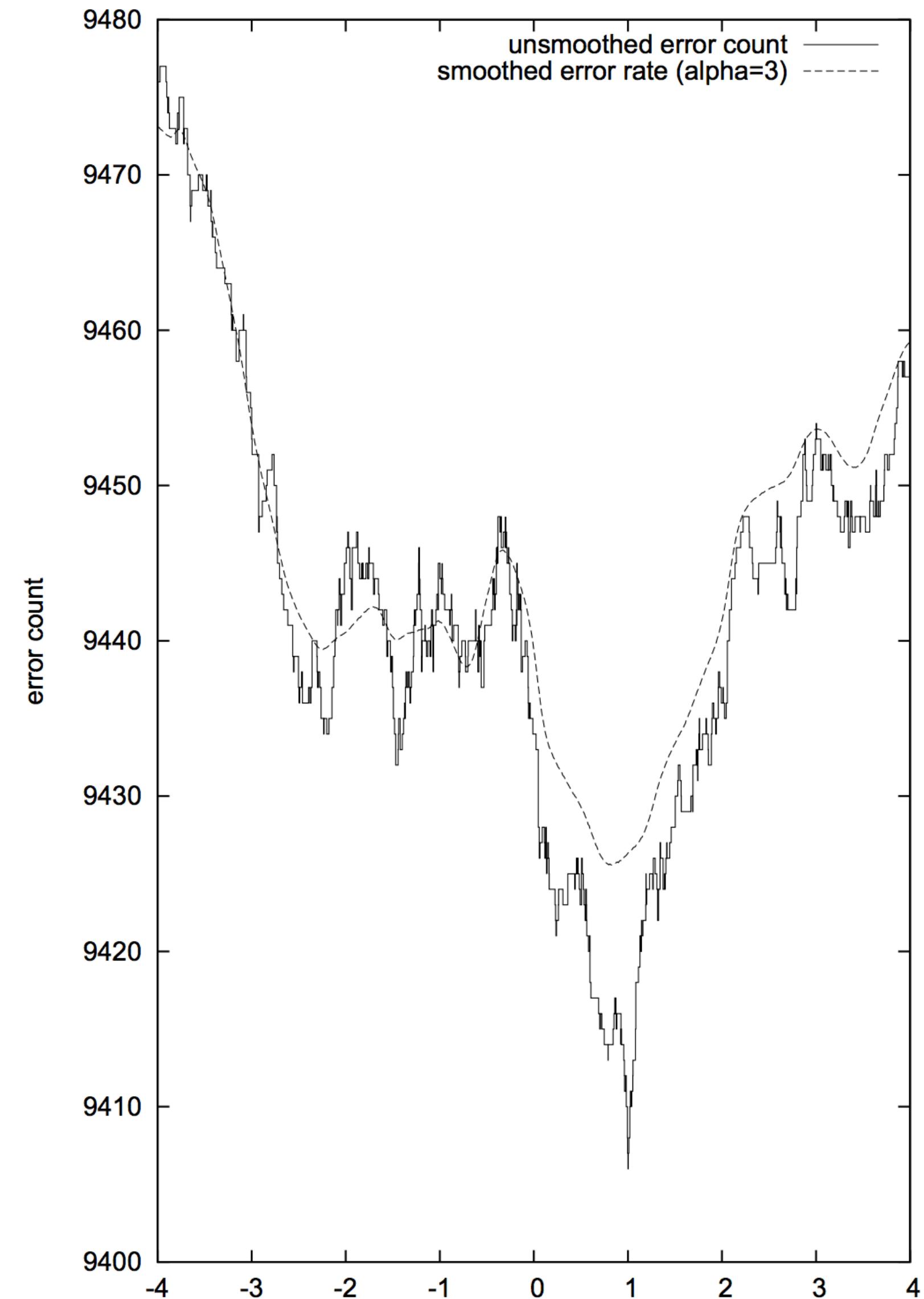


# Training Decoders

$$\text{score} = \alpha \log P(\text{LM}) + \beta \log P(\text{TM})$$

...and TM is broken down into several feature

- ▶ Usually 5-20 feature weights to set, want to optimize for BLEU score which is not differentiable
- ▶ MERT (Och 2003): decode to get 1000-best translations for each sentence in a small training set (<1000 sentences), do line search on parameters to directly optimize for BLEU



# Moses

---

- ▶ Toolkit for machine translation due to Philipp Koehn + Hieu Hoang
  - ▶ Pharaoh (Koehn, 2004) is the decoder from Koehn's thesis
- ▶ Moses implements word alignment, language models, and this decoder, plus \*a ton\* more stuff
  - ▶ Highly optimized and heavily engineered, could more or less build SOTA translation systems with this from 2007-2013
- ▶ Next week: results on these and comparisons to neural methods

# Syntax

# Syntactic MT

- ▶ Rather than use phrases, use a *synchronous context-free grammar*

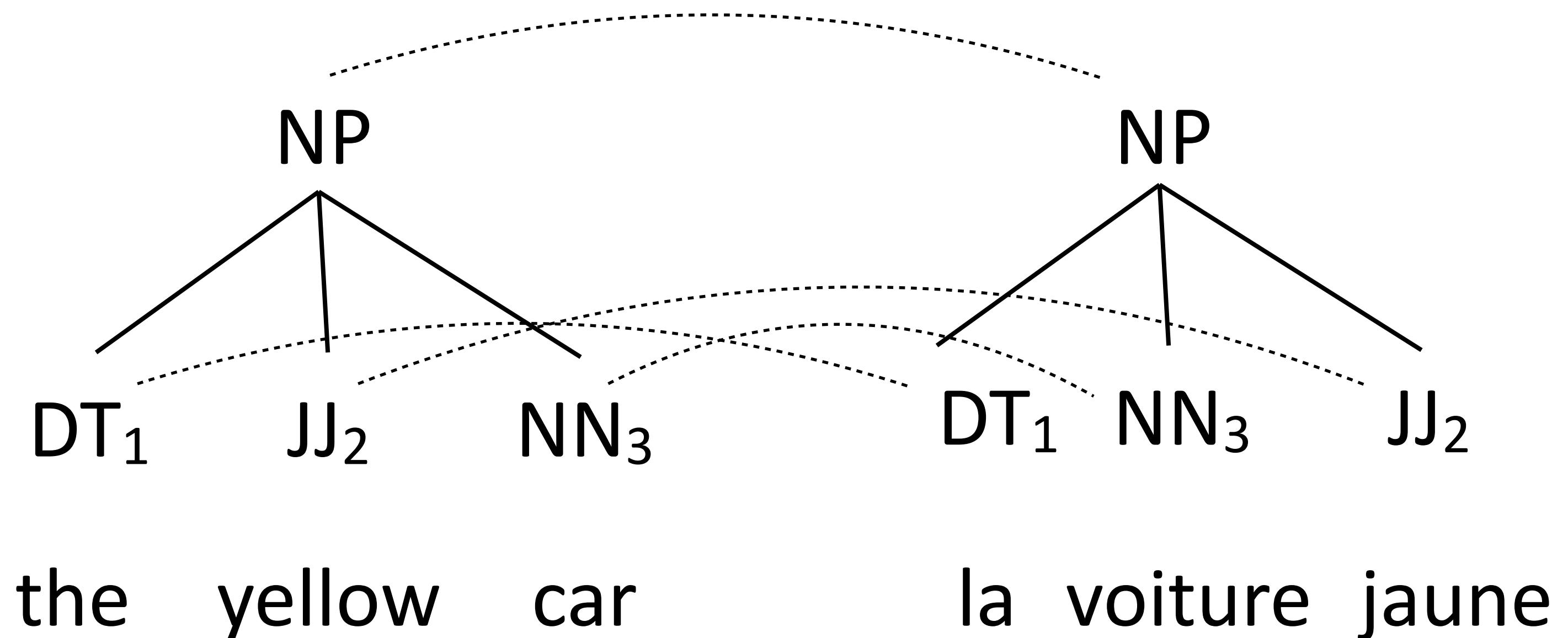
$NP \rightarrow [DT_1\ JJ_2\ NN_3; DT_1\ NN_3\ JJ_2]$

$DT \rightarrow [\text{the}, \text{la}]$

$DT \rightarrow [\text{the}, \text{le}]$

$NN \rightarrow [\text{car}, \text{voiture}]$

$JJ \rightarrow [\text{yellow}, \text{jaune}]$



- ▶ Translation = parse the input with “half” of the grammar, read off the other half
- ▶ Assumes parallel syntax up to reordering

# Syntactic MT

---

- ▶ Rather than use phrases, use a *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>

# Syntactic MT

---

- ▶ Rather than use phrases, use a *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*

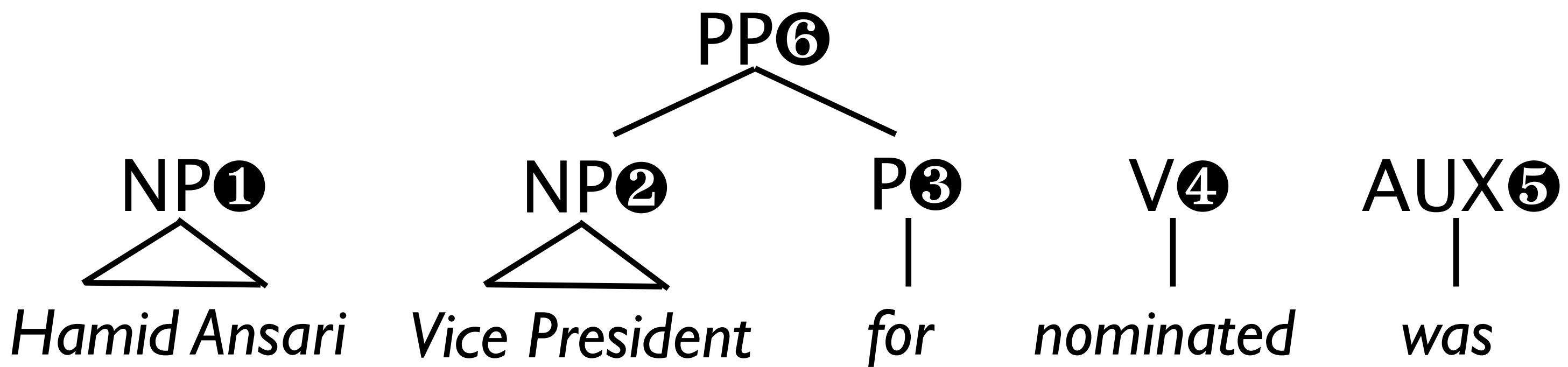
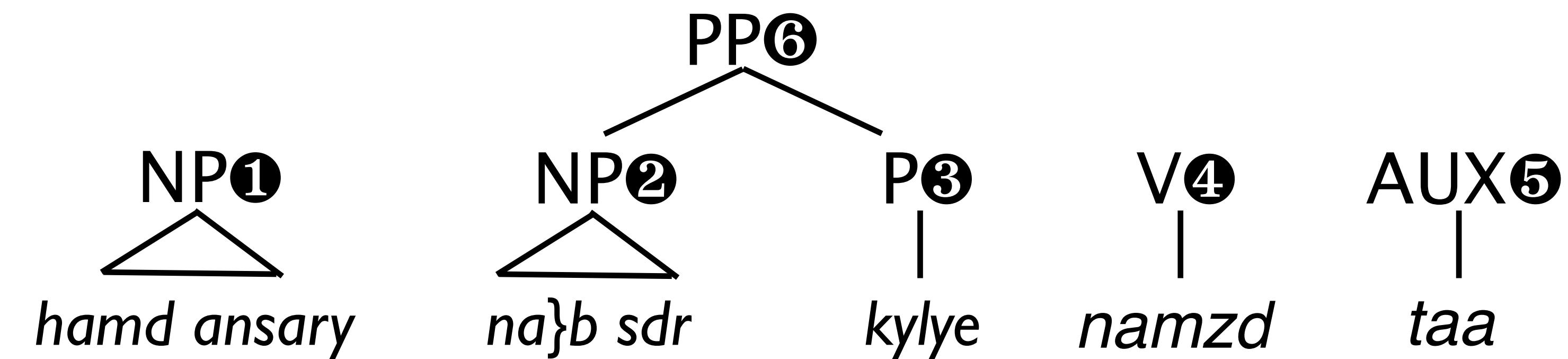
NP①  
  
*Hamid Ansari*

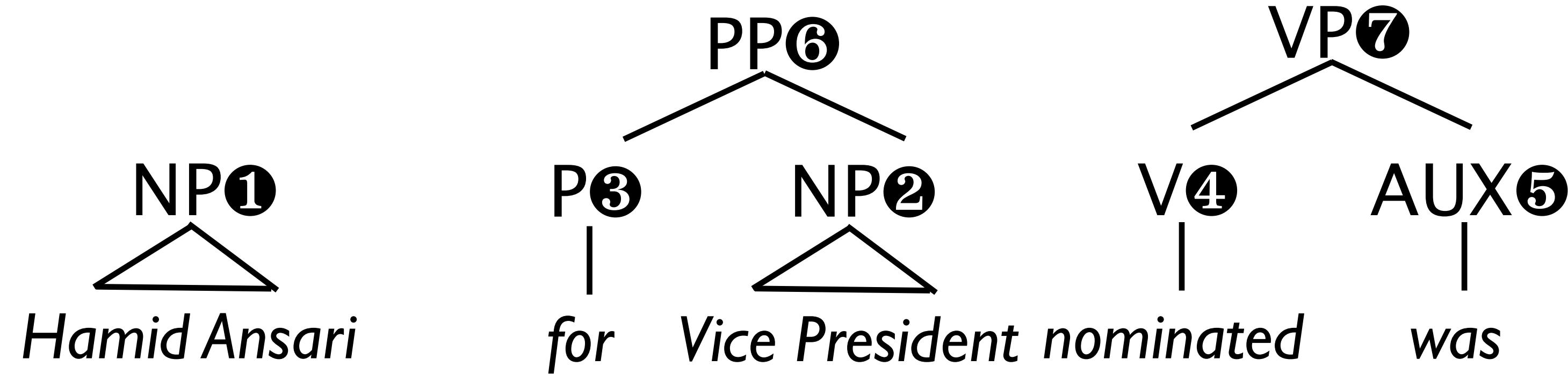
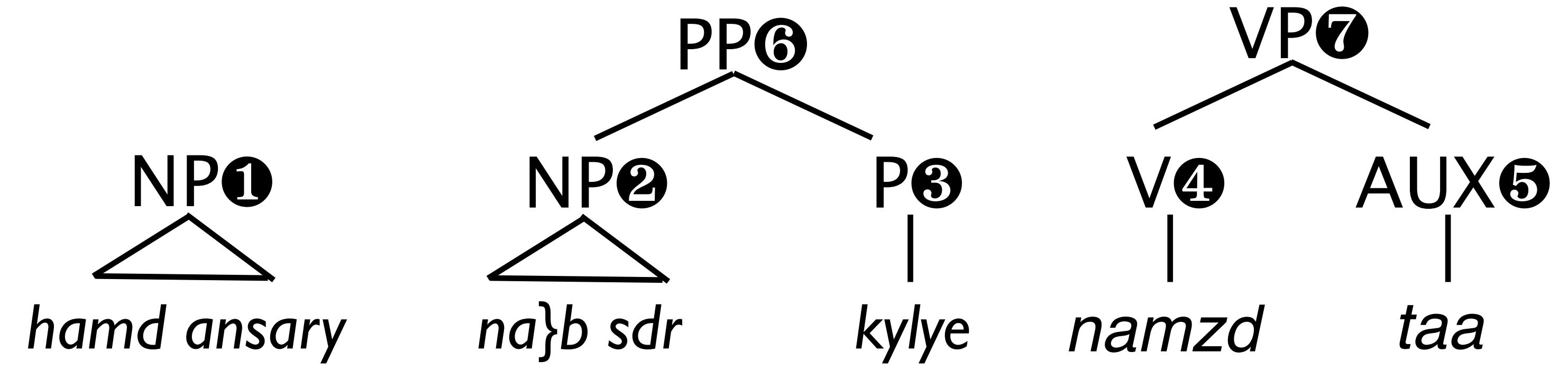
NP②  
  
*Vice President*

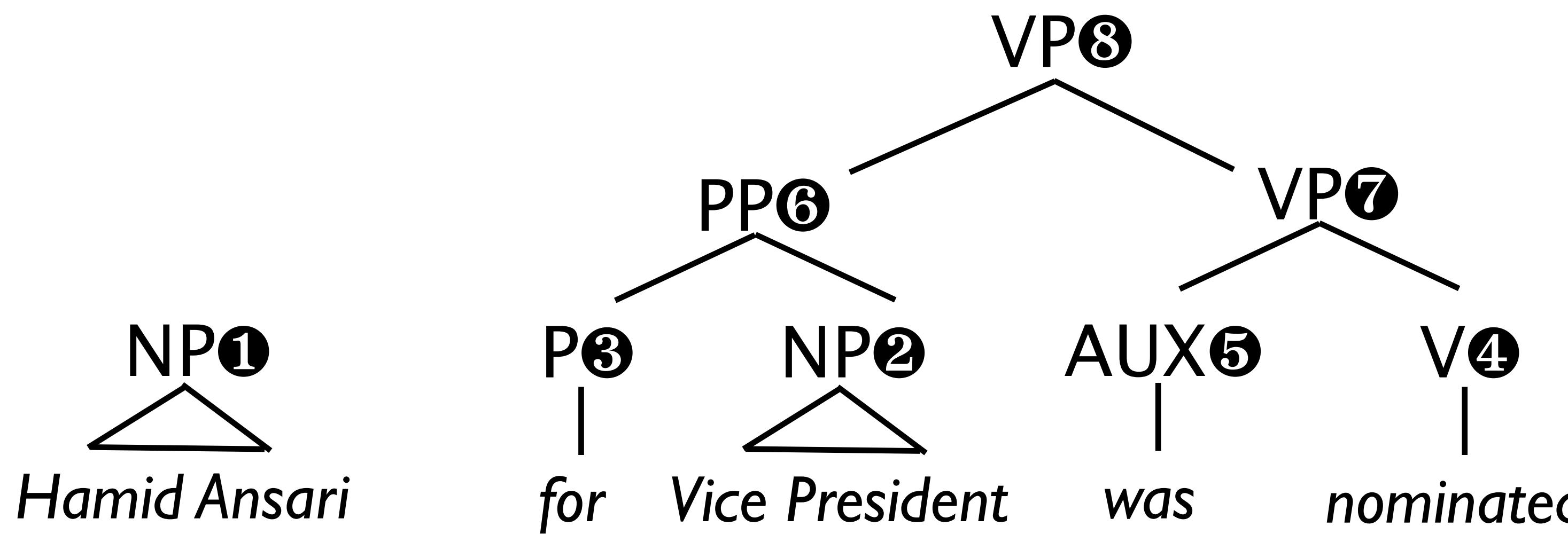
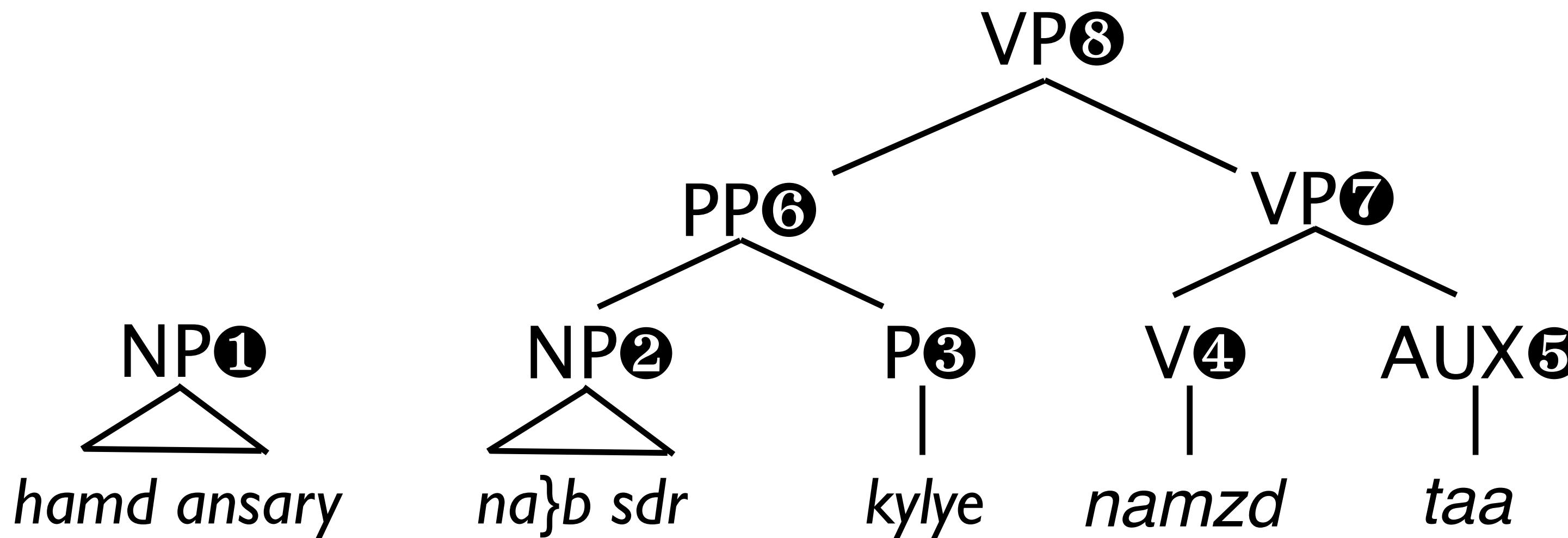
P③  
|  
*for*

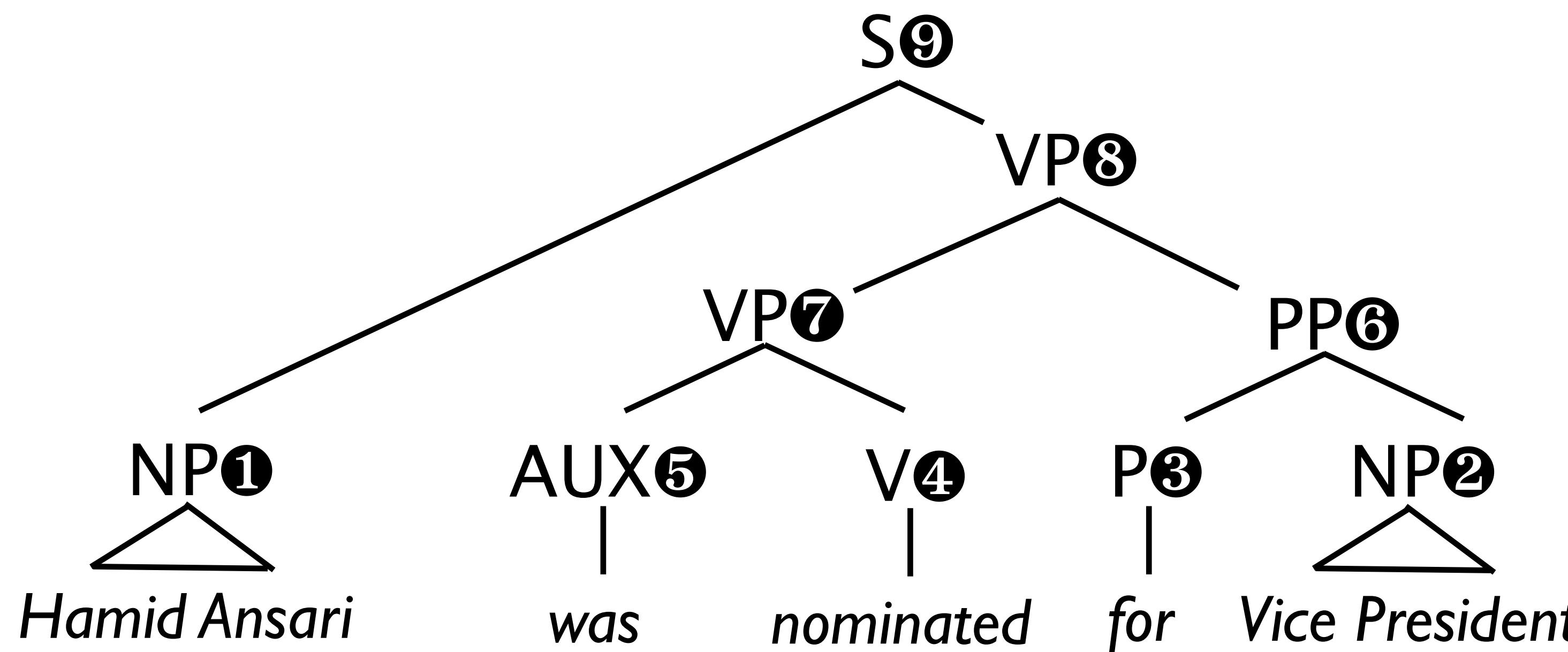
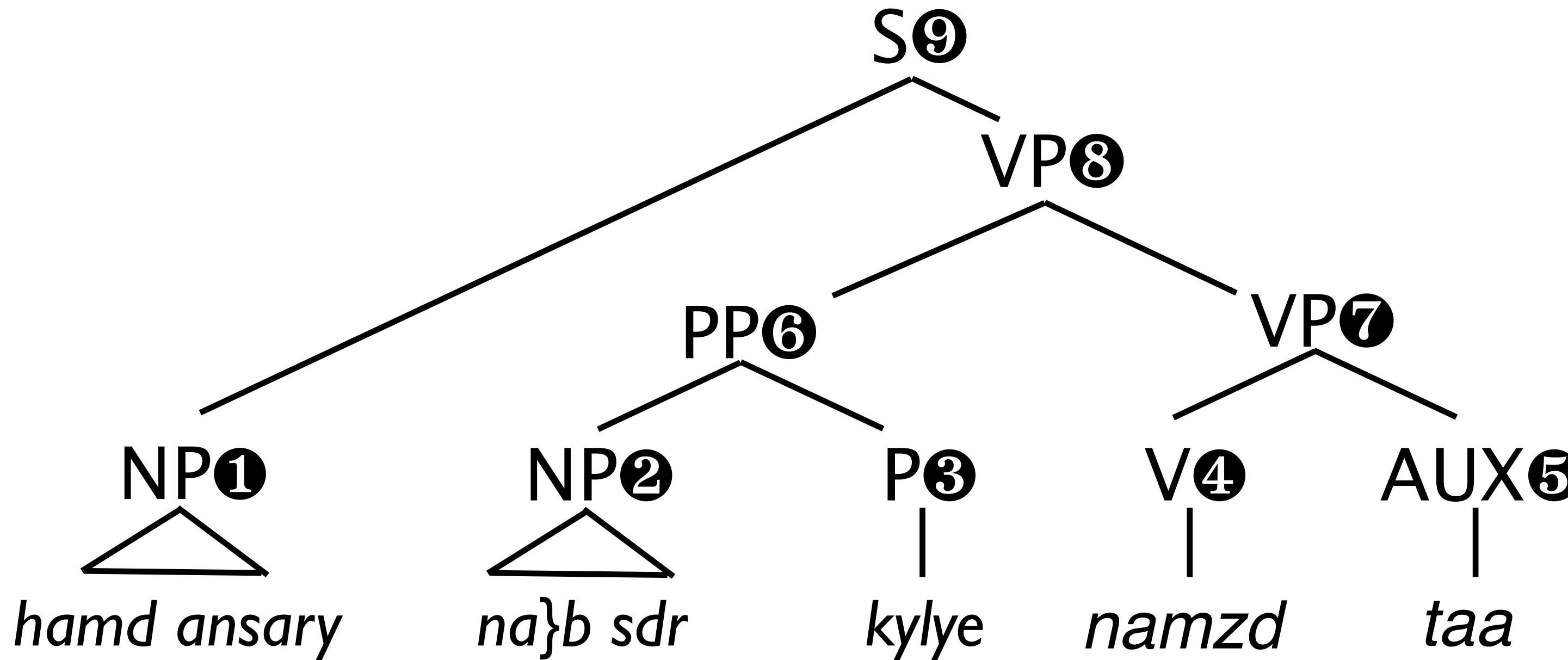
V④  
|  
*nominated*

AUX⑤  
|  
*was*

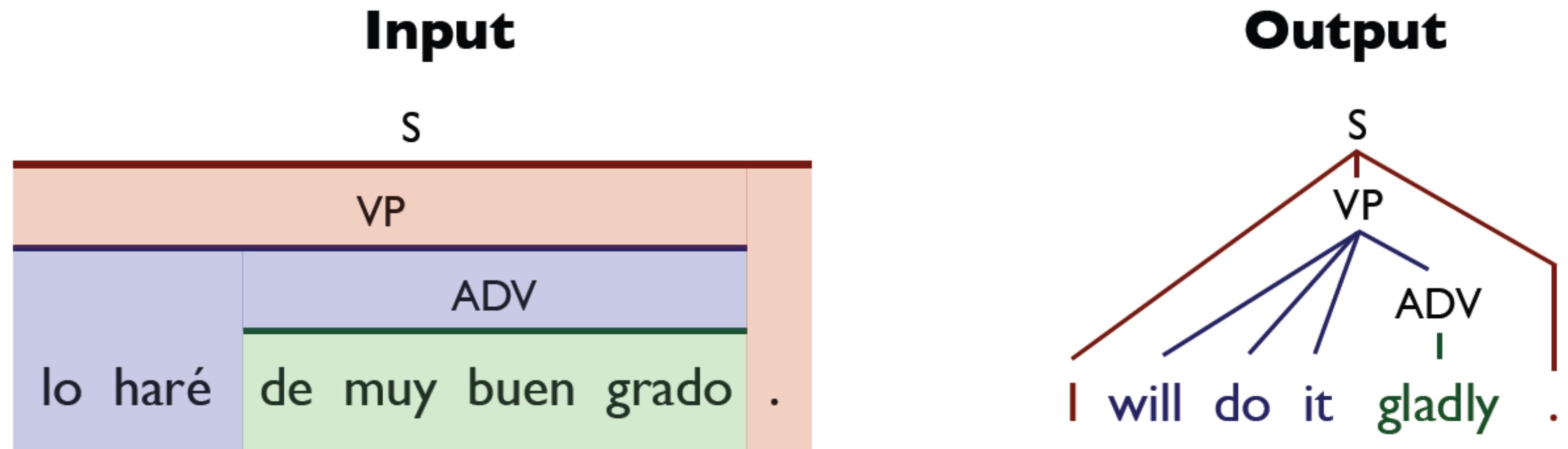








# Syntactic MT



- ▶ Use lexicalized rules, look like “syntactic phrases”

## Grammar

$$S \rightarrow \langle VP . ; I VP . \rangle \quad \text{OR} \quad S \rightarrow \langle VP . ; you VP . \rangle$$
$$VP \rightarrow \langle lo\ haré\ ADV ; will\ do\ it\ ADV \rangle$$
$$S \rightarrow \langle lo\ haré\ ADV . ; I\ will\ do\ it\ ADV . \rangle$$
$$ADV \rightarrow \langle de\ muy\ buen\ grado\ ; gladly \rangle$$

- ▶ Leads to HUGE grammars, parsing is slow

# Joshua

---

- ▶ Toolkit for syntactic machine translation due to many researchers at JHU (Weese, Ganitkevitch, Callison-Burch, Post, Lopez, ...)
- ▶ Joshua implements synchronized grammar extraction (Thrax!), parsing, language modeling, pruning , plus \*a ton\* more stuff
- ▶ Joshua uses two types of SCFG: Hiero grammar (Chiang, 2007), SAMT grammar (Zollmann & Venugopal, 2007)

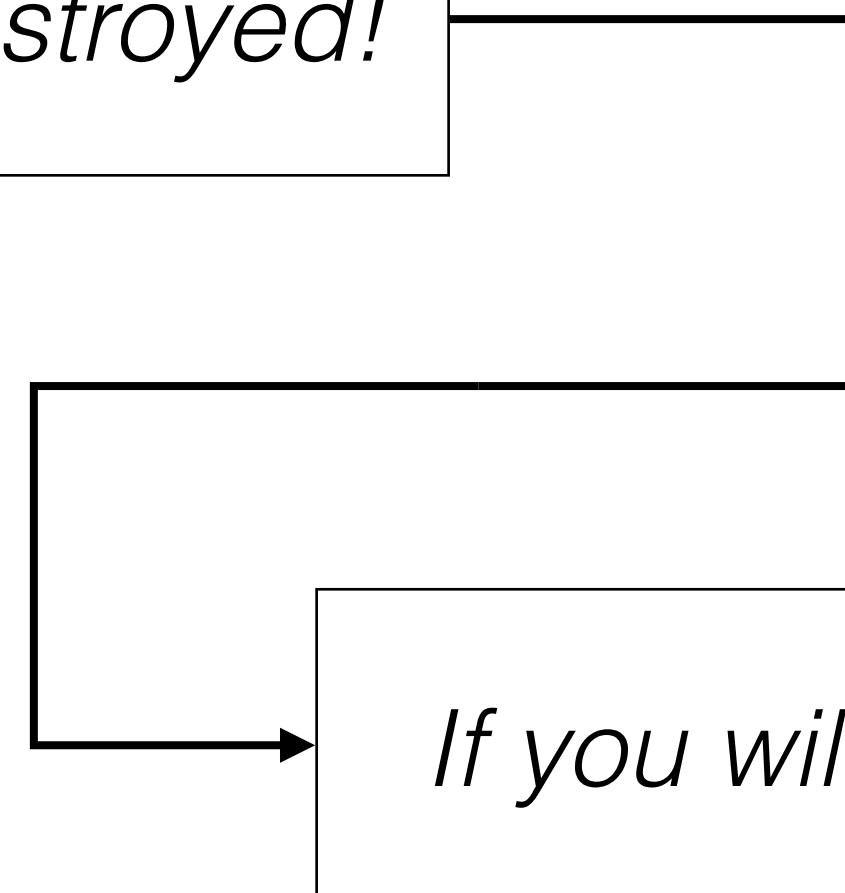
# Case Studies: Monolingual MT

# Style Transfer

---



*If you will not be turned, you will be destroyed!*



- Applied phrase-based MT (Moses Toolkit) to Shakespearean bitext

# Text Simplification

---

*Slightly more fourth-graders nationwide are reading proficiently compared with a decade ago, but only a third of them are now reading well, according to a new report.*



**transform**

*Most fourth-graders are better readers than they were 10 years ago.  
But few of them can actually read well.*

# Text Simplification

---

*Slightly more fourth-graders nationwide are reading proficiently compared with a decade ago, but only a third of them are now reading well, according to a new report.*



**transform**

*Most fourth-graders are better readers than they were 10 years ago.  
But few of them can actually read well.*

# Text Simplification

**Large-scale Paraphrases**  
(lexical, phrasal, syntactic)

**Tuning Data**  
(crowdsourced multi-references)



**Feature Functions**  
(readability, language modeling, etc.)

**Objective Function**

(Xu et al., 2016)

$$SARI = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$
$$p_{add}(n) = \frac{\sum_{g \in O} \min(\#_g(O \cap \bar{I}), \#_g(R))}{\sum_{g \in O} \#_g(O \cap \bar{I})}$$
$$r_{add}(n) = \frac{\sum_{g \in O} \min(\#_g(O \cap \bar{I}), \#_g(R))}{\sum_{g \in O} \#_g(R \cap \bar{I})}$$

**Pairwise Ranking Optimization**

$$\begin{aligned} g(i, j) > g(i, j') &\Leftrightarrow h_w(i, j) > h_w(i, j') \\ &\Leftrightarrow h_w(i, j) - h_w(i, j') > 0 \\ &\Leftrightarrow w \cdot x(i, j) - w \cdot x(i, j') > 0 \\ &\Leftrightarrow w \cdot (x(i, j) - x(i, j')) > 0 \end{aligned}$$

- Implemented by modifying 4 major components of syntax-based MT (Joshua Toolkit)

# Takeaways

---

- ▶ Phrase-based systems consist of 3 pieces: aligner, language model, decoder
  - ▶ HMMs work well for alignment
  - ▶ N-gram language models are scalable and historically worked well
  - ▶ Decoder requires searching through a complex state space
- ▶ Lots of system variants incorporating syntax
- ▶ Next week: neural MT