Evaluation

- How to compare two arbitrary translations?
- Low agreement rate even between reviewers
- BLEU score – a popular automatic technique

Reference: ***E-mail was sent on Tuesday.***

System output: ***The letter was sent on Tuesday.***

1-grams: 4 / 6

2-grams: 3 / 5

3-grams: 2 / 4

4-grams: 1 / 3

Brevity: min(1, 6/5)

$$\mathrm{BLEU} = 1 \cdot \sqrt[4]{\frac{4}{6} \cdot \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3}}$$