Evaluating translation quality - part 2

February 27, 2014

How do we know which MT evaluation metric is best?

- Measure correlation with human judgments
- **How do people evaluation MT quality?**

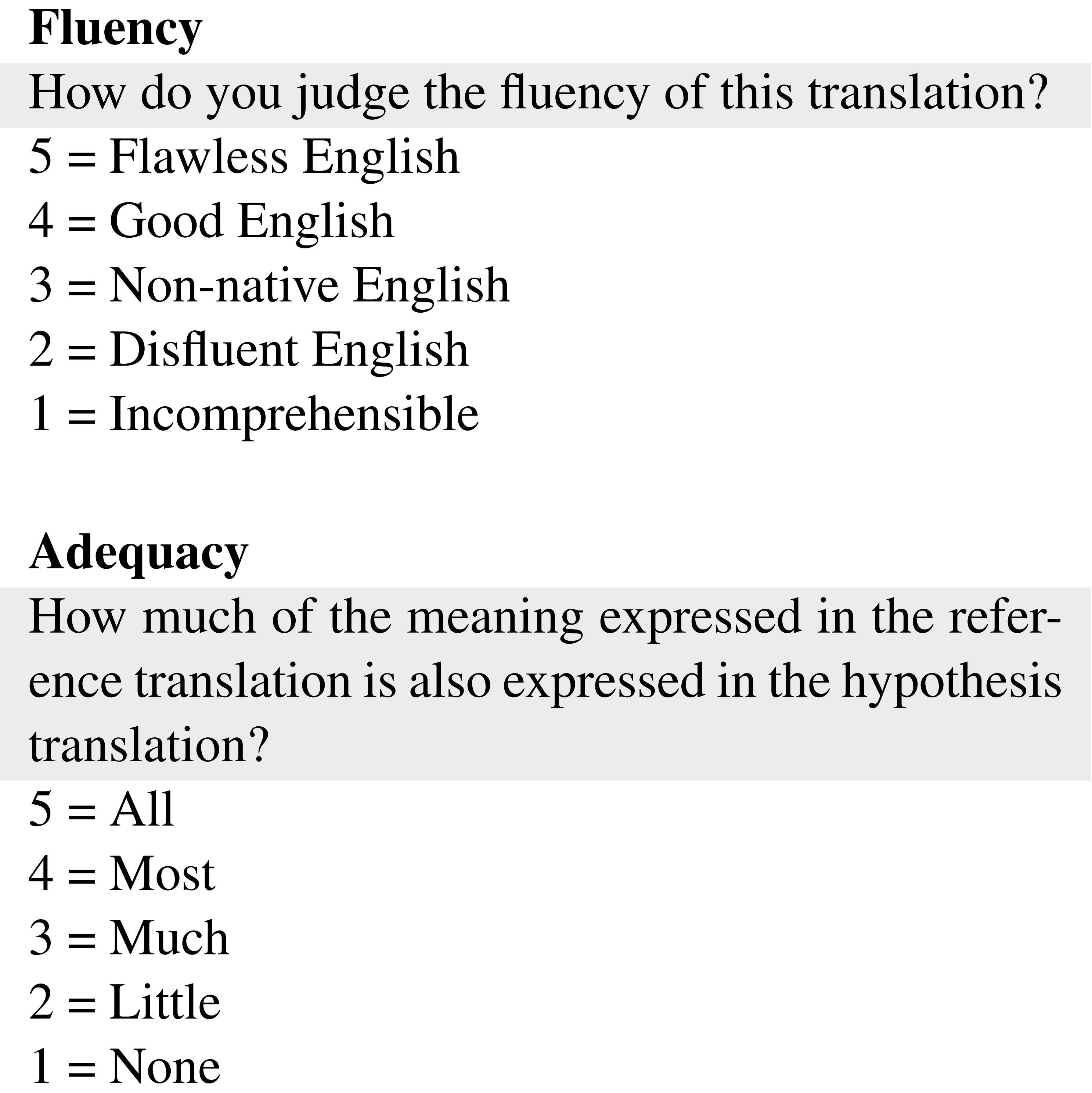
Manual Evaluation

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.	○ 1 Best ○ 2 ○ 3 ○ 4 ● 5 Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ● 3 ○ 4 ○ 5 Worst
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.	○ 1 Best ● 2 ○ 3 ○ 4 ○ 5 Worst
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.	● 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
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.	○ 1 Best ○ 2 ○ 3 ● 4 ○ 5 Worst

5-point scales



Fluency

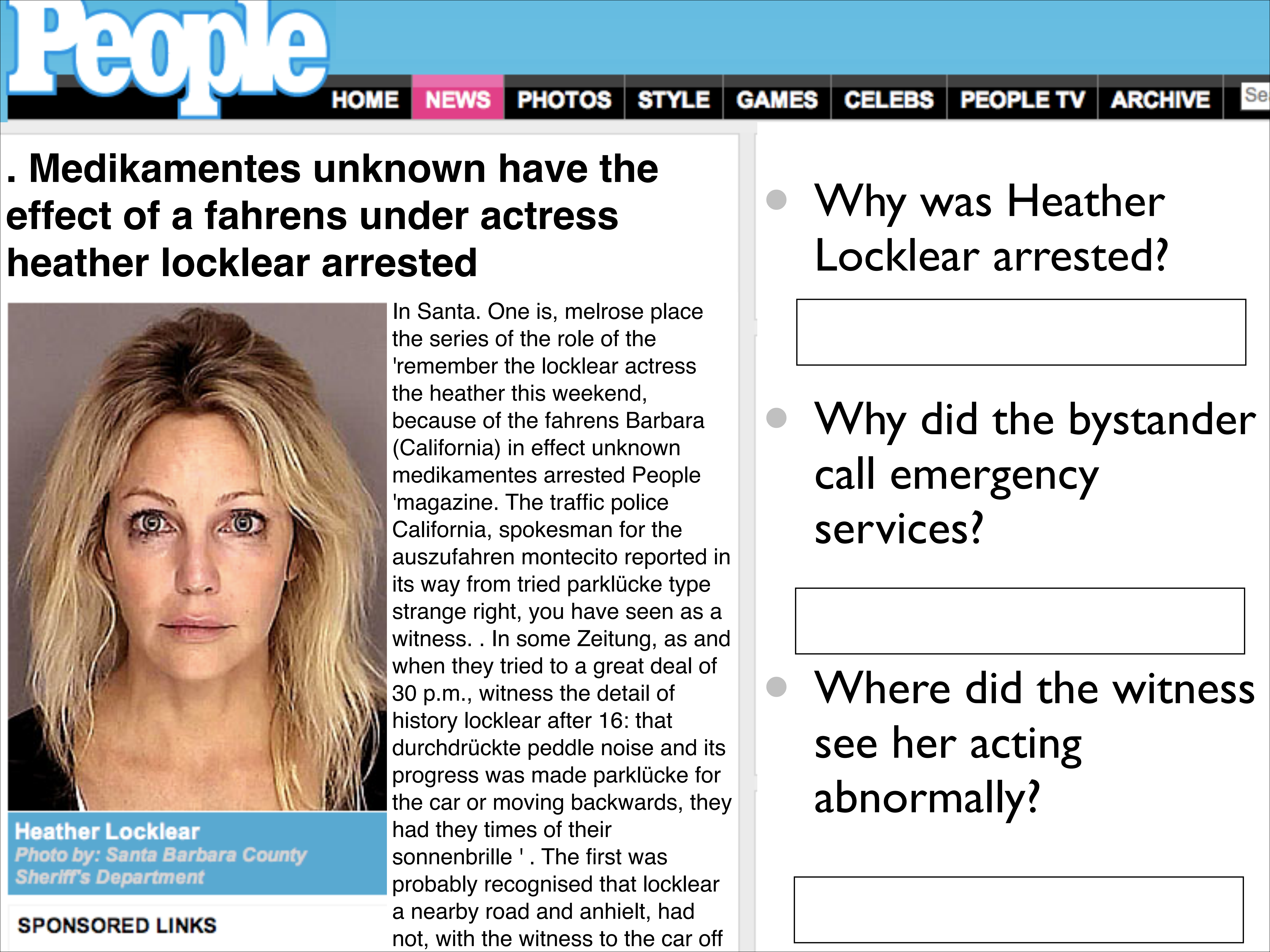
How do you judge the fluency of this translation?

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



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

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

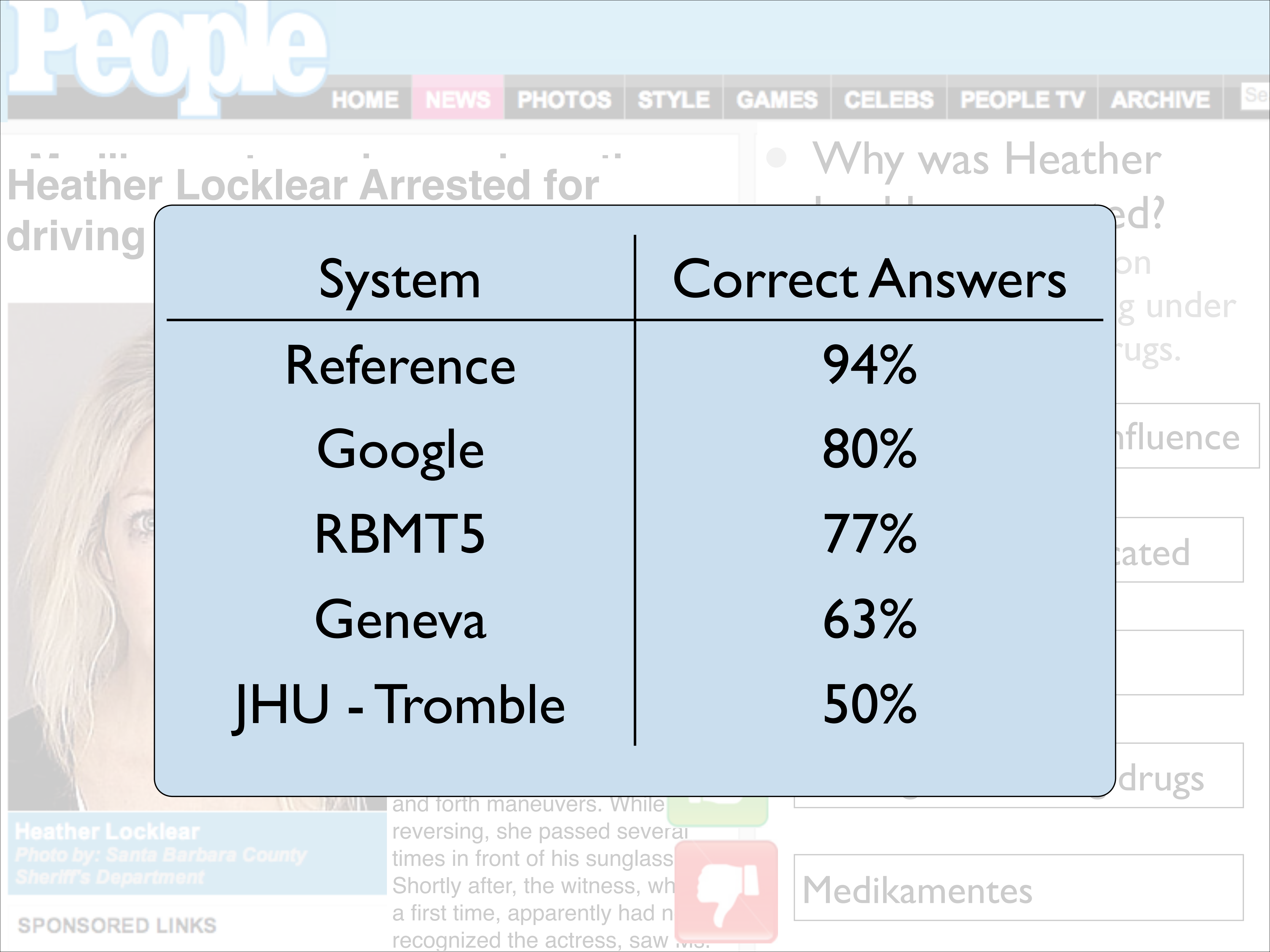
SPONSORED LINKS

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

- Why was Heather

ed?

on

g under

rugs.

nfluence

ated

drugs

Medikamentes

and forth maneuvers. While reversing, she passed several times in front of his sunglass Shortly after, the witness, wh a first time, apparently had n recognized the actress, saw Ms.

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

SPONSORED LINKS

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

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.

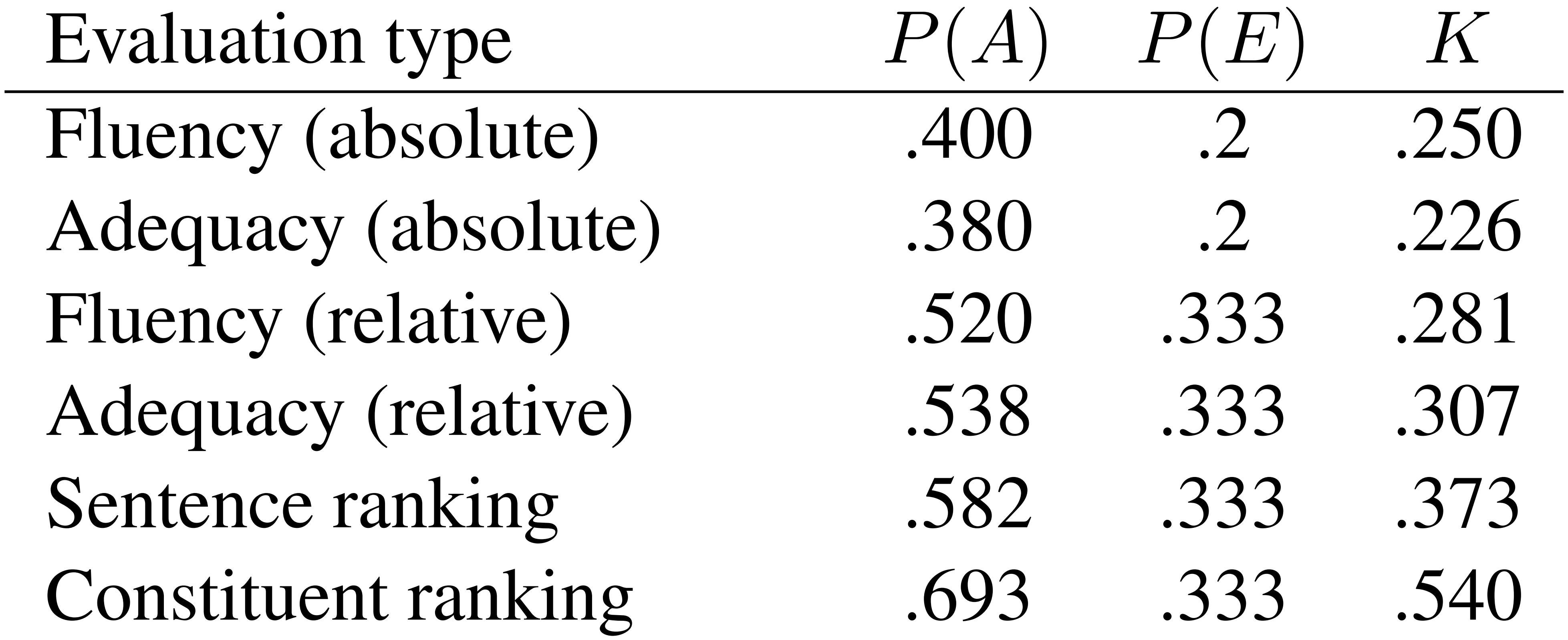
Machine translation

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

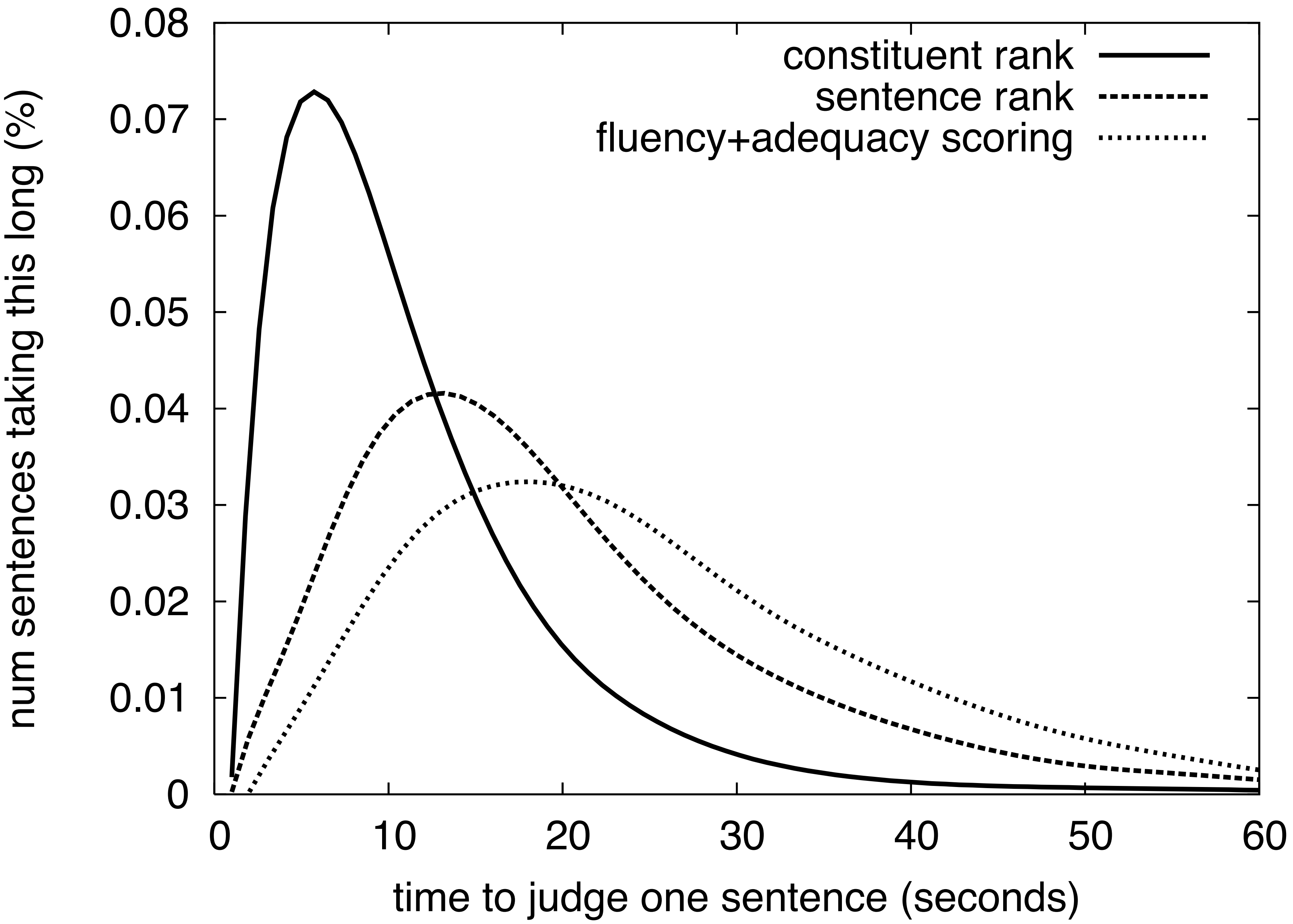
The man was **working for** the Ministry of Defense when he **lost** two extremely confidential documents in a train to Waterloo.

Which type of Human Evaluation is Best?



Evaluation type	$P(A)$	$P(E)$	K
Fluency (absolute)	.400	.2	.250
Adequacy (absolute)	.380	.2	.226
Fluency (relative)	.520	.333	.281
Adequacy (relative)	.538	.333	.307
Sentence ranking	.582	.333	.373
Constituent ranking	.693	.333	.540

Which type of Human Evaluation is Best?



Using manual judgments to evaluate automatic metrics...

- Measure correlation with human judgments
- System-level correlation
- Sentence-level correlation

Calculating Correlation

- The human evaluation metrics provide a ranking of the systems
  - So do the automatic metrics
- Calculate the correlation between the two lists
  - Metrics with higher correlation better predict human judgments

Spearman's rank correlation coefficient

- For system-level correlation

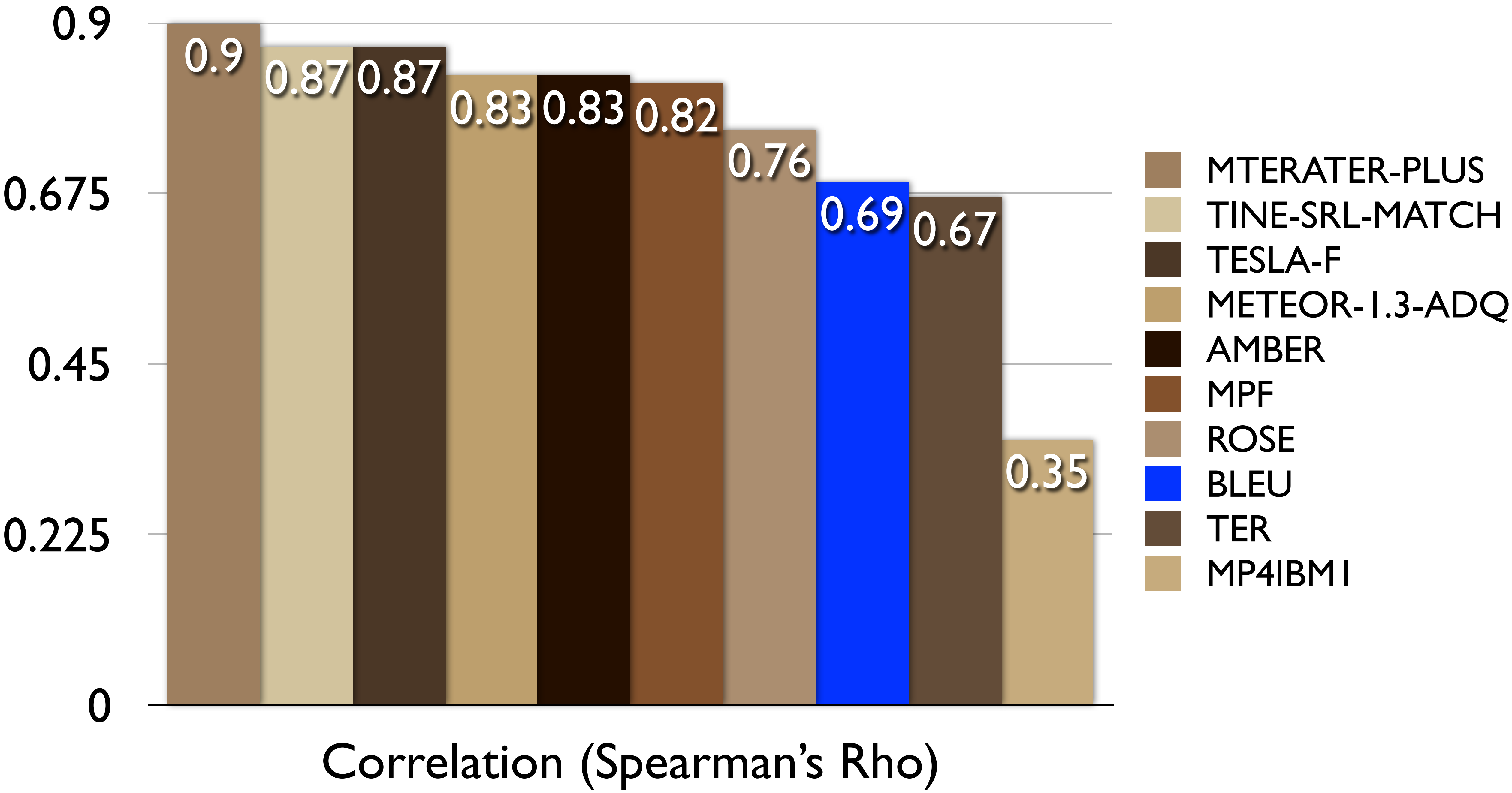
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Kendall’s Tau

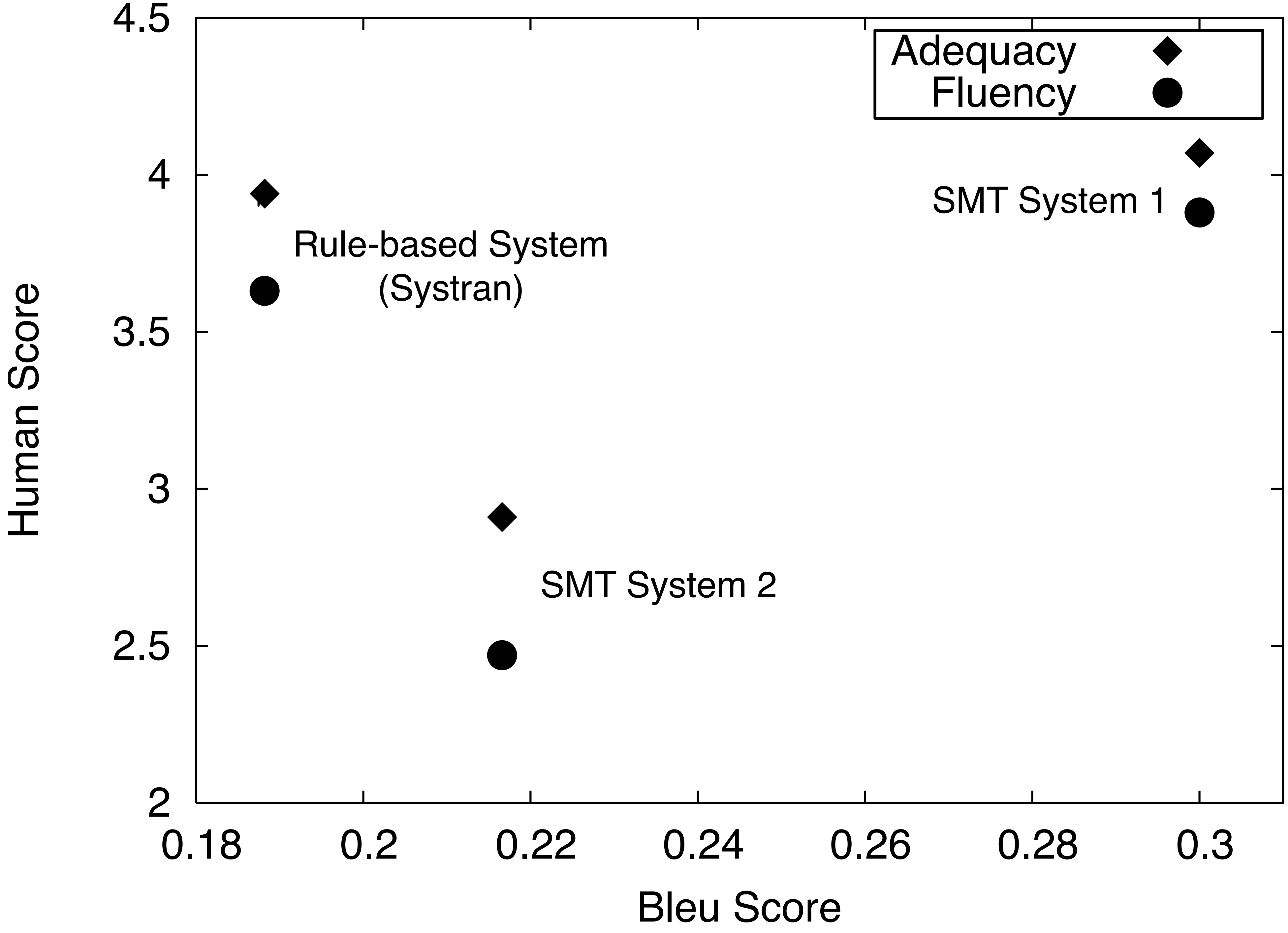
- Segment level evaluation

$$\tau = \frac{\text{num concordant pairs - num discordant pairs}}{\text{total pairs}}$$

Many metrics are better than BLEU

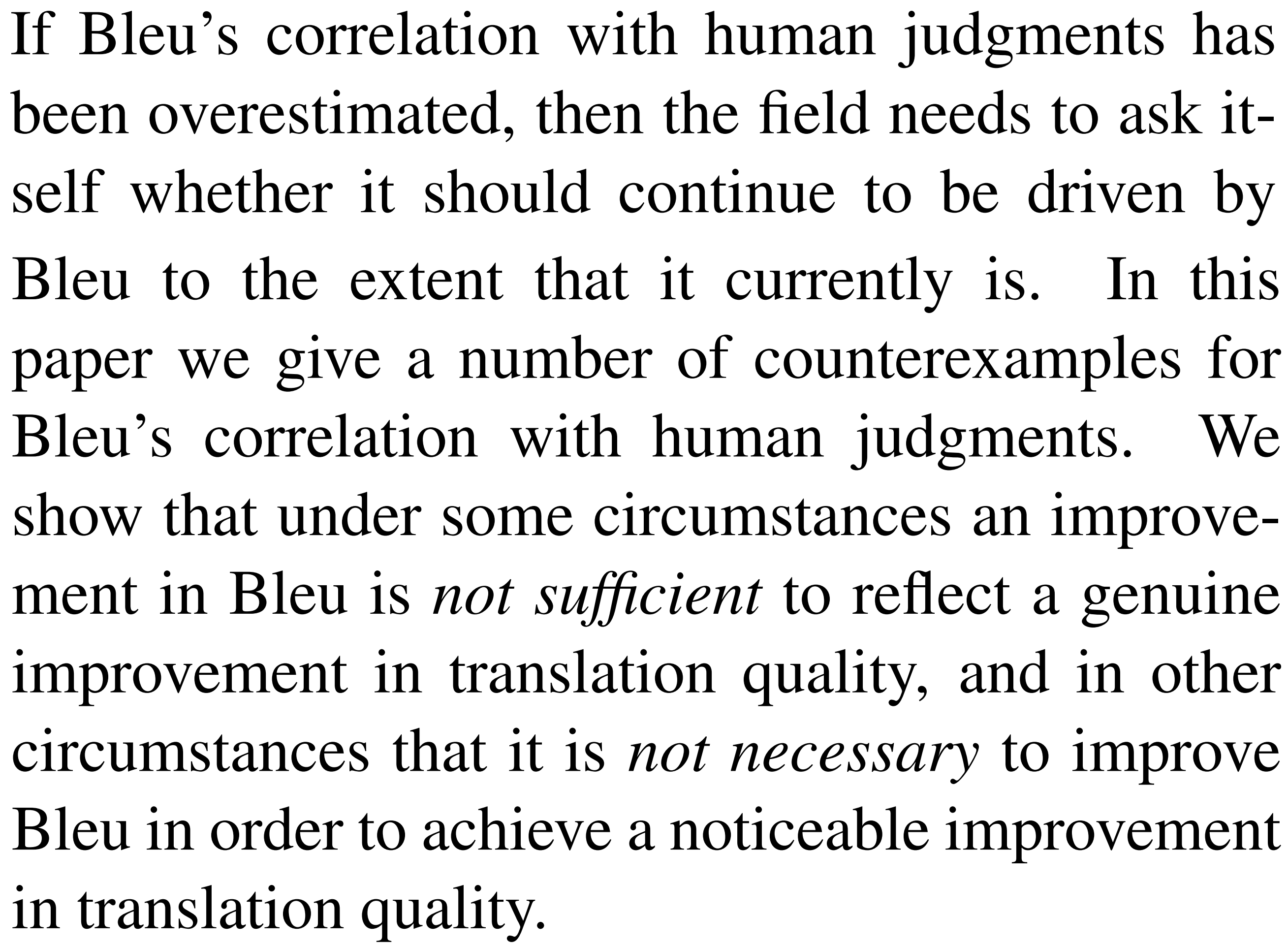


This is bad


4.5
4
3.5
3
2.5
2
0.18
0.2
0.22
0.24
0.26
0.28
0.3
Human Score
Bleu Score
Adequacy ◆
Fluency ●
Rule-based System
(Systran)
SMT System 1
SMT System 2

Re-evaluating the Role of BLEU in Machine Translation Research

Chris Callison-Burch Miles Osborne Philipp Koehn



If Bleu's correlation with human judgments has been overestimated, then the field needs to ask itself whether it should continue to be driven by Bleu to the extent that it currently is. In this paper we give a number of counterexamples for Bleu's correlation with human judgments. We show that under some circumstances an improvement in Bleu is *not sufficient* to reflect a genuine improvement in translation quality, and in other circumstances that it is *not necessary* to improve Bleu in order to achieve a noticeable improvement in translation quality.

Final thoughts on Evaluation

When writing a paper

- If you're writing a paper that claims that
  - one approach to machine translation is better than another, or that
  - some modification you've made to a system has improved translation quality
- Then you need to back up that claim
- Evaluation metrics can help, but good experimental design is also critical

Experimental Design

- Importance of separating out training / test / development sets
- Importance of standardized data sets
- Importance of standardized evaluation metric
- Error analysis
- Statistical significance tests for differences between systems

Evaluation drives MT research

- Metrics can drive the research for the topics that they evaluate
- NIST MT Eval -> DARPA Funding
- Bleu has lead to a focus on phrase-based translation
- Minimum error rate training (next lecture!)
- Other metrics may similarly change the community's focus

Invent your own evaluation metric

- If you think that Bleu is inadequate then invent your own automatic evaluation metric
- Can it be applied automatically?
- Does it correlate better with human judgment?
- Does it give a finer grained analysis of mistakes?

Goals for Automatic Evaluation

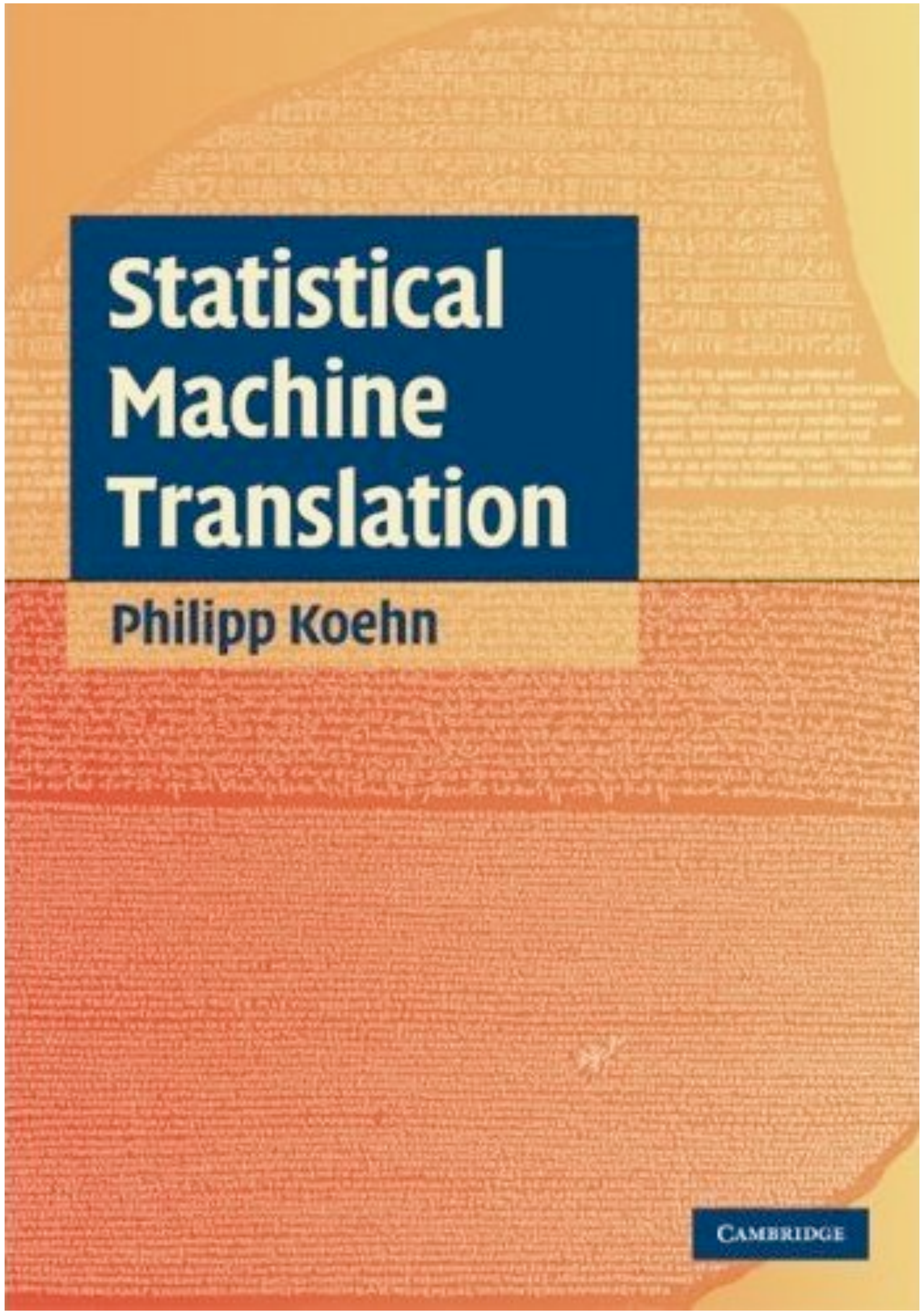
- No cost evaluation for incremental changes
- Ability to rank systems
- Ability to identify which sentences we're doing poorly on, and categorize errors
- Correlation with human judgments
- Interpretability of the score
- Quick to calculate for MERT

Questions?

- Tons of data available at
- http://statmt.org/wmt10/results.html
- http://statmt.org/wmt11/results.html
- http://statmt.org/wmt12/results.html
- http://statmt.org/wmt13/results.html

Reading

- Read 8 from the textbook



Announcements

- Upcoming language-in-10
  - Tuesday: **Edward** - Greek and **Rui** - Urdu
  - Thursday: **Anshul** - Japanese and **Rigel** - Javanese

Term project

- **Problem description** – similar to the descriptions on the HW assignments
- **Data collection** – used to train a model, and evaluate its performance
- **Objective function** – score submissions on a leaderboard
- **Default system** – An implementation of the simplest possible solution
- **Baseline system** – An implementation of a published baseline

Term project schedule

Spring Break	Choose your topic, start doing research on it
March 18	Draft write-up is due. (Problem description, citations to literature, descriptions of objective function and data)
March 27	Data collection is done. Submit it to TA.
April 1	Revised write up is due, taking into account any feedback from instructor and TA.
April 3	Objective function implementation is due
April 8	Default system is due
April 10	Baseline system is due
April 15	Term project is due (final writeup, data, objective fn, default system, and baseline system).

Term project schedule

April 17	HW5 assignments released (do one of the other student's term projects)
April 29	HW5 due: turn in your implementation of another student's term project
Until Final	Extra credit opportunities: do additional HWs from other students, add extensions to your own project that beat the baseline

http://mt-class.org/penn/project.html