

# Machine Translation



Dan Klein, John DeNero  
UC Berkeley

# Translation Task

---

- Text as input & text as output.
- Input & output have roughly the same information content.
- Output is more predictable than a language modeling task.
- Lots of naturally occurring examples (but not much metadata).

# Translation Examples

## English-German News Test 2013 (a standard dev set)

---

Republican leaders justified their policy by the need to combat electoral fraud.

Die	Führungskräfte	der	Republikaner
The	Executives	of the	republican
rechtfertigen	ihre	Politik	mit der
justify	your	politics	With of the
Notwendigkeit	,	den Wahlbetrug	zu
need	,	the election fraud	to
bekämpfen	.		
fight	.		

## Variety in Human-Generated Translations

---

An asteroid large enough to destroy a mid-size city brushed the Earth within a short distance of 463,000 km without being detected in advance. Astronomers did not know the event until four days later. About 50 meters in diameter, the asteroid came from the direction of the sun, making it very difficult for astronomers to discover it.

An asteroid, large enough to flatten an average city, brushed past the Earth within a short range of 463,000 kilometers, but was not discovered in time. It was four days after the close shave could astronomers tell about it. This asteroid, about 50 meters in diameter, was flying from the direction of the sun, thus astronomers could hardly detect it.

An asteroid big enough to ruin a mid-sized city passed by in a close range of 463,000 kilometres off Earth without being noticed in advance. Astronomers learned of the event four days later. The asteroid, about 50 metres in diameter, came in the direction of Sun, which made it hard for astronomers to discover.

## Variety in Machine Translations

A small planet, whose is as big as could destroy a middle sized city, passed by the earth with a distance of 463 thousand kilometers. This was not found in advance. The astronomers got to know this incident 4 days later. This small planet is 50m in diameter. The astronomers are hard to find it for it comes from the direction of sun.

Human-generated reference translation

A volume enough to destroy a medium city small planet is big, flit earth within 463,000 kilometres of close however were not in advance discovered, astronomer just knew this matter after four days. This small planet diameter is about 50 metre, from the direction at sun, therefore astronomer very hard to discovers it.

A commercial system from 2002

An asteroid that was large enough to destroy a medium-sized city, swept across the earth at a short distance of 463,000 kilometers, but was not detected early. Astronomers learned about it four days later. The asteroid is about 50 meters in diameter and comes from the direction of the sun, making it difficult for astronomers to spot it.

Google Translate, 2020

# Evaluation

# BLEU Score

BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty (harshly penalizes translations shorter than the reference).

$$\text{Matched}_i = \sum_{t_i} \min \left\{ C_h(t_i), \max_j C_j(t_i) \right\}$$

If "of the" appears twice in hypothesis  $h$  but only at most once in a reference, then only the first is "correct"

$$P_i = \frac{\text{Matched}_i}{H_i}$$

"Clipped" precision of  $n$ -gram tokens

$$B = \exp \left\{ \min \left( 0, \frac{n - L}{n} \right) \right\}$$

Brevity penalty only matters if the hypothesis **corpus** is shorter than the shortest reference.

$$\text{BLUE} = B \left( \prod_{i=1}^4 P_i \right)^{\frac{1}{4}}$$

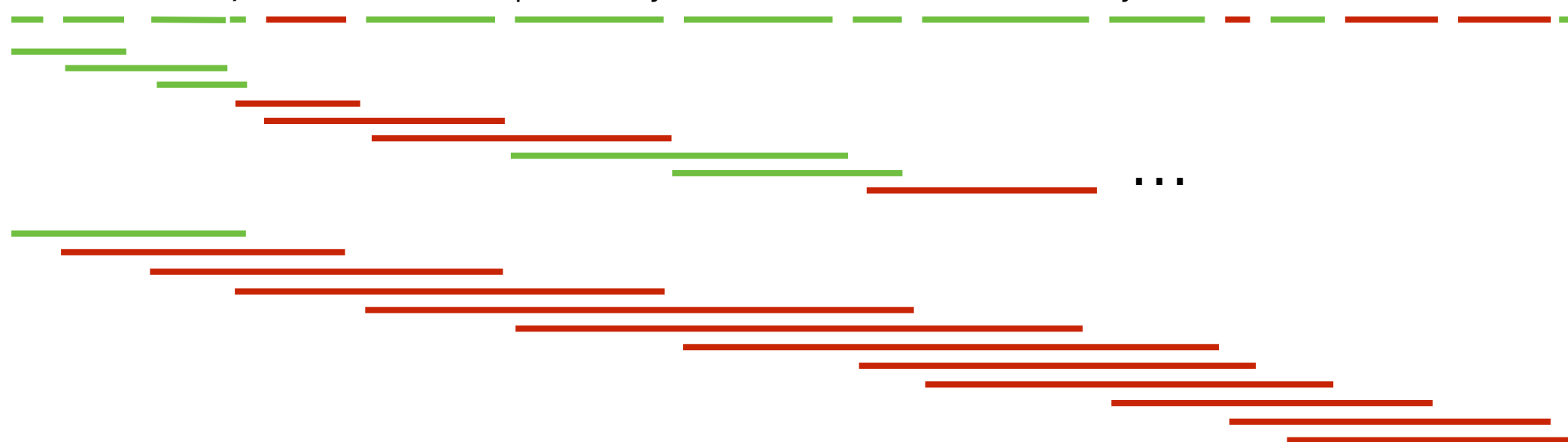
BLUE is a mean of clipped precisions, scaled down by the brevity penalty.



# Evaluation with BLEU

In this sense, the measures will partially undermine the American democratic system.

In this sense, these measures partially undermine the democratic system of the United States.



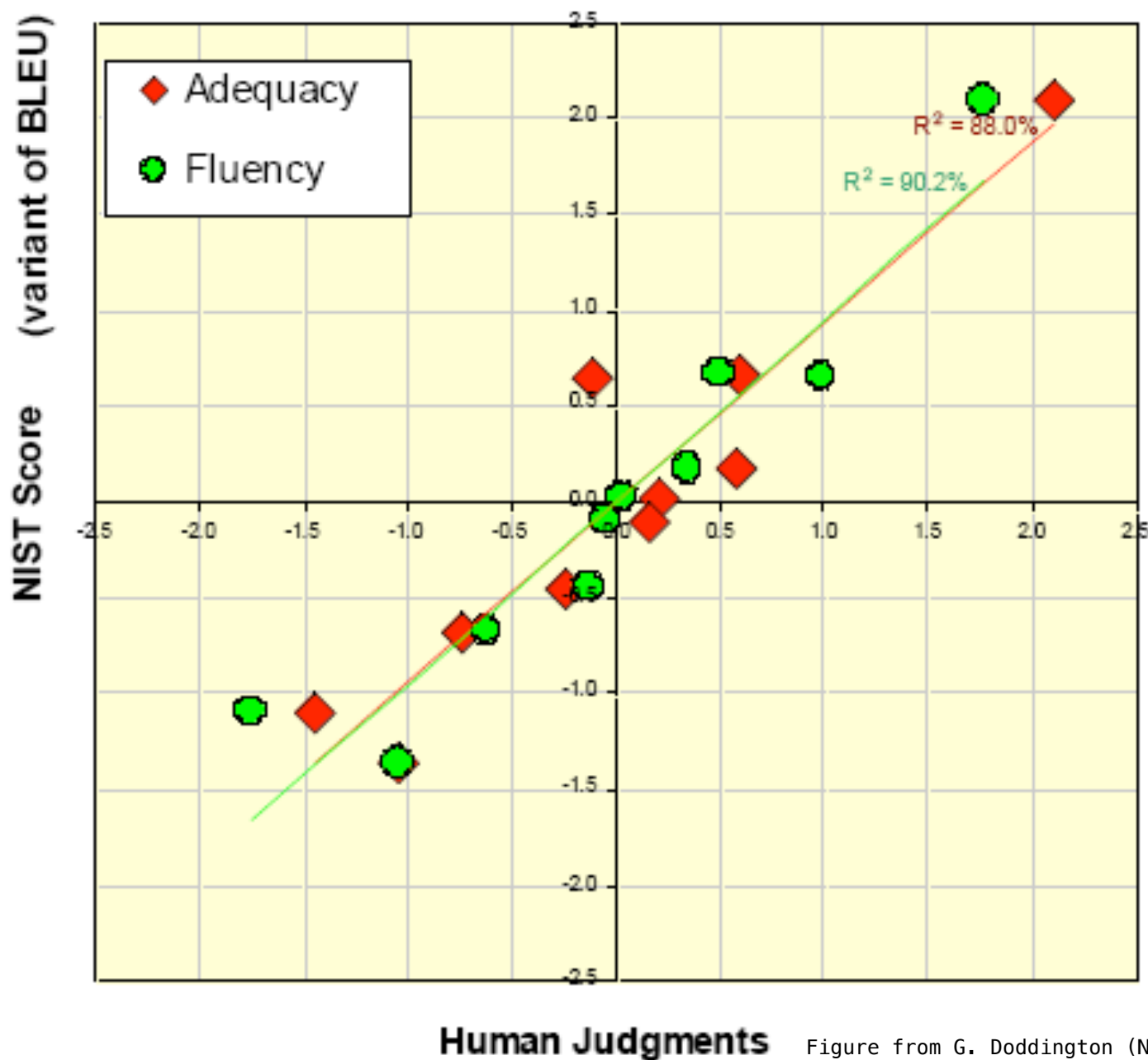
BLEU = 26.52, 75.0/40.0/21.4/7.7 (BP=1.000, ratio=1.143, hyp\_len=16, ref\_len=14)

# Corpus BLEU Correlations with Average Human Judgments

These are ecological correlations over multiple segments; segment-level BLEU scores are noisy.

Commercial machine translation providers seem to all perform human evaluations of some sort.

(Ma et al., 2019)  
Results of the WMT19 Metrics Shared Task:  
Segment-Level and Strong MT Systems  
Pose Big Challenges



Human Judgments

Figure from G. Doddington (NIST)

# Human Evaluations

---

## **Direct assessment:** adequacy & fluency

- Monolingual: Ask humans to compare machine translation to a human-generated reference. (Easier to source annotators)
- Bilingual: Ask humans to compare machine translation to the source sentence that was translated. (Compares to human quality)
- Annotators can assess segments (sentences) or whole documents.
- Segments can be assessed with or without document context.

## **Ranking assessment:**

- Raters are presented with 2 or more translations.
- A human-generated reference may be provided, along with the source.
- "In a pairwise ranking experiment, human raters assessing adequacy and fluency show a stronger preference for human over machine translation when evaluating documents as compared to isolated sentences." (Laubli et al., 2018)

## **Editing assessment:** How many edits required to reach human quality

# Translationese and Evaluation

---

Translated text can: (Baker et al., 1993; Graham et al., 2019)

- be more explicit than the original source
- be less ambiguous
- be simplified (lexical, syntactically and stylistically)
- display a preference for conventional grammaticality
- avoid repetition
- exaggerate target language features
- display features of the source language

"If we consider only original source text (i.e. not translated from another language, or translationese), then we find evidence showing that human parity has not been achieved."

(Toral et al., 2018)

# WMT 2019 Evaluation

---

2019 segment-in-context direct assessment (Barrault et al, 2019):

- ✓ German to English: many systems are tied with human performance;
- × English to Chinese: all systems are outperformed by the human translator;
- × English to Czech: all systems are outperformed by the human translator;
- × English to Finnish: all systems are outperformed by the human translator;
- ✓ English to German: Facebook-FAIR achieves super-human translation performance; several systems are tied with human performance;
- × English to Gujarati: all systems are outperformed by the human translator;
- × English to Kazakh: all systems are outperformed by the human translator;
- × English to Lithuanian: all systems are outperformed by the human translator;
- ✓ English to Russian: Facebook-FAIR is tied with human performance.

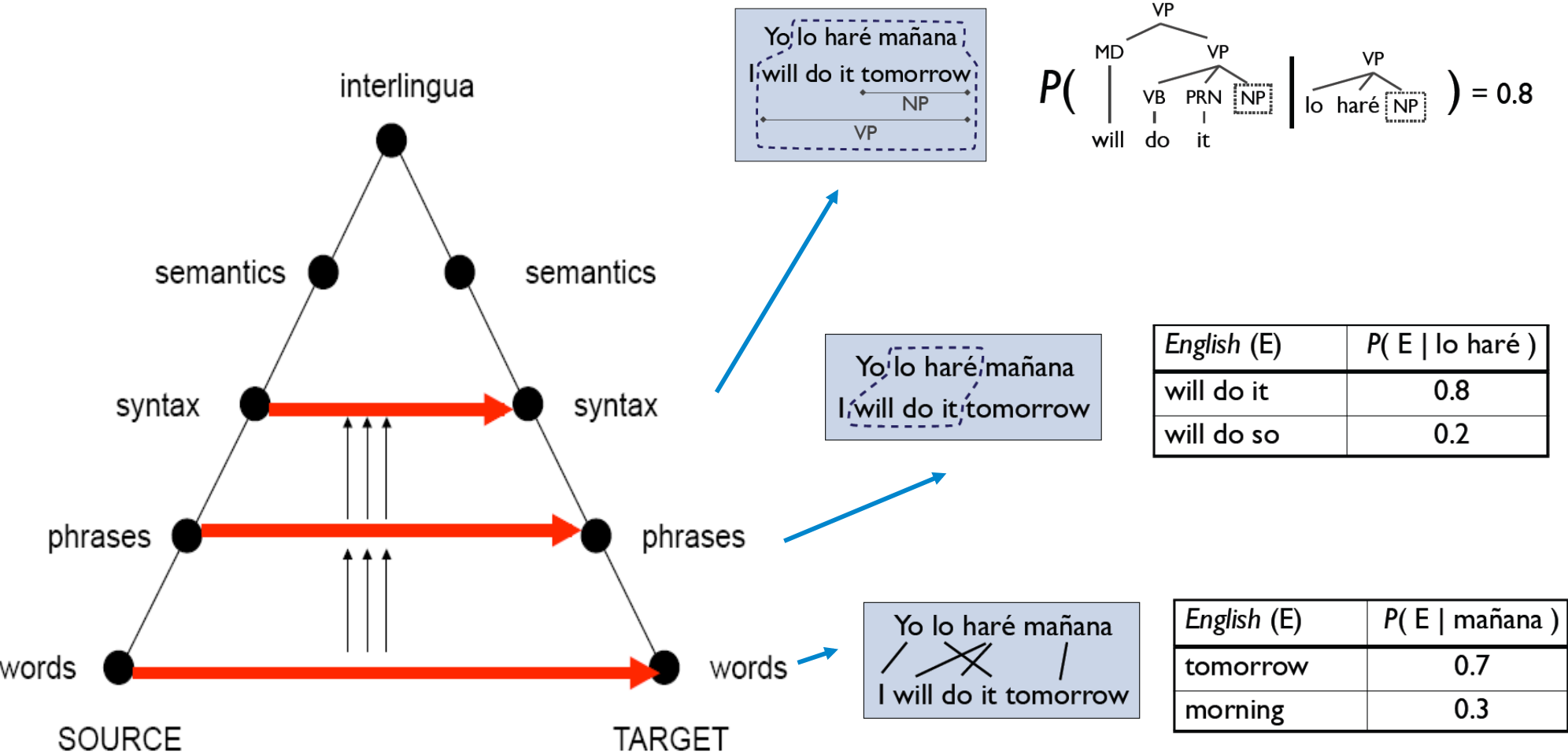
# Statistical Machine Translation (1990 - 2015)



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)

# Levels of Transfer: Vauquois Triangle (1968)





# Data-Driven Machine Translation

*Target language corpus gives examples of well-formed sentences*

I will get to it later

See you later

He will do it

*Parallel corpus gives translation examples*

I will do it gladly

Yo lo haré de muy buen grado

You will see later

Después lo veras

*Machine translation system:*

**Source language**

Yo lo haré después

NOVEL SENTENCE

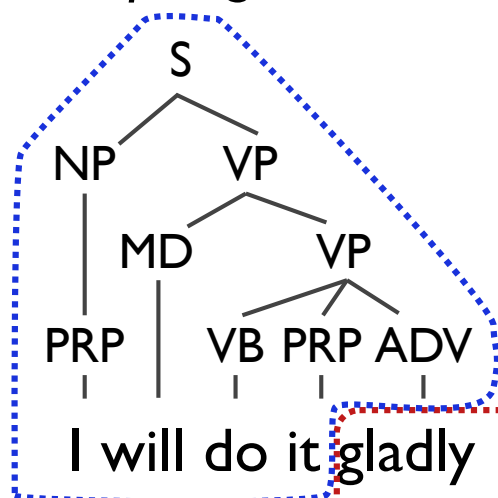
Model of  
translation

**Target language**

I will do it later

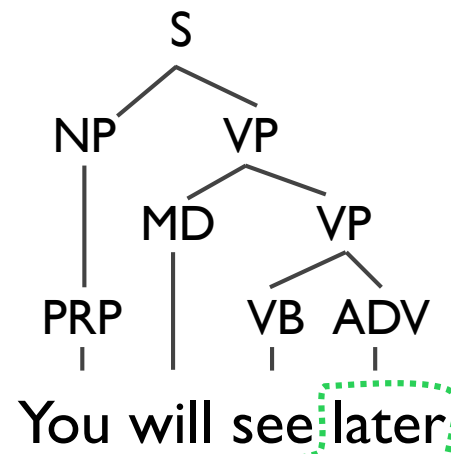
# Stitching Together Fragments

*Parallel corpus gives translation examples*



I will do it gladly

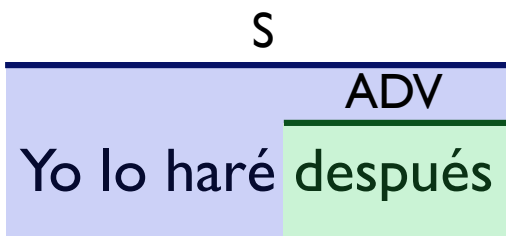
Yo lo haré de muy buen grado



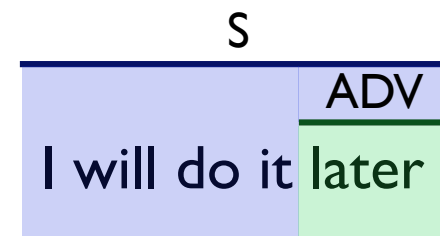
You will see later

Después lo veras

*Machine translation system:*



Model of translation



## Evolution of the Noisy Channel Model

---

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

$$\max_e P(e|f) = \max_e P(f|e) \cdot P(e)$$

$$P(e|f) \propto P(f|e)^{\phi_{\text{tm}}} \cdot P(e)^{\phi_{\text{lm}}}$$

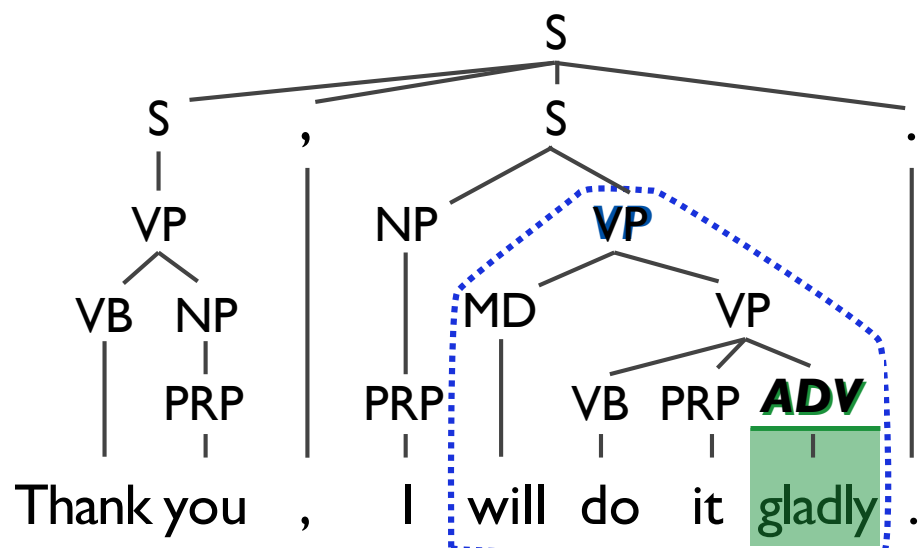
$$P(e|f) \propto \exp \left\{ \sum_i w_i \cdot f_i(e, f) \right\}$$

Chosen to minimize loss

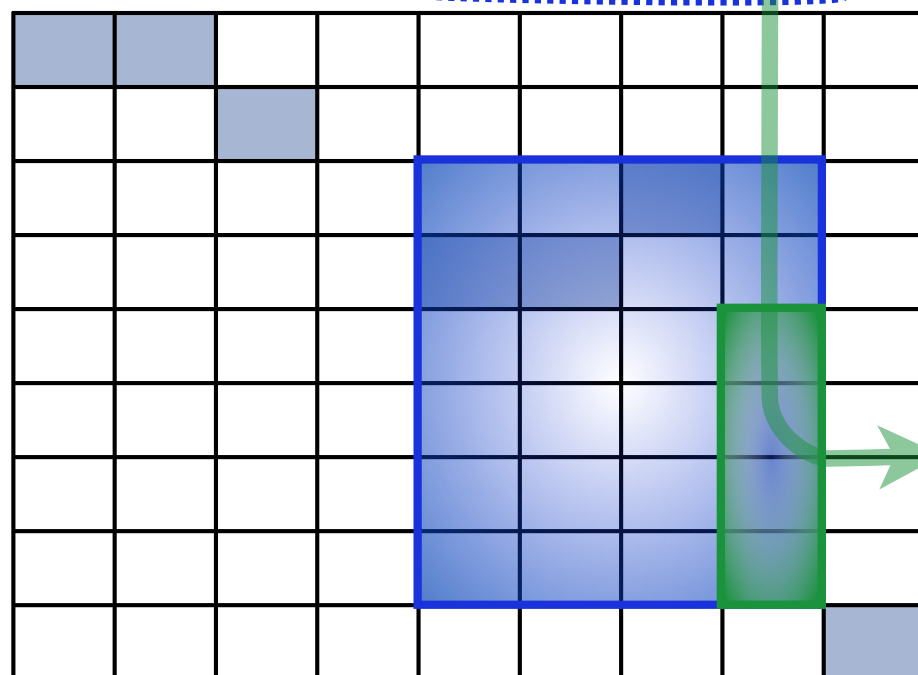
E.g.,  $\log P(e)$

# Word Alignment

## Extracting Translation Rules



Frequency statistics on these rules serve as features in a translation model



# Gracias

lo

haré

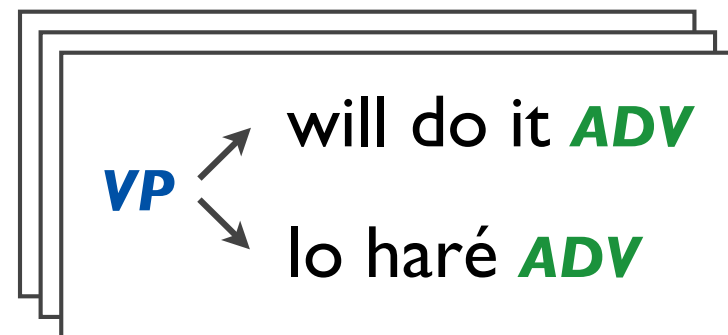
de

muv

buen

grado

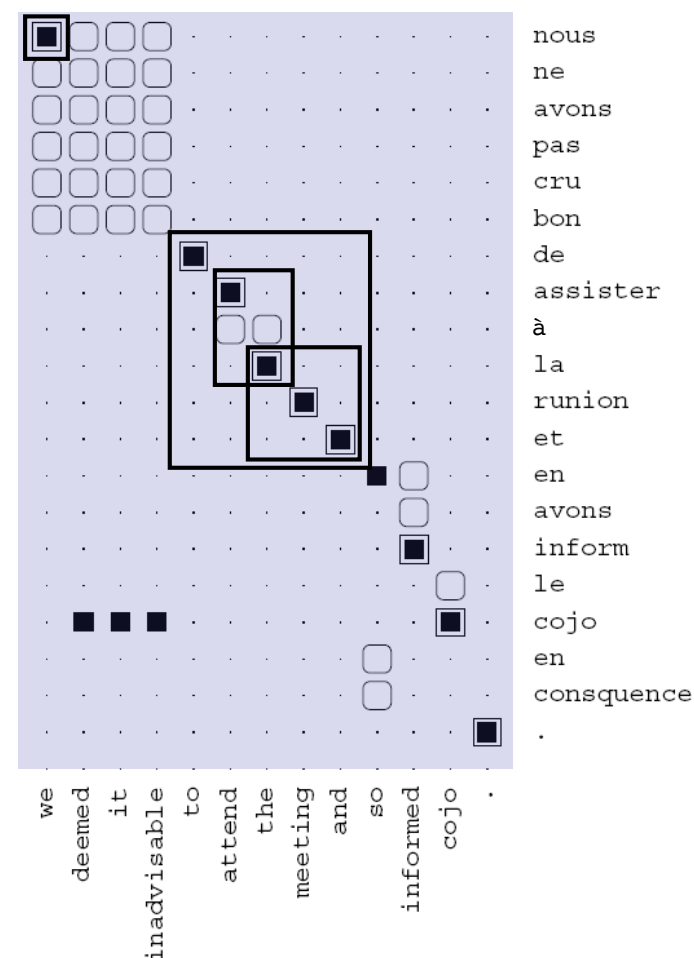
ADV



# Counting Aligned Phrases

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  
...

- Relative frequencies are the most important features in a phrase-based or syntax-based model.
- Scoring a phrase under a lexical model is the second most important feature.
- Estimation does not involve choosing among segmentations of a sentence into phrases.



## Interlude: Lexical Translation Models

## What's Next?

---

Searching over the space of translations

Neural models: attention and the transformer architecture

Tricks of the trade: back-translation, knowledge distillation, subword models, and coverage vectors