

# Evaluating translation quality

## Machine Translation Lecture 9

**Instructor: Chris Callison-Burch**  
**TAs: Mitchell Stern, Justin Chiu**

**Website: [mt-class.org/penn](http://mt-class.org/penn)**



# Goals for this lecture

- Understanding advantages of human versus automatic evaluation
- Details of BLEU
- How to validate automatic evaluation metrics
- What makes a good {manual / automatic} evaluation?

# Evaluating MT Quality

- Why do we want to do it?
  - ▶ Want to rank systems
  - ▶ Want to evaluate incremental changes
  - ▶ What to make scientific claims
- How not to do it
  - ▶ “Back translation”
  - ▶ The vodka is not good

# Human Evaluation of MT v. Automatic Evaluation

- Human evaluation is
  - ▶ Ultimately what we're interested in, but
  - ▶ Very time consuming
  - ▶ Not re-usable
- Automatic evaluation is
  - ▶ Cheap and reusable, but
  - ▶ Not necessarily reliable

# Manual Evaluation

**Source:** Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

**Reference:** These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	<div><div><div></div><div>1</div><div>Best</div></div><div><div></div><div>2</div><div></div></div><div><div></div><div>3</div><div></div></div><div><div></div><div>4</div><div></div></div><div><div><div></div></div><div>5</div><div>Worst</div></div></div>

# Goals for Automatic Evaluation

- No cost evaluation for incremental changes
- Ability to rank systems
- Ability to identify which sentences we're doing poorly on, and categorize errors
- Correlation with human judgments
- Interpretability of the score

# Methodology

- Comparison against reference translations
- Intuition: closer we get to human translations, the better we're doing
- Could use WER like in speech recognition

# Word Error Rate

- Levenshtein Distance (also "edit distance")
- Minimum number of insertions, substitutions, and deletions needed to transform one string into another
- Useful measure in speech recognition
  - ▶ This shows how easy it is to recognize speech
  - ▶ This shows how easy it is to wreck a nice beach



# Problems with using WER for translation?

- (discuss with your neighbor)

# Problems with WER

- Unlike speech recognition we don't have the assumption of
  - ▶ exact match against the reference
- In machine translation there can be many possible (and equally valid) ways of translating a sentence
  - ▶ This shows how easy it is to recognize speech
  - ▶ It illustrates how simple it is to transcribe the spoken word

# Problems with WER

- Unlike speech recognition we don't have the assumption of
  - ▶ linearity
- Clauses can move around, since we're not doing transcription
  - ▶ This shows how easy it is to recognize speech
  - ▶ It is easy to recognize speech, as this shows
  - ▶ This shows that recognizing speech is easy

# Solutions?

- (Talk to your neighbor)

# Solutions

- Compare against lots of test sentences
- Use multiple reference translations for each test sentence
- Look for phrase / n-gram matches, allow movement

# BLEU

- BiLingual Evaluation Understudy
- Uses multiple reference translations
- Look for n-grams that occur anywhere in the sentence

# Multiple references

<b>Ref 1</b>	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
<b>Ref 2</b>	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
<b>Ref 3</b>	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
<b>Ref 4</b>	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

# n-gram precision

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

- BLEU modifies this precision to eliminate repetitions that occur across sentences.



# Modified precision

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him <b>to Miami</b> , Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him <b>to Miami</b> , Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him <b>to Miami</b> in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him <b>to Miami</b> in Florida.

“to Miami” can only be counted as correct once

<b>Ref 1</b>	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
<b>Ref 2</b>	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
<b>Ref 3</b>	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
<b>Ref 4</b>	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

<b>Hyp</b>	appeared calm when he was taken to the American plane, which will to Miami, Florida.
------------	--

**American, Florida, Miami, Orejuela,**  
**appeared,** as, being, calm, carry, escorted, he,  
 him, in, led, **plane,** quite, seemed, take, that,  
 the, **to, to,** to, **was** , was, **which,** while, **will,**  
 would, ,, .

I-gram precision = 15/18

Hyp	<b>appeared calm when he was taken to the American plane , which will to Miami , Florida .</b>
-----	--

American plane, Florida ., Miami , Miami  
 in, Orejuela appeared, Orejuela seemed,  
**appeared calm**, as he, being escorted, being  
 led, calm as, calm while, carry him, escorted  
 to, **he was**, him to, in Florida, led to, plane  
 that, plane which, quite calm, seemed quite,  
 take him, that was, that would, **the American**,  
 the plane, **to Miami**, to carry, **to the**, was  
 being, was led, was to, **which will**, while  
 being, will take, would take, , Florida

2-gram precision = 10/17

Hyp	<b>appeared calm when he was taken to the American plane , which will to Miami , Florida .</b>
-----	--

# n-gram precision

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.
-----	--

1-gram precision =  $15/18 = .83$

2-gram precision =  $10/17 = .59$

3-gram precision =  $5/16 = .31$

4-gram precision =  $3/15 = .20$

- Geometric average

$$(0.83 * 0.59 * 0.31 * 0.2)^{(1/4)} = 0.417$$

or equivalently

$$\exp(\ln .83 + \ln .59 + \ln .31 + \ln .2/4) = 0.417$$

<b>Ref 1</b>	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
<b>Ref 2</b>	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
<b>Ref 3</b>	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
<b>Ref 4</b>	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

<b>Hyp</b>	to the American plane
------------	-----------------------

# Better?

Hyp	to the American plane
-----	-----------------------

1-gram precision =  $4/4 = 1.0$

2-gram precision =  $3/3 = 1.0$

3-gram precision =  $2/2 = 1.0$

4-gram precision =  $1/1 = 1.0$

$$\exp(\ln 1 + \ln 1 + \ln 1 + \ln 1) = 1$$

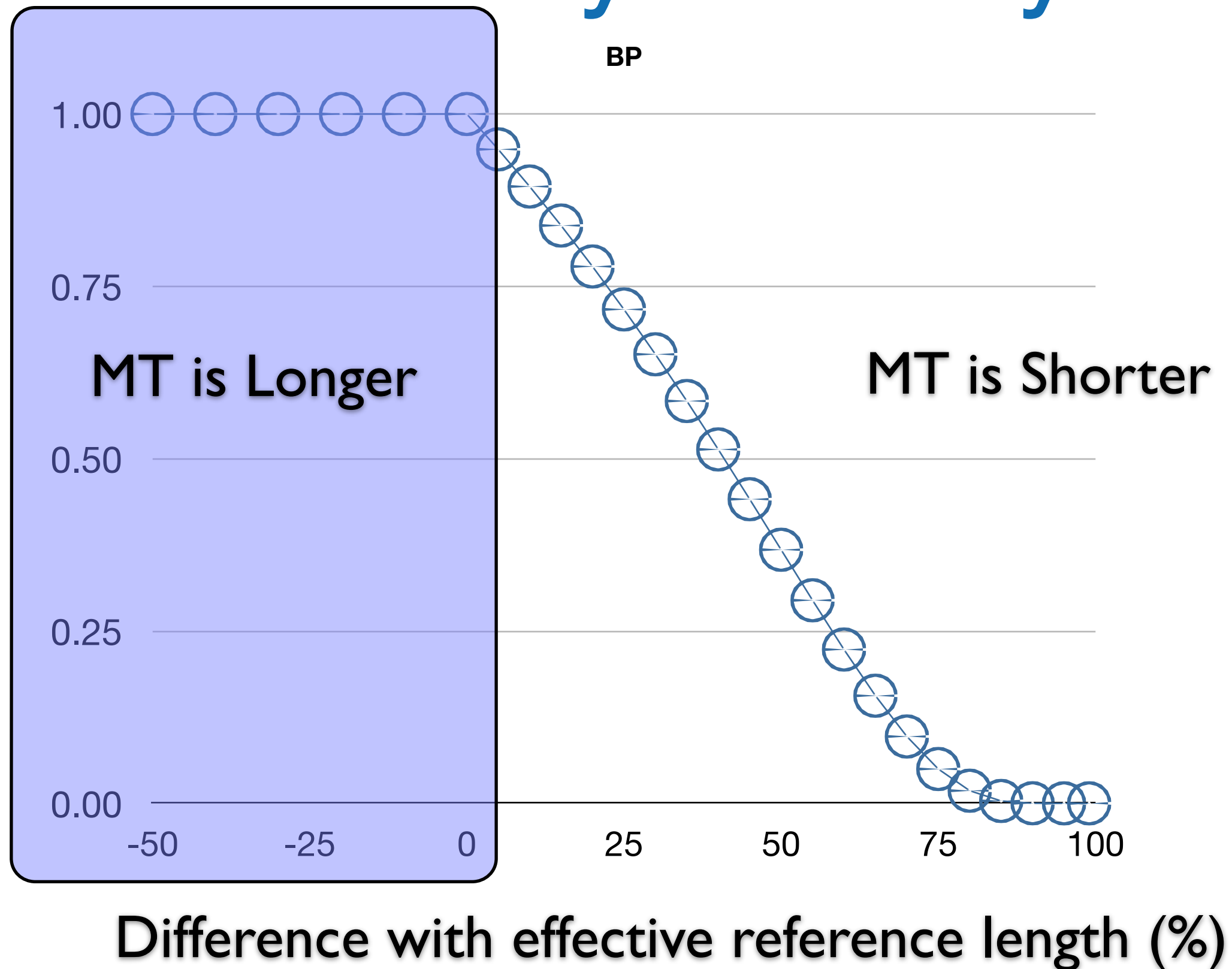
# Brevity Penalty

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

- $c$  is the length of the corpus of hypothesis translations
- $r$  is the effective reference corpus length
- The effective reference corpus length is the sum of the single reference translation from each set that is closest to the hypothesis translation.



# Brevity Penalty



<b>Ref I</b>	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. <b><math>r = 20</math></b>
<b>Hyp</b>	appeared calm when he was taken to the American plane, which will to Miami, Florida. <b><math>c = 18</math></b>

$$BP = \exp(1 - (20/18)) = 0.89$$

<b>Ref I</b>	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. <b><math>r = 20</math></b>
<b>Hyp</b>	to the American plane <b><math>c = 4</math></b>

$$BP = \exp(1 - (20/4)) = 0.02$$

# BLEU

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

- Geometric average of the n-gram precisions
- Optionally weight them with  $w$
- Multiplied by the brevity penalty

# BLEU

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.
-----	--

$$\exp(1-(20/18)) * \exp((\ln .83 + \ln .59 + \ln .31 + \ln .2)/4) = 0.374$$

Hyp	to the American plane
-----	-----------------------

$$\exp(1-(20/4)) * \exp((\ln 1 + \ln 1 + \ln 1 + \ln 1)/4) = 0.018$$

# Problems with BLEU

- (Discuss with your neighbor)

# Problems with BLEU

- **Synonyms and paraphrases** are only handled if they are in the set of multiple reference translations
- The scores for **words** are **equally weighted** so missing out on content-bearing material brings no additional penalty.
- The brevity penalty is a stop-gap measure to compensate for the fairly serious problem of not being able to calculate **recall**.

# More Metrics

- WER - word error rate
- PI-WER - position independent WER
- METEOR - **M**etric for **E**valuation of **T**ranslation with **E**xplicit **O**Rdering
- TERp - Translation **E**dit **R**ate plus

# Even More Metrics

Metric IDs	Participant
AMBER, AMBER-NL, AMBER-IT	National Research Council Canada
F15, F15G3	Koç University (Bicici and Yuret, 2011)
METEOR-1.3-ADQ, METEOR-1.3-RANK	Carnegie Mellon University (Denkowski and Lavie, 2011)
MTERATER, MTERATER-PLUS	Columbia / ETS (Parton et al., 2011)
MP4IBM1, MPF, WMPF	DFKI (Popović, 2011; Popović et al., 2011)
PARSECONF	DFKI (Avramidis et al., 2011)
ROSE, ROSE-POS	The University of Sheffield (Song and He, 2011)
TESLA-B, TESLA-F, TESLA-M	National University of Singapore (Rajkumar et al., 2011)
TINE	University of Wolverhampton (Rico and He, 2011)
BLEU	provided baseline (Papineni et al., 2002)
TER	provided baseline (Snover et al., 2006)



# How do we know which metric is best?

- Measure correlation with human judgments
- How do people evaluate MT quality

# Manual Evaluation

**Source:** Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

**Reference:** These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	<div><div><div></div><div>1</div><div>Best</div></div><div><div></div><div>2</div><div></div></div><div><div></div><div>3</div><div></div></div><div><div></div><div>4</div><div></div></div><div><div><div></div></div><div>5</div><div>Worst</div></div></div>
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	<div><div><div></div><div>1</div><div>Best</div></div><div><div></div><div>2</div><div></div></div><div><div><div></div></div><div>3</div><div></div></div><div><div></div><div>4</div><div></div></div><div><div></div><div>5</div><div>Worst</div></div></div>
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	<div><div><div></div><div>1</div><div>Best</div></div><div><div><div></div></div><div>2</div><div></div></div><div><div></div><div>3</div><div></div></div><div><div></div><div>4</div><div></div></div><div><div></div><div>5</div><div>Worst</div></div></div>
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	<div><div><div><div></div></div><div>1</div><div>Best</div></div><div><div></div><div>2</div><div></div></div><div><div></div><div>3</div><div></div></div><div><div></div><div>4</div><div></div></div><div><div></div><div>5</div><div>Worst</div></div></div>
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	<div><div><div></div><div>1</div><div>Best</div></div><div><div></div><div>2</div><div></div></div><div><div></div><div>3</div><div></div></div><div><div><div></div></div><div>4</div><div></div></div><div><div></div><div>5</div><div>Worst</div></div></div>

# 5-point scales

## **Fluency**

How do you judge the fluency of this translation?

5 = Flawless English

4 = Good English

3 = Non-native English

2 = Disfluent English

1 = Incomprehensible

## **Adequacy**

How much of the meaning expressed in the reference translation is also expressed in the hypothesis translation?

5 = All

4 = Most

3 = Much

2 = Little

1 = None

# Reading Comprehension of Machine Translation

- Jones et al (2005) - Measured translation quality by testing English speakers on a Defense Language Proficiency Test for Arabic
- Read the MT output, and assess how many questions were answered correctly
- Nice, intuitive gauge of how good MT quality actually is

## Heather Locklear Arrested for driving under the influence of drugs



The actress Heather Locklear, Amanda of the popular series *Melrose Place*, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to *People* magazine by a spokesman for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses." Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw Ms.

- Why was Heather Locklear arrested?
  - ▶ She was arrested on suspicion of driving under the influence of drugs.
- Why did the bystander call emergency services?
  - ▶ He was concerned for Ms. Locklear's life.
- Where did the witness see her acting abnormally?
  - ▶ Pulling out of parking in Montecito

**Heather Locklear**

*Photo by: Santa Barbara County Sheriff's Department*



## Was arrested actress Heather Locklear because of the driving under the effect of an unknown medicine



The actress Heather Locklear that is known to the Amanda through the role from the series "Melrose Place" was arrested at this weekend in Santa Barbara (Californium) because of the driving under the effect of an unknown medicine. A female witness observed she attempted in quite strange way how to go from their parking space in Montecito, speaker of the traffic police of californium told the warehouse `People'. The female witness told in detail, that Locklear 'pressed `after 16:30 clock accelerator and a lot of noise did when she attempted to move their car towards behind or forward from the parking space, and when it went backwards, she pulled itself together unites Male at their sunglasses'. A little later the female witness that did probably

**Heather Locklear**

Photo by: Santa Barbara County Sheriff's Department

**SPONSORED LINKS**

- Why was Heather Locklear arrested?

Driving while medicated

- Why did the bystander call emergency services?

There was a lot of noise

- Where did the witness see her acting abnormally?

In a parking lot

## . Medikamentes unknown have the effect of a fahrens under actress heather locklear arrested



**Heather Locklear**

Photo by: Santa Barbara County Sheriff's Department

In Santa. One is, melrose place the series of the role of the 'remember the locklear actress the heather this weekend, because of the fahrens Barbara (California) in effect unknown medikamentes arrested People 'magazine. The traffic police California, spokesman for the auszufahren montecito reported in its way from tried parklücke type strange right, you have seen as a witness. . In some Zeitung, as and when they tried to a great deal of 30 p.m., witness the detail of history locklear after 16: that durchdrückte peddle noise and its progress was made parklücke for the car or moving backwards, they had they times of their sonnenbrille '. The first was probably recognised that locklear a nearby road and anhielt, had not, with the witness to the car off

- Why was Heather Locklear arrested?

- Why did the bystander call emergency services?

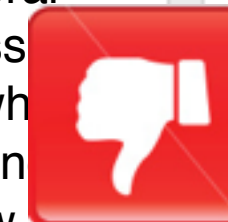
- Where did the witness see her acting abnormally?



## Heather Locklear Arrested for driving under the influence of drugs



The actress Heather Locklear, Amanda of the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from parking in Montecito, as revealed to People magazine by a spokesman for the California Highway Police. The witness stated that around 4.30pm Heather Locklear "hit the accelerator violently, making excessive noise while trying to take her car out from the parking with abrupt and forth maneuvers. While reversing, she passed several times in front of his sunglasses. Shortly after, the witness, who a first time, apparently had not recognized the actress, saw her.



- Why was Heather Locklear arrested?
  - She was arrested on suspicion of driving under the influence of drugs.

Driving under the influence

Driving while medicated

DUI

Driving while using drugs

Medikamentes

**Heather Locklear**

Photo by: Santa Barbara County Sheriff's Department

**SPONSORED LINKS**



## Heather Locklear Arrested for driving



Why was Heather Locklear arrested?

on  
g under  
rugs.

influence

cated

drugs

System	Correct Answers
Reference	94%
Google	80%
RBMT5	77%
Geneva	63%
JHU - Tromble	50%

and forth maneuvers. While reversing, she passed several times in front of his sunglasses. Shortly after, the witness, who a first time, apparently had not recognized the actress, saw me.



Medikamentes

**Heather Locklear**  
Photo by: Santa Barbara County Sheriff's Department

SPONSORED LINKS

# HTER - costs to edit

## Reference translation

The man was on assignment from the Ministry of Defense when he left two highly classified documents on a train to Waterloo.

## Machine translation

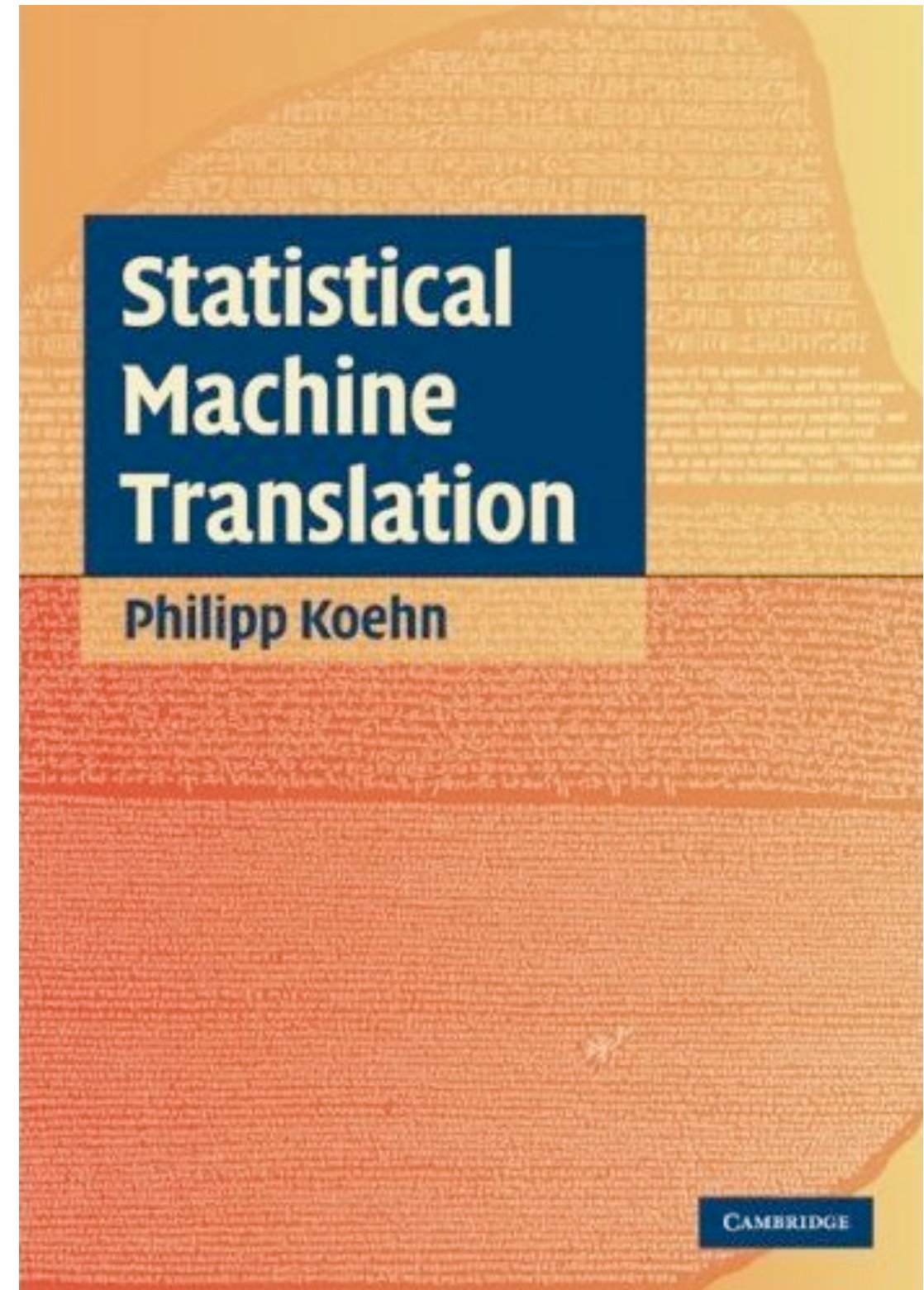
The man was seconded by the Ministry of Defense when he was two extremely confidential documents in a train to Waterloo lost.

## Edited machine translation

The man was **working for** the Ministry of Defense when he **lost** two extremely confidential documents in a train to Waterloo.

# Reading

- Read 8 from the textbook



# Announcements

- HW2 due in 1 week
- Mitchell will be holding a special office hour session tomorrow from noon-1:30 to review the HW2 sample code and show how it relates to the equations