

# Lecture 6.1: MT Evaluation

# What is MT Evaluation?

- How good is MT?
  - Define what is the “expected” quality
- Define metric(s)
  - => Human evaluation
  - => Automatic evaluation

# Human evaluation scoring: Adequacy / Fluency

- **Adequacy:** how much information is transferred between the original and the translation.

*Is the original meaning preserved in English?*

- **Fluency:** how good is the translation?

*(Ignoring original) How fluent is the English?*

# What you should not do!

- Round-trip evaluation RTT / reverse translation / back-and-forth translation (bad technique!)
  - Machine translate a source text
  - “back-translate” the MT
  - Look how similar it is

Very bad method, translation are not “transitive”, a 100% transfer MT would be best than any other.

“RTT is good ... for nothing” [Sommer 2006]

# Difficult problem

- Many translations are acceptable
  - Subjective human judgment
  - Challenge of automatic evaluation
- Quality criteria:
  - Adequacy
  - Fluency
  - Is the original information understandable?
  - Is it easy/quicker to rephrase the output to get it publishable?
  - Is this new MT system better than the previous?
  - *A* better than *B*?
  - Speed? Accessibility? Usability?
- Some non-scientific criteria:
  - I am a translator, is that product going to take out my job? 😊
  - I am a computer scientist, do I really care what MT quality means? 😊
  - I am spending \$ in translation, is quality that important? 😊

# Evaluation criteria, scale

- **Adequacy**

*How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?*

- 5 Everything / All meaning is preserved
- 4 Most meaning is preserved
- 3 Much meaning is still there
- 2 Little meaning
- 1 None of the original meaning is there

- **Fluency**

*To what extent the translation is grammatically well formed, contains correct spellings, is close to natural English (including terminology, names etc.)*

- 5 Flawless
- 4 Good
- 3 Non-native
- 2 Disfluent
- 1 Incomprehensible

# MT evaluation in real life

- What a MT developer cares about:  
Instant quality measure of a system
- What a user of your gist MT wants: original information is preserved (adequacy). But often can only judge the fluency
- What a translator needs: post-editing the MT. Both fluency & adequacy are important
- What your manager cares about:  
Costs

# Human evaluation scoring: Examples

- Fr: Le chat est dans la maison bleue  
Gloss: [the] [cat] [is] [in] [the] [house] [blue]
- Mt: The dog is in the brown house
  - adequacy: 2 – little meaning is preserved
  - fluency: 5 - flawless English
- Mt: In the house blue is the cat
  - adequacy: 5 – all meaning has been transferred
  - fluency: 2 - disfluent English



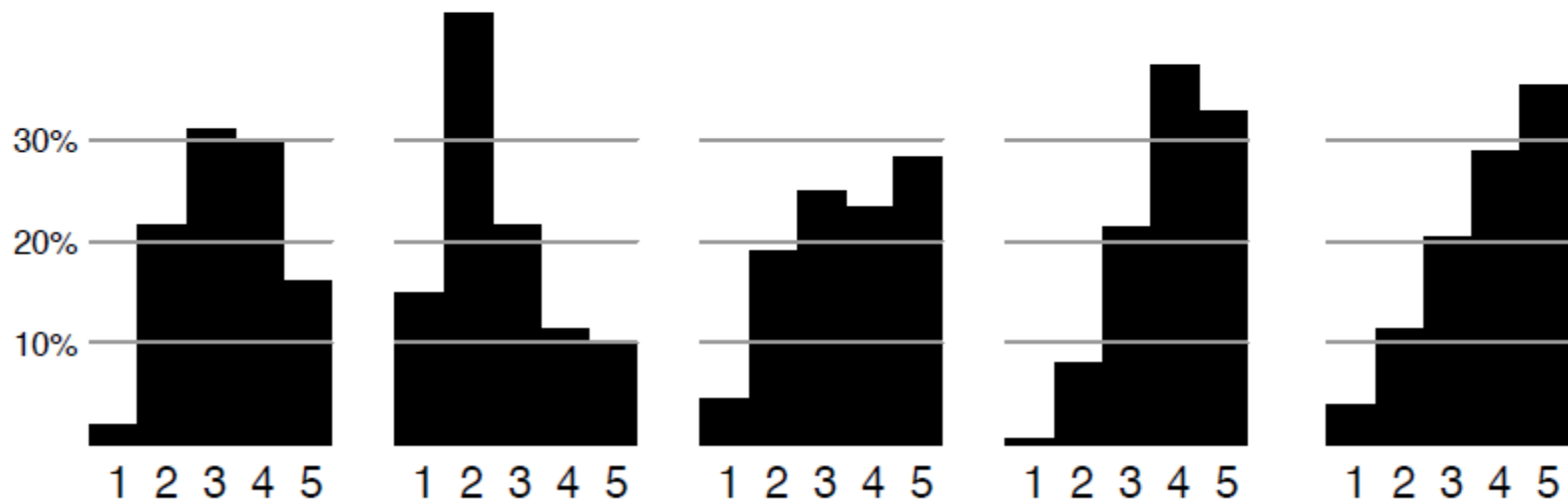
# Evaluators disagree

## Evaluators Disagree



10

Histogram of adequacy judgments by different human evaluators



(from WMT 2006 evaluation)

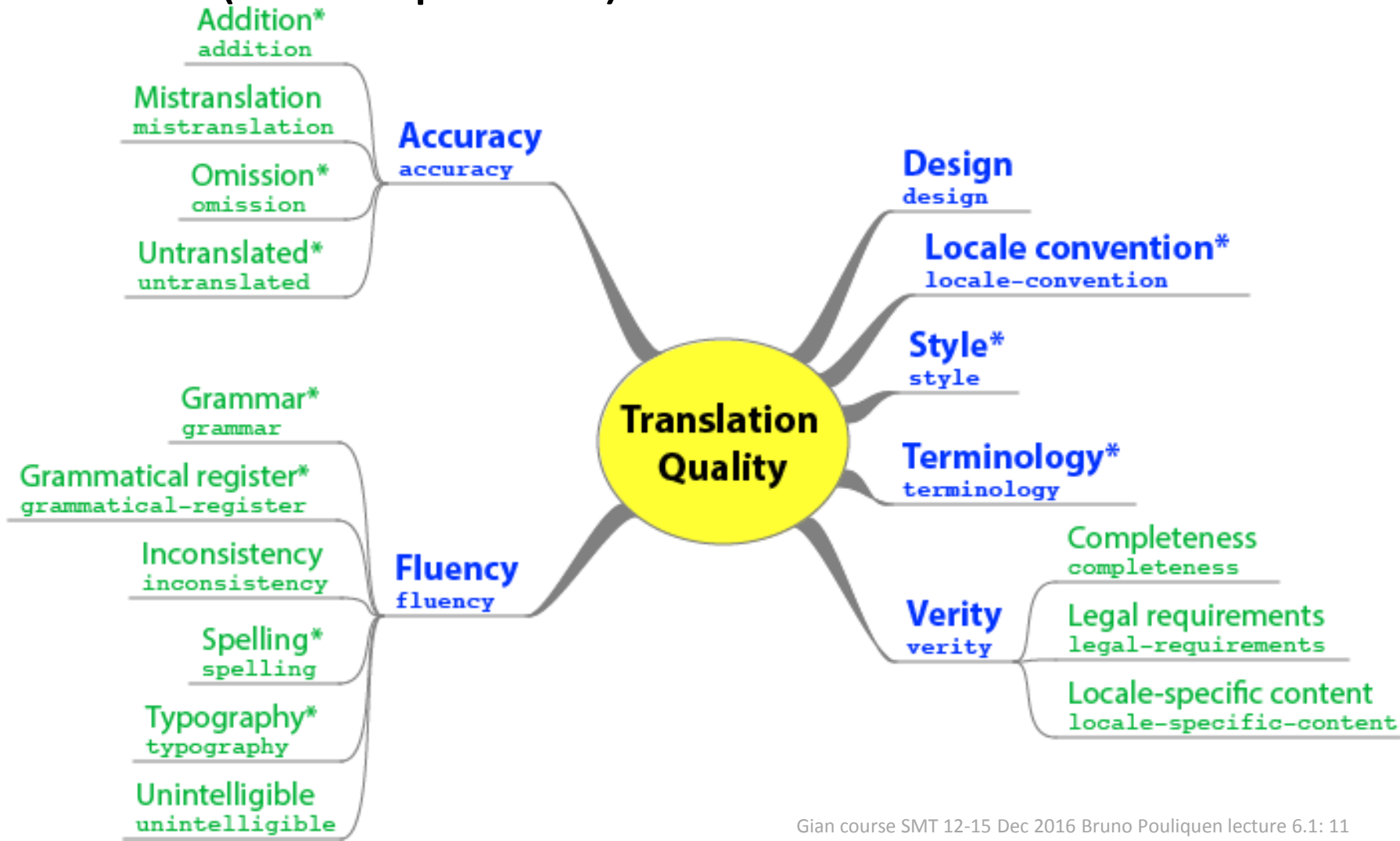
# Human evaluation, ranking:

- Is translation from system *A* better than system *B* (and/or systems *C,D...*)?
- “Double blind”
- Evaluators are more consistent

Evaluation type	$P(A)$	$P(E)$	$K$
Fluency	.400	.2	.250
Adequacy	.380	.2	.226
Sentence ranking	.582	.333	.373

# Human evaluation: error annotation

- Qt21 ([www.qt21.eu](http://www.qt21.eu))



# Human evaluation > task based

- Instead of asking evaluators “how good/bad is MT?”, let them use MT and look at the output
- Two families:
  - Collect post-editions, look at speed (average seconds per word) / number of edits (HTER)
  - Ask users to read a MT text
    - Ask questions about what they understood
    - Look at eyetracking

# Human evaluation > task based > HTER

- Human-targeted Translation Error Rate (Snover et al. 2006)
- $\text{HTER} = (\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}) / \text{Reference Words}$

# Human evaluation>task based>reading

- Comprehension

The screenshot shows a digital reading interface. On the left, there's a text area with a 'Highlight Color' toolbar (blue, yellow, pink, green) and a 'Clear Highlights' button. The text discusses the debate on retiring the penny. Below the text is a section titled 'ELIMINATE THE PENNY' with two paragraphs. On the right, there's a question area with a '1 / 10' indicator and a navigation arrow. It contains a question about the beliefs of those who support eliminating the penny, followed by four multiple-choice options (A, B, C, D) each with a dropdown arrow for selection.

Highlight Color: ■ ■ ■ ■ Clear Highlights

In recent years, there has been a growing movement to "retire" the penny or take it out of circulation. This movement has been countered by people passionate about preserving the penny. There are compelling reasons to eliminate the penny and to preserve it. What do you think?

**ELIMINATE THE PENNY**

According to the U.S. Mint, it costs 2.4 cents to produce one penny. In other words, the cost of making a penny is more than double its value. Since the United States Mint produced \$50 million worth of pennies in 2010 at a cost of \$120 million dollars, \$70 million was wasted.

Advocates of "retiring" the penny claim the coin is obsolete and virtually worthless. Nothing can realistically be bought for a penny anymore. In addition, simply handling pennies

1 / 10 =>

Those who support eliminating the penny believe....

A. ? pennies make the economy more efficient

B. ? nickels should be eliminated too

C. ? making pennies is a waste of money

D. ? pennies can still buy things today

- Eye tracking

From [Specia, 2016]

# Automatic evaluation

- Certainly not the most accurate, but the cheapest! (objective and flexible)
- Based on “reference translation(s)”, compute similarity between MT output and the original (human) translation