CENTER FOR ADVANCED STUDY OF LANGUAGE

JOHNS HOPKINS UNIVERSITY

human language technology center of excellence

Putting the "human" back in HLT:

The importance of human evaluation in assessing the quality and potential uses of translation technology

Erica Michael & Petra Bradley

University of Maryland CASL

Paul McNamee & Matt Post

Johns Hopkins University HLT COE

Outline

- **Part 1:** Human comprehension of machine translation (MT) output
- **Part 2:** Use of MT for translation and comprehension of Chinese texts
- **Part 3:** Utility of translation memory (TM) in an operational context

CENTER FOR ADVANCED STUDY OF LANGUAGE

JOHNS HOPKINS UNIVERSITY

human language technology center of excellence

Human comprehension of MT output: Findings from 2016 SCALE

SCALE: Summer Camp for Applied Language Exploration

SCALE 2016a: Knowledge Rich Statistical Machine Translation

Approaches to MT

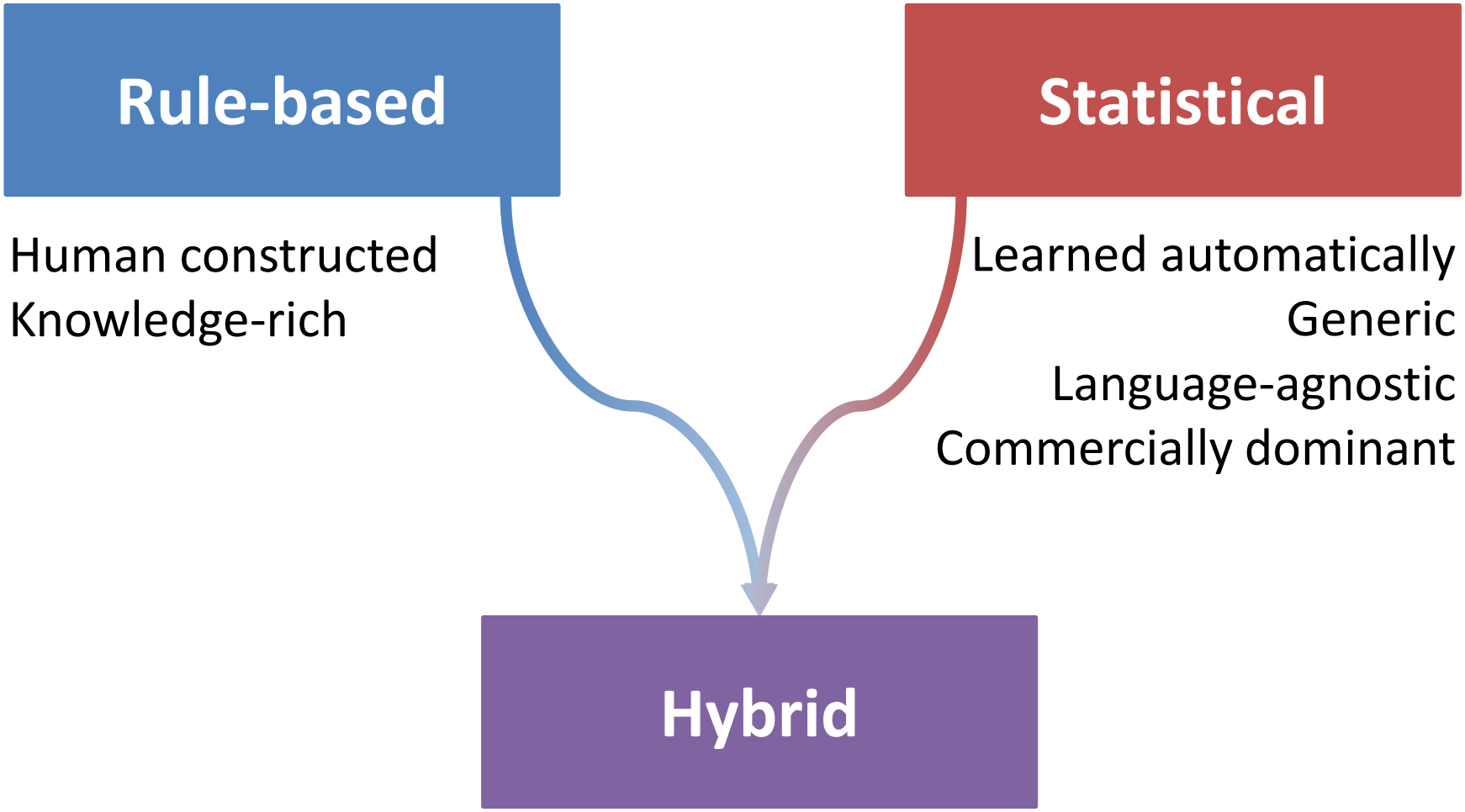
Rule-based	Human experts create rules
Example-based	Sentences are matched against previous translations
Statistical	Phrase translations are learned from many examples
Deep Learning	Neural nets are trained from many examples

Features of each approach

Rule-based	Statistical
• Rules are composed by language experts • Performs a deep source language analysis • Easy to update, adapt to new domains • Very fast	• Learns automatically from example translations • Doesn't require language-specific knowledge • Leverages Big Data

Workshop questions

- How do the Rule-based and Statistical paradigms compare in terms of translation quality?
- Can we improve translation quality by combining them?


Rule-based
Statistical
Human constructed
Knowledge-rich
Learned automatically
Generic
Language-agnostic
Commercially dominant
Hybrid
Best of both worlds ?

Workshop themes

- Explore Hybrid MT *in a number of languages*
- *Augmenting* SMT using linguistic information from Rule-based MT
- Evaluate *both* intrinsic MT quality and document comprehensibility
- Make *engineering changes* to support fast updates, easy adoption

Statistical translation

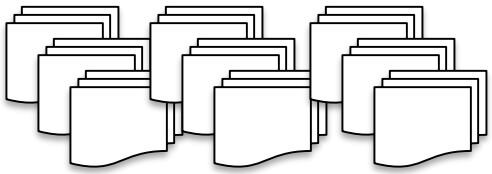
Translation Model (learned from parallel texts)

'Это было ошеломляющее зрелище' – сказал Керри после посещения парка.

'It was a stunning sight' - Kerry said after visiting the park.

p(канада | canada) = 0.7
p(канада | canadian) = 0.1
p(канада | montreal) = 0.1
p(украине | ukraine) = 0.6

Language Model (learned from text)



p(“the eyes of texas”) = high
p(“eyes texas the of”) = very low
p(“like to eat crabs”) = high
p(“like to eat forks”) = low

Used Open Source Apache Joshua during workshop:
joshua.incubator.apache.org

Hybrid approaches - I

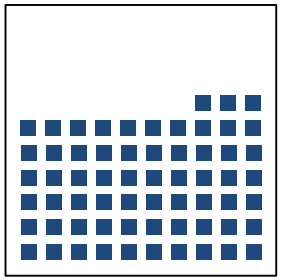
- Dictionary Extraction
  - Add translation pairs from RBT lexicon to augment the SMT translation model
  - Pros: direct; can be done once; reduces OOVs
  - Cons: requires lots of morphological analysis to accurately convert lexicon base forms to the surface forms needed in SMT

Hybrid approaches - II

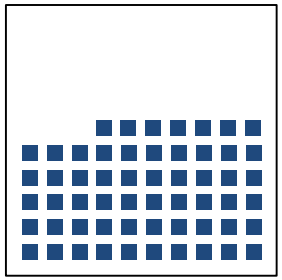
- "Black Box"
  - Apply RBT to a set of foreign sentences. Use generated translations as supplementary training data ("imperfect bitext")
  - Pros: simple; can take advantage of additional RBT processing (e.g., pre- or post-corrections, transliteration of unknown words)
  - Cons: possibly adding errorful training data

Training data

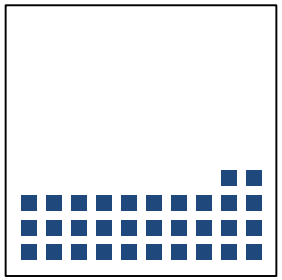
Higher Resource



Russian (63)

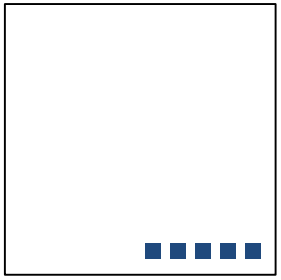


Arabic (57)

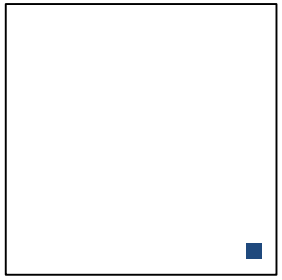


Portuguese (32)

Lower Resource



Farsi (5)



Swahili (0.2)

■ ***= 1 million sentences with English translations***

Two evaluations

- Automatic Metrics
  - Given reference translations, compute scores that characterize the fidelity and fluency of MT system output
- Human Evaluation
  - More directly measures quality and the ability to use MT to support analyst workflows

Motivation for human evaluation

- **Research questions**
  - How well can human users comprehend machine-translated text?
  - Are differences in BLEU scores reflected in human comprehension scores?
- **Goals for implementing the study**
  - Materials: as authentic as possible
  - Task: similar to some of the ways in which analysts might use MT output

Design overview

- **5 languages**
  - Arabic, Farsi, Portuguese, Russian, Swahili
  - Range of morphological complexity and resources
- **4 translation types**
  - **Human**, **Rule-based**, **Statistical**, **Hybrid**
- **16 passages per language**
  - For most languages, 8 News + 8 Conversation
- **2-6 questions per passage**
  - Topic, Main Idea, Fact, Inference
  - Questions based on human translation

Materials

- **Arabic & Farsi**
  - Created by Doug Jones (MITLL)
  - All news items
  - Similar to Voice of America and BBC items
  - Originally in English, translated to Arabic and Farsi
  - Average number of words per passage = 340.8
  - 2 or 3 questions per passage

Materials

	Portuguese	Russian	Swahili
News			
Sources	Global Voices, BBC	Voice of America	BBC
Original language	English	English	Mostly English
Avg. # of words	258.4	308.8	138.4
Conversation			
Sources	Global Voices	Global Voices, course materials	JSPS Global COE, swahiliweb.net
Original language	Portuguese, English, other	Russian	Swahili
Avg. # of words	221.4	224.1	232.7

Types of questions

Similar to triage tasks

Topic	Main Idea
The topic of this passage is:	Which of the following is the best title for this article?
a. Polygamy	a. Obama promotes voting on American Idol finale
b. Marriage	b. Obama performs on American Idol finale
c. Divorce	c. Obama asks for votes on American Idol finale
d. Children	d. American Idol comes to an end

Types of questions

More relevant for finding essential elements of information

Fact	Inference
According to the IMF, what would increase indebted-ness for emerging market economics?	What does Person A imply about Kennedy and Krushchev?
a. High budget deficits	a. They were prudent and averted a war.
b. Increasing 50% of their gross domestic product	b. They met because of the Caribbean Crisis.
c. Strong revenues with slow growth	c. They caused the Caribbean Crisis.
d. World government debt	d. They could have worked together to avoid catastrophe.

Pilot testing

- **Guessability**
  - Tested questions without passages to ensure correct responses could not be guessed
    - Criterion: ≤ .70
- **Difficulty**
  - Tested questions with human-translated passages to ensure questions were not too difficult
    - Criterion: ≥ .50

MT examples

Portuguese

À medida que as pessoas envelhecem, os relógios biológicos começam a voltar a despertar mais cedo,

Human Translation

As people age, the biological clock starts waking up earlier again,

Statistical MT

As people age, the biological clocks are beginning to return to awaken sooner,

Rule-based MT

while the people age, the biological clocks begin to return to awaken earlier,

Hybrid MT

While people age, the biological clocks are beginning to return to wake up earlier

MT examples

Swahili

A: Labla Marekani katika jimbo gani? Marekani ni kubwa.

Human Translation

A: To be more precise, which state in America? America is vast.

Statistical MT

A: “Maybe America in what state? The United States is the greatest.

Rule-based MT

A: LABLA America in/at what region? America is big.

Hybrid MT

A: Maybe America in what region? The United States is big.

MT examples (with sample question)

According to the story, in what way are planes and trains alike?

a. They both rarely have accidents.
b. They both offer tea.
c. They both have good window seats.
d. They both have duty free.

Russian

B: Но на поезде почти никогда не бывает аварий...
A: Самолёты тоже падают очень редко.

MT examples (with sample question)

According to the story, in what way are planes and trains alike?

a. They both rarely have accidents.
b. They both offer tea.
c. They both have good window seats.
d. They both have duty free.

Rule-based MT

B: but on the train almost never is wrecks...
A: aircraft also fall very far-between

Russian

B: Но на поезде почти никогда не бывает аварий...
A: Самолёты тоже падают очень редко.

MT examples (with sample question)

According to the story, in what way are planes and trains alike?

a. They both rarely have accidents.
b. They both offer tea.
c. They both have good window seats.
d. They both have duty free.

Russian

B: Но на поезде почти никогда не бывает аварий...
A: Самолёты тоже падают очень редко.

Rule-based MT

B: but on the train almost never is wrecks...
A: aircraft also fall very far-between

Stat MT / Hybrid MT (identical)

B: but the train is almost never crashes...
A: the planes also fall very rarely.

MT examples (with sample question)

According to the story, in what way are planes and trains alike?

a. They both rarely have accidents.
b. They both offer tea.
c. They both have good window seats.
d. They both have duty free.

Russian

B: Но на поезде почти никогда не бывает аварий...
A: Самолёты тоже падают очень редко.

Rule-based MT

B: but on the train almost never is wrecks...
A: aircraft also fall very far-between

Stat MT / Hybrid MT (identical)

B: but the train is almost never crashes...
A: the planes also fall very rarely.

Human Translation

B: But by train there's almost never an accident...
A: Planes fall very rarely too.

MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

a. Her belief that obese children can become seriously ill.
b. A program to raise funds for terminally ill Americans.
c. The need to help millions of children without health care.
d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءًا مستمراً على مرض سمنة "هيا نتحرك"وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرّضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

a. Her belief that obese children can become seriously ill.
b. A program to raise funds for terminally ill Americans.
c. The need to help millions of children without health care.
d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءًا مستمراً على مرض سمنة "هيا نتحرك"وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرّضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it Mrs/lady Obama ``come on, we move`` and that launched her last February to focusing light continuing on disease fat(ness)/Samnah the children in the United States and also on the millions of guys/al-Shabab the exhibitions to the injury danger with illnesses dangerous related with the fatness.

MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

a. Her belief that obese children can become seriously ill.
b. A program to raise funds for terminally ill Americans.
c. The need to help millions of children without health care.
d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءًا مستمراً على مرض سمنة "هيا نتحرك"وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرّضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it Mrs/lady Obama `come on, we move`` and that launched her last February to focusing light continuing on disease fat(ness)/Samnah the children in the United States and also on the millions of guys/al-Shabab the exhibitions to the injury danger with illnesses dangerous related with the fatness.

MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

a. Her belief that obese children can become seriously ill.
b. A program to raise funds for terminally ill Americans.
c. The need to help millions of children without health care.
d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءًا مستمراً على مرض سمنة "هيا نتحرك"وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرّضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it Mrs/lady Obama `come on, we move`` and that launched her last February to focusing light continuing on disease fat(ness)/Samnah the children in the United States and also on the millions of guys/al-Shabab the exhibitions to the injury danger with illnesses dangerous related with the fatness.

Statistical MT

2 The Obama campaign Ms. " come on move " last February, which unleashed by continuous casting ضوءًا سمنة satisfactory to children in the United States and also on the millions of youth المعرّضينat risk of contracting serious diseases related to fat

Proceedings of AMTA 2016, vol. 2: MT Users' Track
Austin, Oct 28 - Nov 1, 2016 | p. 482

MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

a. Her belief that obese children can become seriously ill.
b. A program to raise funds for terminally ill Americans.
c. The need to help millions of children without health care.
d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءًا مستمراً على مرض سمنة "هيا نتحرك"وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرّضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Rule-based MT

2 And aim(s) campaign/download_it Mrs/lady Obama `come on, we move`` and that launched her last February to focusing light continuing on disease fat(ness)/Samnah the children in the United States and also on the millions of guys/al-Shabab the exhibitions to the injury danger with illnesses dangerous related with the fatness.

Statistical MT

2 The Obama campaign Ms. " come on move " last February, which unleashed by continuous casting ضوءًا سمنة satisfactory to children in the United States and also on the millions of youth المعرّضينat risk of contracting serious diseases related to fat

MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

a. Her belief that obese children can become seriously ill.
b. A program to raise funds for terminally ill Americans.
c. The need to help millions of children without health care.
d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءًا مستمراً على مرض سمنة "هيا نتحرك"وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرّضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Hybrid MT

2 The Obama campaign Ms. " come on move " by last February which continued to shed light on disease fat children in the United States and also on the millions of youth the exhibitions at risk of contracting serious diseases related to fat.

MT examples (with sample question)

What portion of Mrs. Obama's speech is highlighted in Lines 1-2?

a. Her belief that obese children can become seriously ill.
b. A program to raise funds for terminally ill Americans.
c. The need to help millions of children without health care.
d. The NAACP's campaign to improve US medical care.

Arabic

والتي أطلقتها في فبراير الماضي إلى تسليط ضوءًا مستمراً على مرض سمنة "هيا نتحرك"وتهدف حملة السيدة أوباما الأطفال في الولايات المتحدة و أيضاً على ملايين الشباب المعرّضين لخطر الإصابة بأمراض خطيرة ذات صلة بالسمنة.

Hybrid MT

2 The Obama campaign Ms. " come on move " by last February which continued to shed light on disease fat children in the United States and also on the millions of youth the exhibitions at risk of contracting serious diseases related to fat.

Human Translation

2 Mrs. Obama's "Let's Move" campaign, which she launched this past February, aims to shine a constant spotlight on childhood obesity in the United States, and the millions of young people at risk of developing related serious health conditions.

Participant overview

- Recruited via Amazon's Mechanical Turk
  - Required to be in the United States
- 77-78 participants per language
  - Each language tested separately
    - No information about what the source language was
    - Participants could do experiment in multiple languages
- All participants saw all translation types
  - Told that some were translated by humans and some by machines, but not which was which

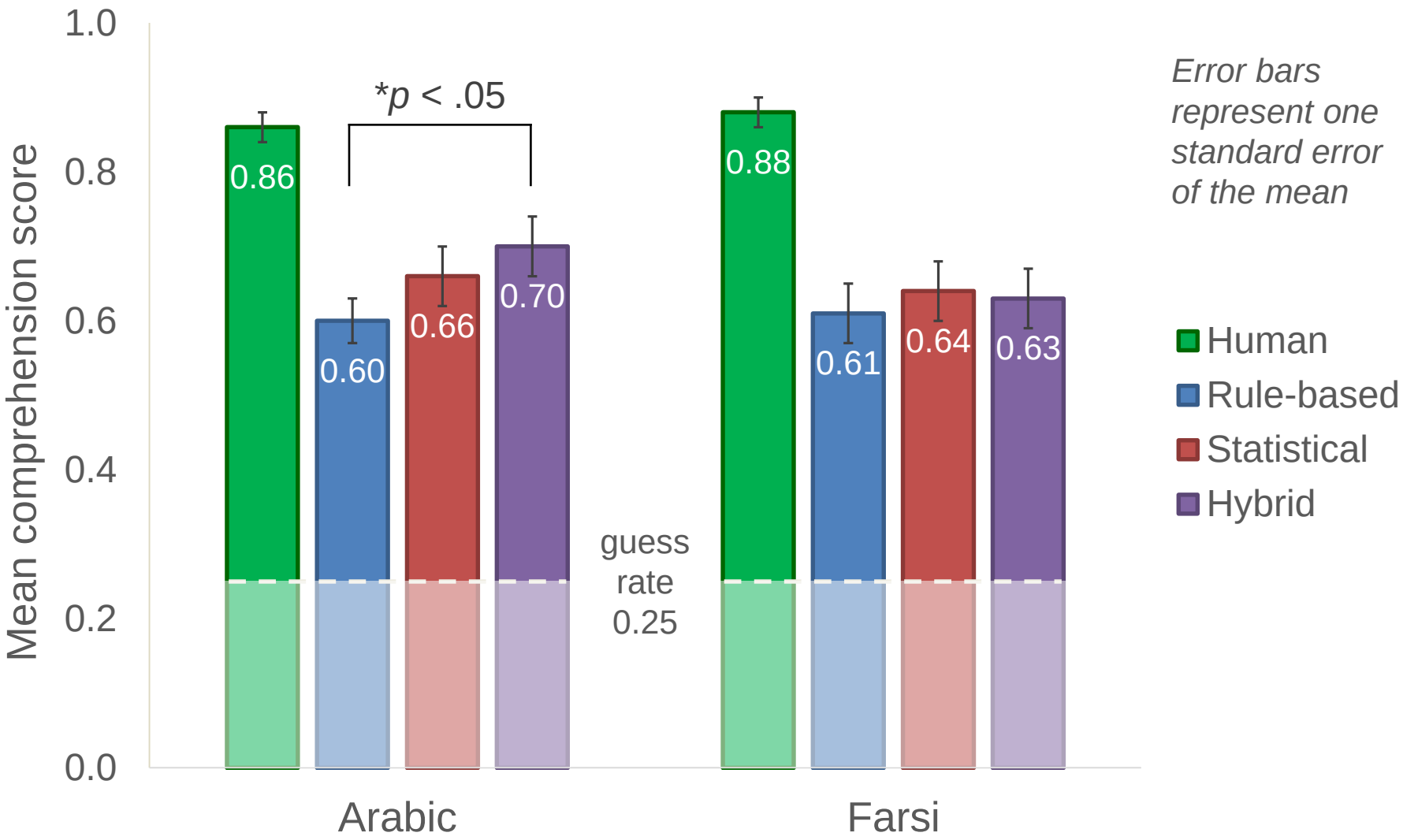
Participant demographics

	Arabic	Farsi	Portuguese	Russian	Swahili
n =	77	78	78	78	78
M/F	40/37	38/40	34/44	43/34	46/32
Average age	36.8	38.9	38.9	40.6	36.8
Age range	20-69	22-67	20-68	20-69	19-69
At least some college	81%	91%	81%	79%	81%
Native language	English	English	English	English	English
Some knowledge of source language	n = 1 *proficiency rating = 3 (1-10 scale)*		n = 1 *proficiency rating = 3 (1-10 scale)*		

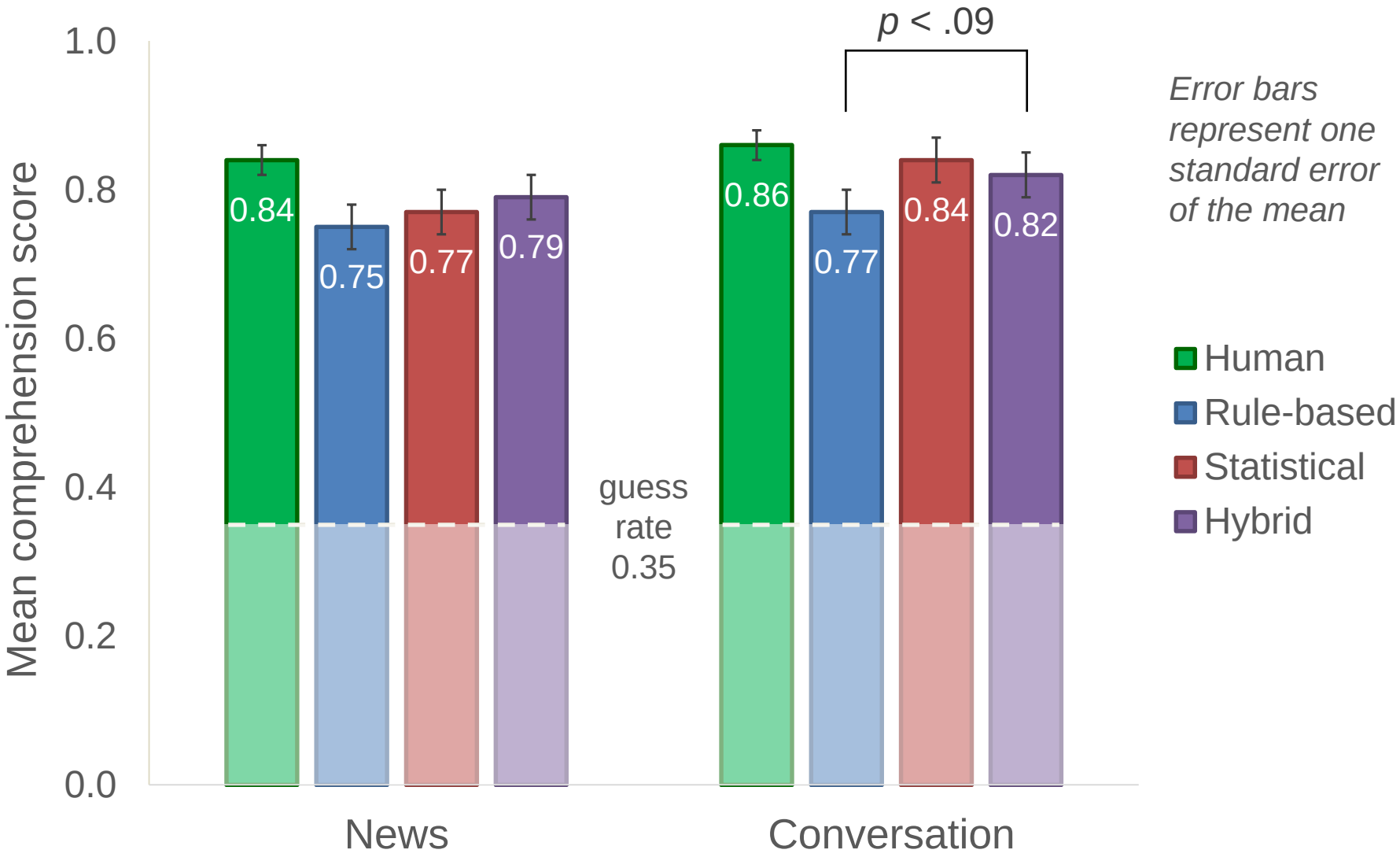
Participant comments

- I found the computer generated text very difficult to **get specific information** from
- This wasn't as easy as I thought it would be. Most of it was **gibberish** but I tried my best to understand the meaning.
- The reading was **really hard to follow** and often I had to read it 4-5 times just to make a guess.
- This **melted my brain**. / My **brain is fried**. / My **brain hurts!**
- Some questions asked for **reference to a specific person or phrase** and a couple of those **weren't actually found in the texts**.
- This task was really interesting, like I was **trying to decode something**. I often had to do this with my family, as my entire family came from Puerto Rico and spoke with very **broken English**.

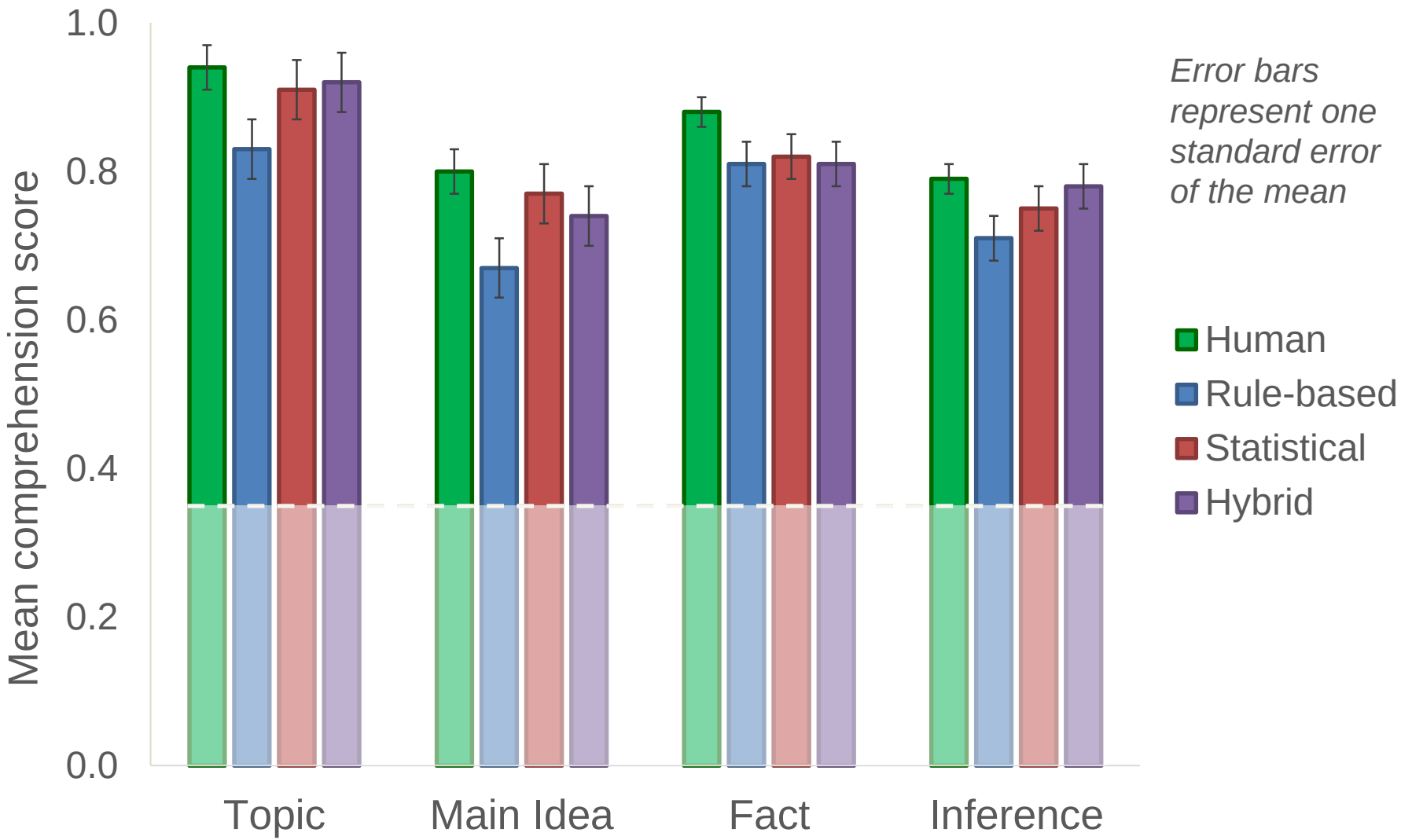
Arabic & Farsi comprehension



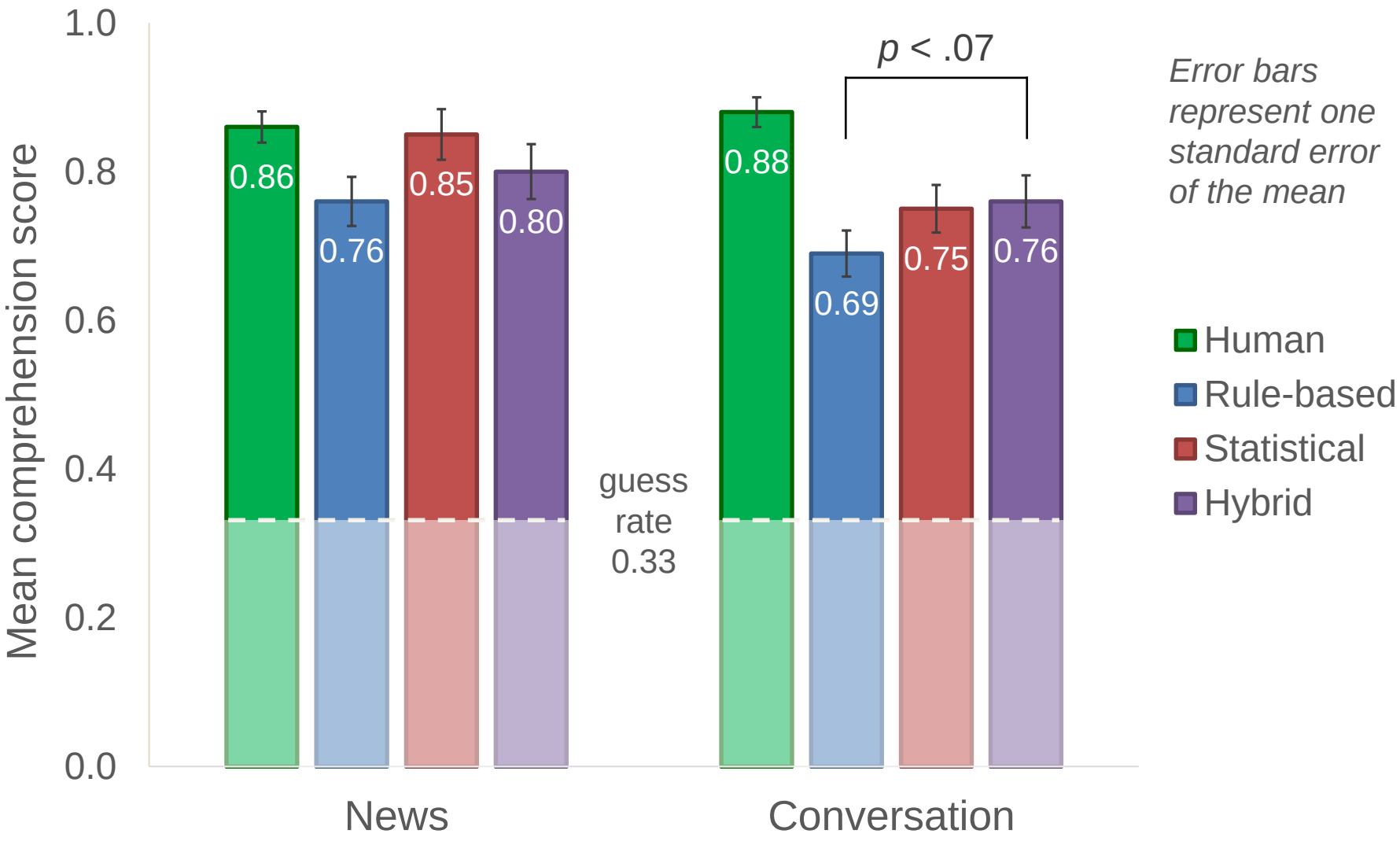
Portuguese comprehension



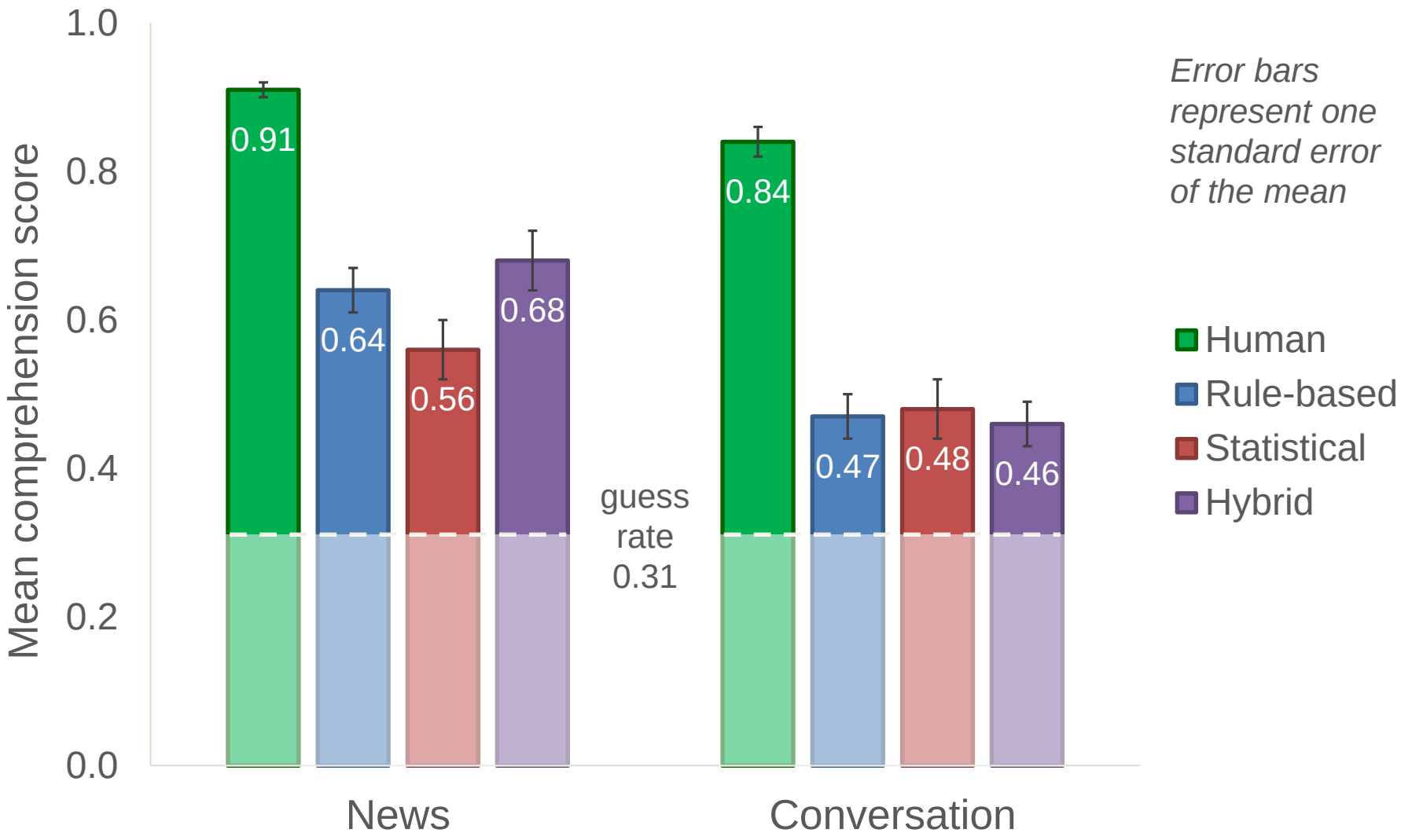
Portuguese comprehension



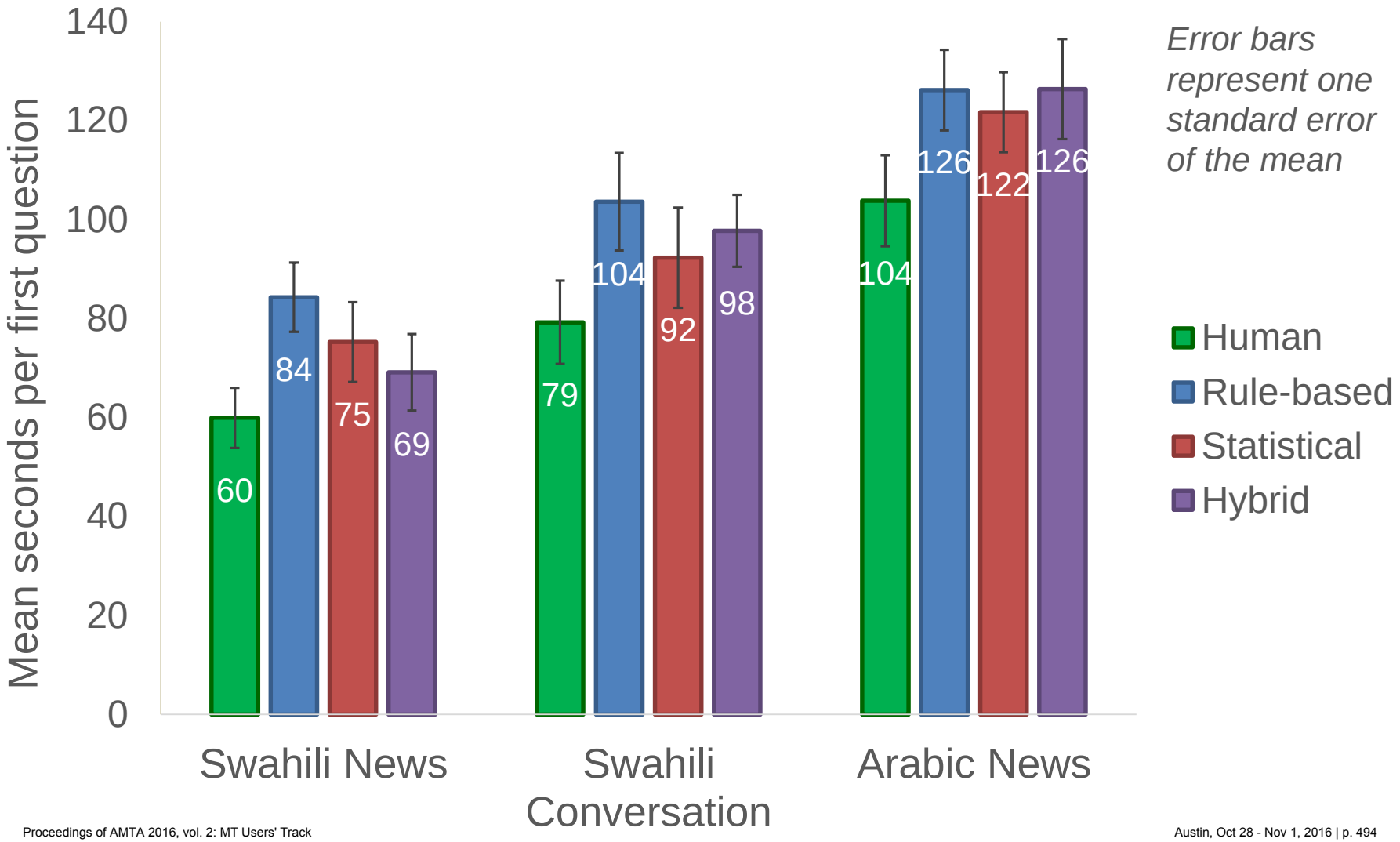
Russian comprehension



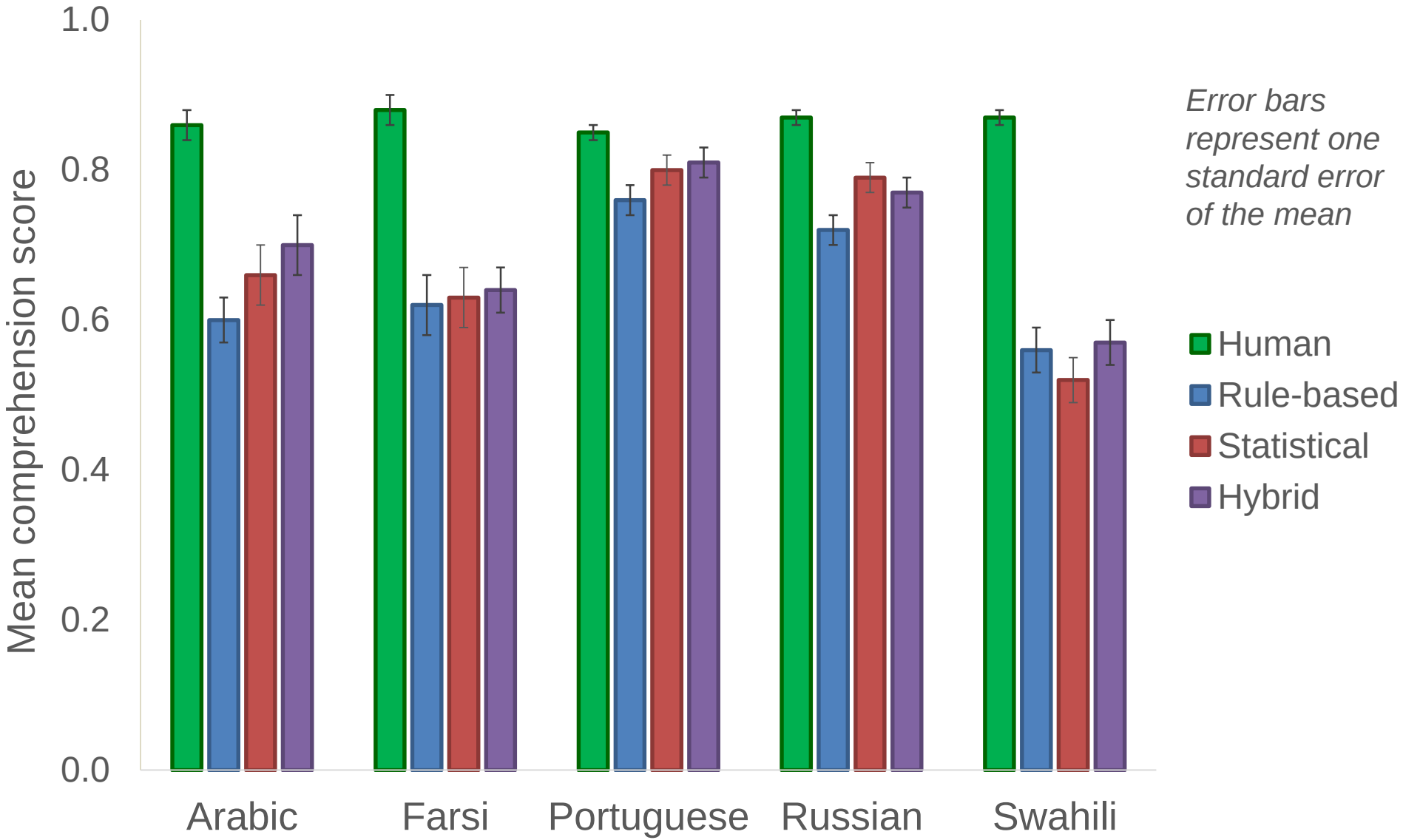
Swahili comprehension



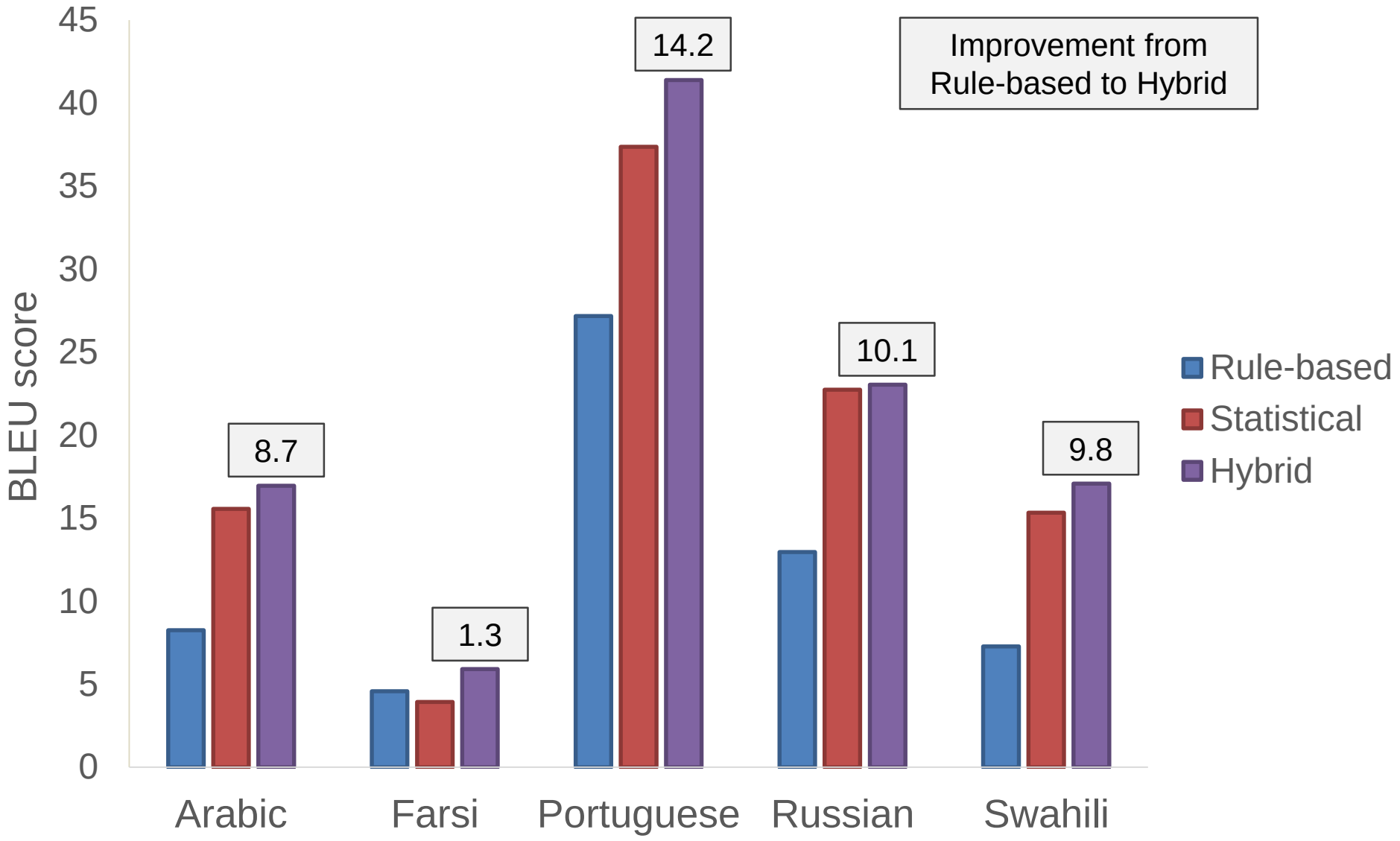
Time per first question (Swahili, Arabic)



Comprehension summary



BLEU scores


BLEU score
45
40
35
30
25
20
15
10
5
0
Arabic
Farsi
Portuguese
Russian
Swahili
8.7
1.3
14.2
10.1
9.8
Improvement from Rule-based to Hybrid
Rule-based
Statistical
Hybrid

Best comprehension scores

	News	Conversation
Arabic	**Hybrid**	*Not tested*
Farsi	**No sig differences**	*Not tested*
Portuguese	**No sig differences**	**Stat/Hybrid** (marginal)
Russian	**StatMT**	**Hybrid** (marginal)
Swahili	**Hybrid***	**No sig differences**

**Significantly higher than StatMT, but not significantly different from Rule-based*

Automatic scoring metrics for MT

- TER – Translation Error Rate
- BLEU – BiLingual Evaluation Understudy
- METEOR – Metric for Evaluation of Translation with Explicit ORdering

TERPa

- Translation Error Rate Plus
  - Like other TER scores, measures the edits required to change MT output to match a reference translation
  - Edits: shifts, substitutions, insertions, deletions
  - Improves on TER to be more correlated with human judgments of translation quality compared to BLEU (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006)

TERPa scoring

- Scoring

  0 = perfect match (can have synonyms)

  1.0 = default cap

  2+ = theoretical maximum

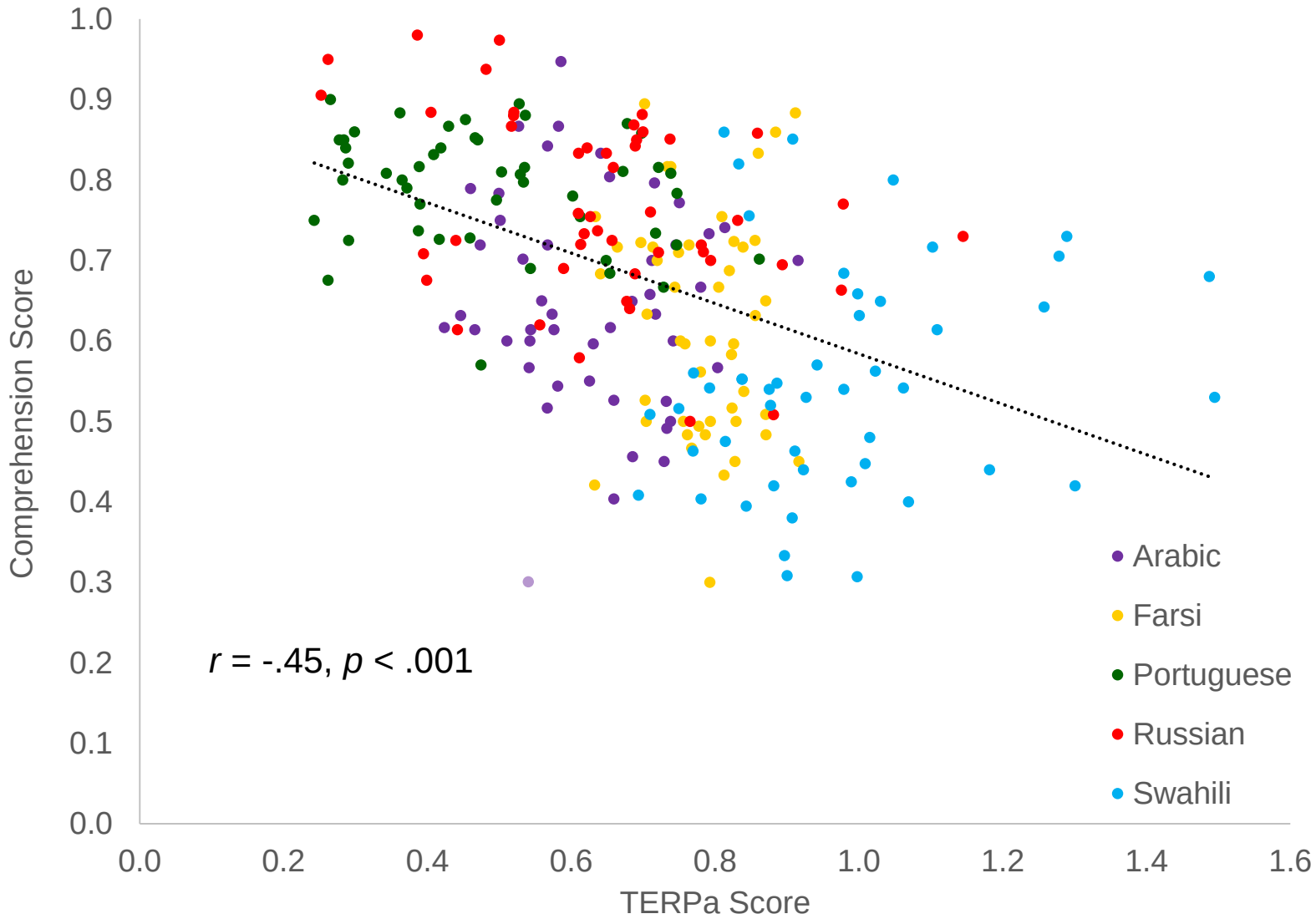
- Can use 1/TERPa for easier interpretation

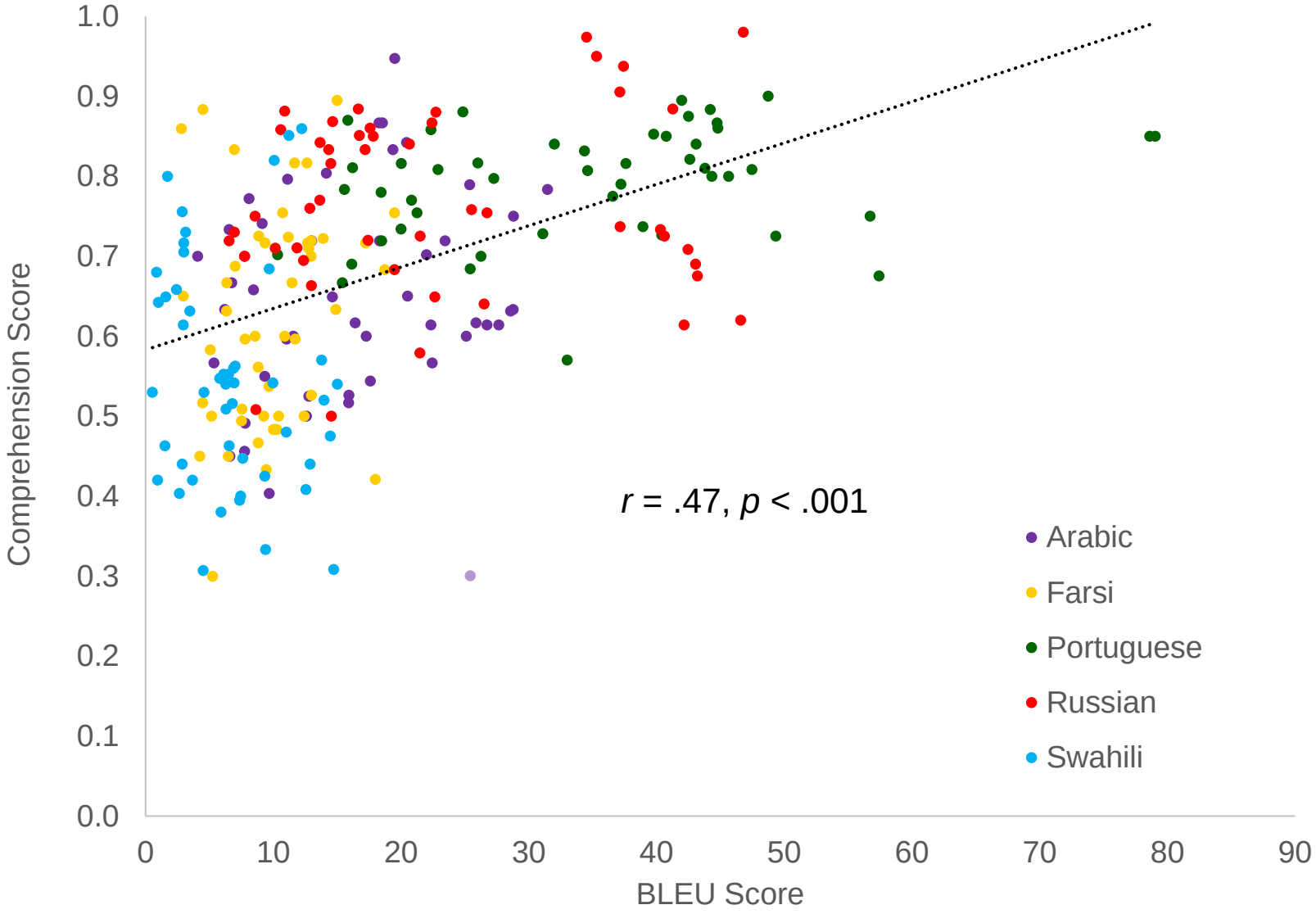
BLEU

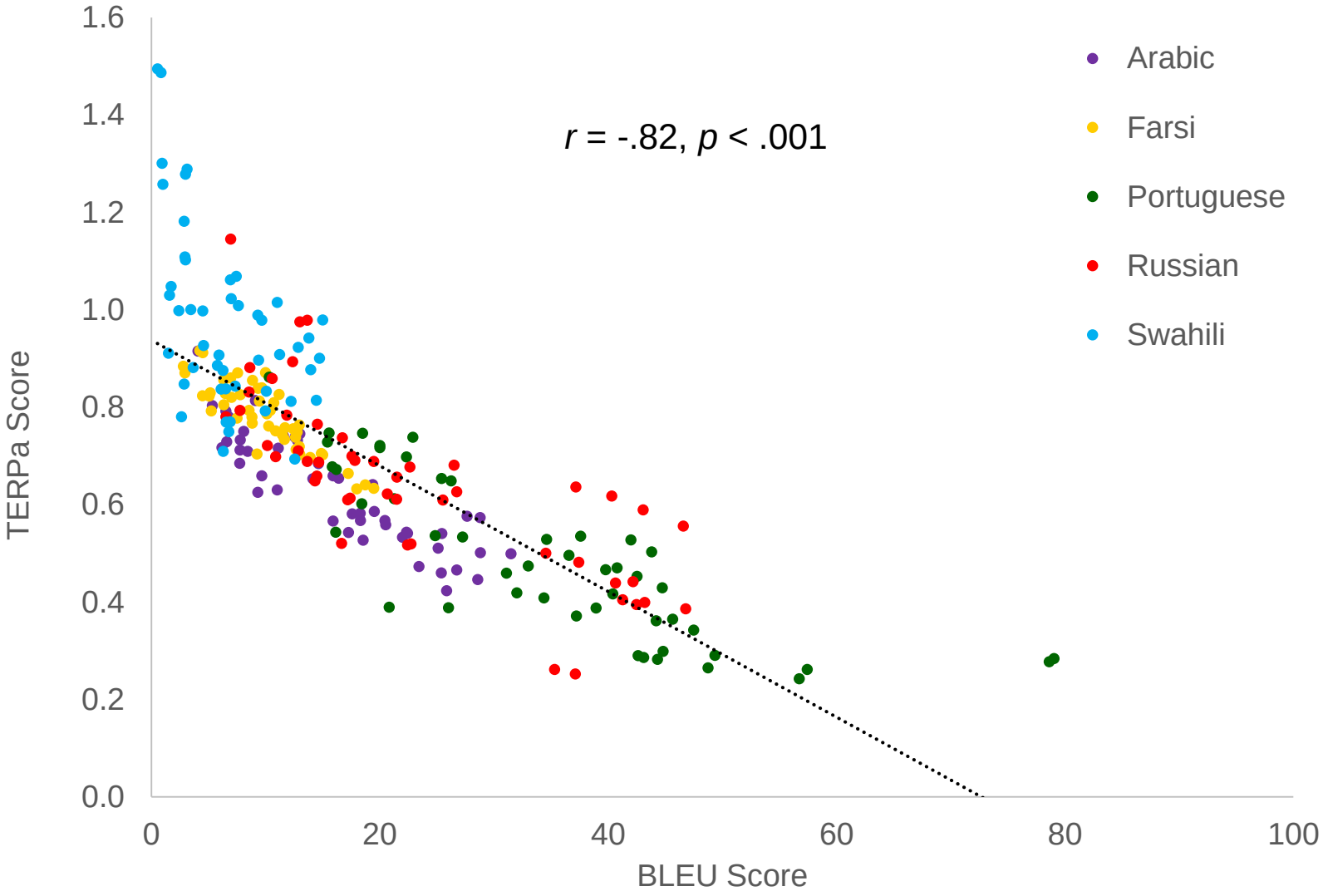
- Measures the number of n-grams from the MT that occur within the reference(s)
- Should be used with a large number of references and large number of sentences in order to correlate with human judgments

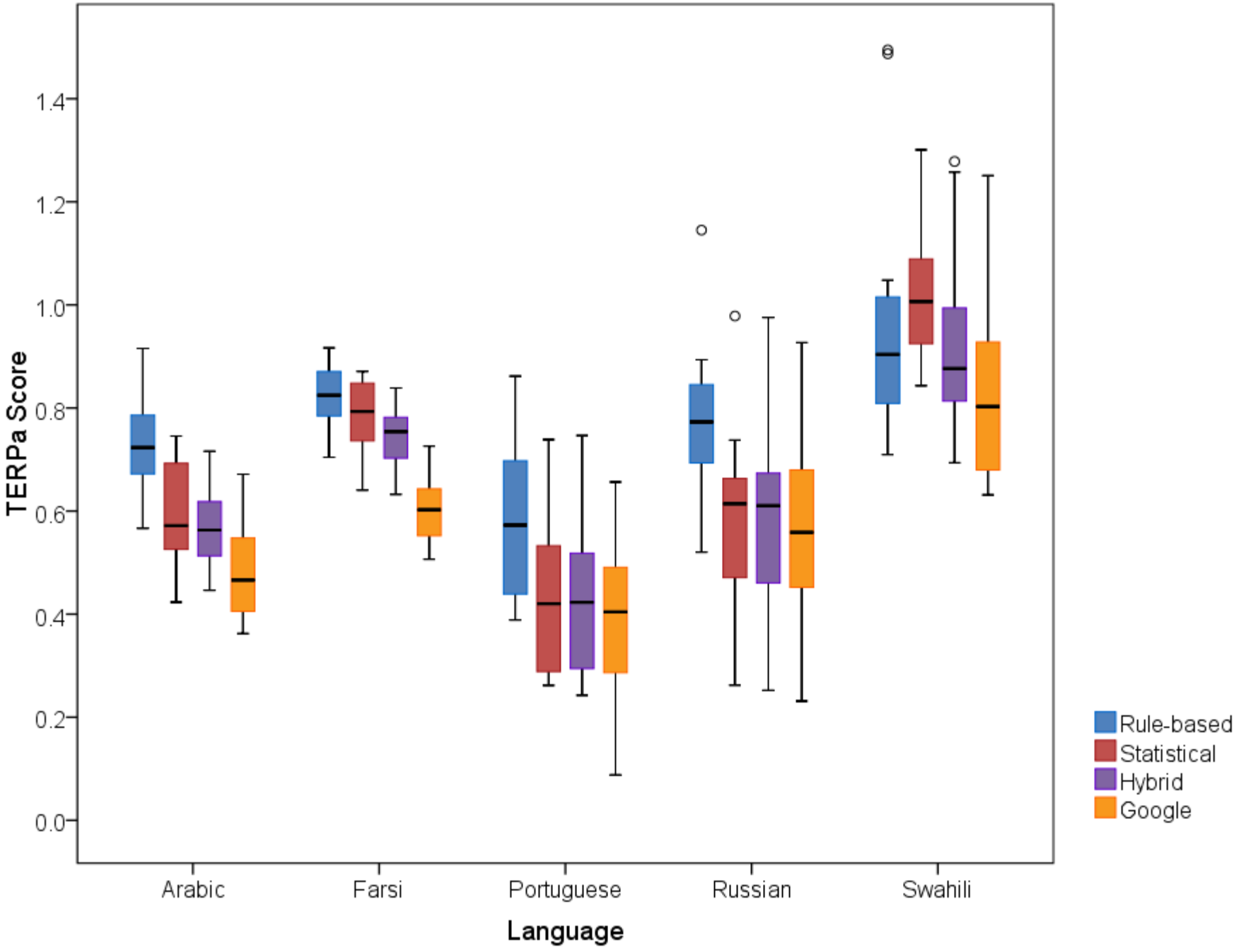
BLEU scoring

- 0 = poorest match
- 100 = perfect match
- Changes of 1-2 points considered “publishable”

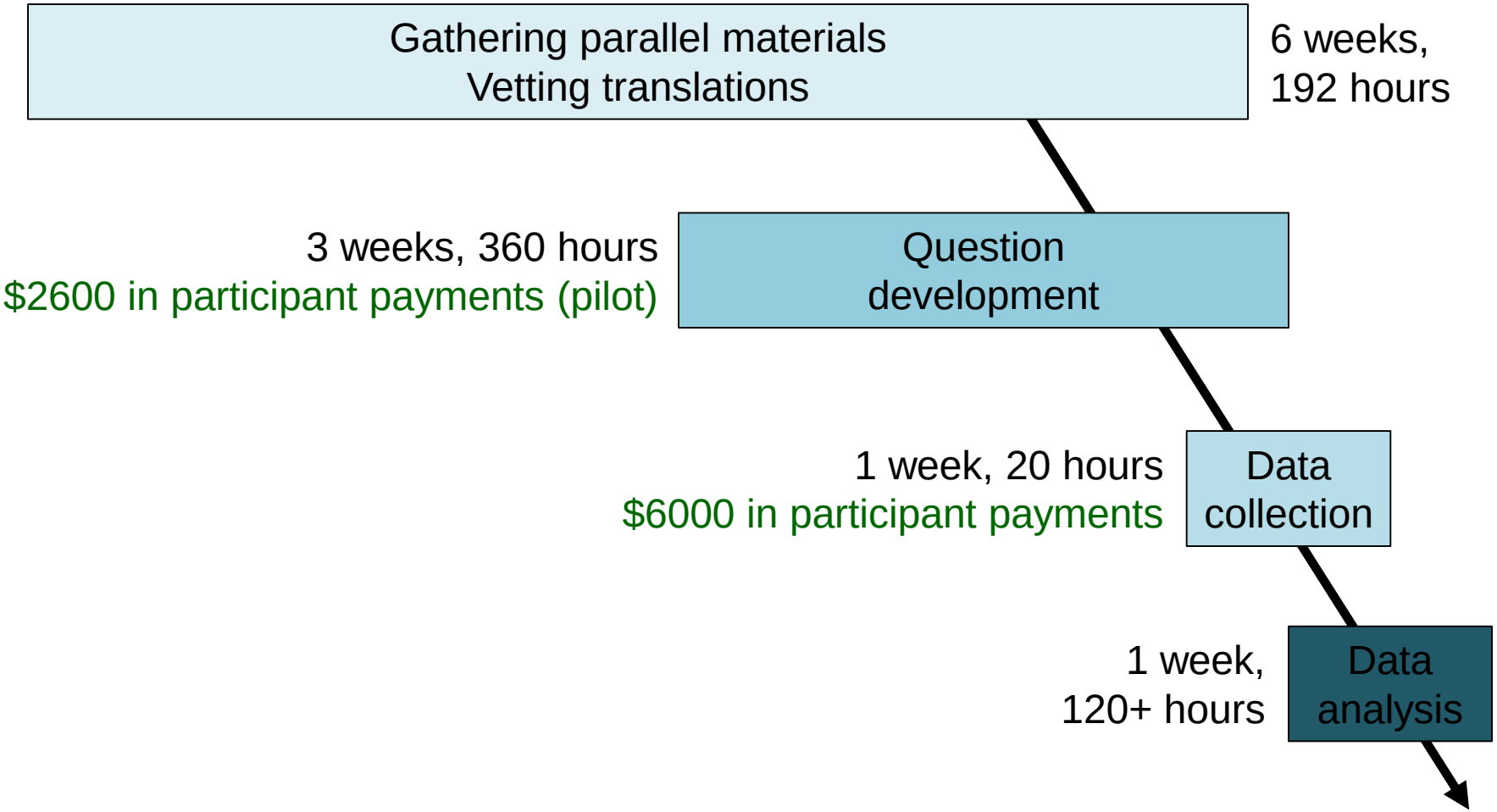

Comprehension Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
TERPa Score
r = -.45, p < .001
Arabic
Farsi
Portuguese
Russian
Swahili


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Comprehension Score
0
10
20
30
40
50
60
70
80
90
BLEU Score
r = .47, p < .001
Arabic
Farsi
Portuguese
Russian
Swahili


1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
TERPa Score
0
20
40
60
80
100
BLEU Score
r = -.82, p < .001
Arabic
Farsi
Portuguese
Russian
Swahili


TERPa Score
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Arabic
Farsi
Portuguese
Russian
Swahili
Language
Rule-based
Statistical
Hybrid
Google

Level of effort required



Summary

- Comprehension of MT output was well above chance, meaning users were able to get some information from the MT even though they found it difficult to understand.
- Improvements in BLEU scores were associated with improvements in comprehension.
- In most conditions, comprehension scores for Hybrid MT output were better than for Rule-based MT.

Future directions

- Collect data with documents produced in the source language
  - Requires time for translation(s) and verification
- Design comprehension questions that focus on essential elements of information
  - EEIs = information critical to decision making
  - Vary based on organization/mission (e.g., FEMA EEIs might differ from law enforcement EEIs)

Future directions (cont'd)

- Compare comprehension scores to paired comparison scores
  - Human judgment of MT output usually done at the sentence level, often with paired comparison of overall quality (e.g., WMT16)
  - Process is less labor-intensive than developing comprehension questions

CENTER FOR ADVANCED STUDY OF LANGUAGE

JOHNS HOPKINS
UNIVERSITY

human language technology
center of excellence

Questions on Part 1?

CENTER FOR ADVANCED STUDY OF LANGUAGE

Use of MT for translation and comprehension of Chinese texts

Design

- Two task conditions
  - Full translation
  - Comprehension questions (multiple choice, short answer)
- Three MT conditions
  - From Scratch: No MT (access to online dictionaries only)
  - Post-Editing: Static MT output (from Google Translate)
  - Interactive: Opportunity to interact with MT system
- Post-task and post-experiment questionnaires

Why interactive condition?

Example of how MT output changes depending on how the input is segmented

Translation 1	
在集成电路芯片设计过程中考虑不周全或恶意植入不受使用方控制的程序或电路	Not considered comprehensive or implant malicious programs or the circuit from the control of the use of the integrated circuit chip design process
Translation 2	
在集成电路芯片	In the integrated circuit chip
设计过程中	The design process
考虑	Consider
不周全或恶意植入	Not comprehensive or malicious implants
不受使用方控制的程序或电路	Without the use of the control of the program or circuit

Topics

- Six scientific abstracts on highly technical topics from mainland China journals (125-175 characters)
  - Text A: Embedding information in satellite imagery
  - Text B: Detecting malware intrusions in Windows
  - Text C: Advances in helicopter collision avoidance radar
  - Text D: Antimissile command and control systems
  - Text E: Anti-interference capabilities of radar
  - Text F: Optimization models for firing effectiveness

Text characteristics

No correlation, i.e., the fluency of MT output does not necessarily give you a clue to other aspects of its quality

Text	Inverse TERp scores[a]	Fluency Ratings[b]	Difficulty Ratings[c]	Familiarity Ratings[d]
A	1.61	2.3 (0.50)	4.1 (1.01)	3.6 (0.67)
B	1.30	2.3 (0.82)	3.8 (0.95)	3.1 (0.93)
C	1.49	3.2 (0.75)	3.2 (0.98)	3.2 (0.78)
D	1.33	2.8 (0.75)	4.2 (0.74)	3.7 (0.66)
E	1.56	2.2 (0.83)	4.5 (0.76)	3.8 (0.52)
F	2.08	2.5 (0.93)	3.6 (1.00)	3.4 (0.76)

[a] TERp (Translation Error Rate plus) = relative to gold standard; higher = more similar

[b] 1 = not at all fluent; 5 extremely fluent (12 independent raters)

[c] 1 = very easy; 5 = very difficult

[d] 1 = very familiar; 4 = not at all familiar

Text characteristics

Significant correlation, i.e., more difficult texts were also rated as less familiar

Text	Inverse TERp scores[a]	Fluency Ratings[b]	Difficulty Ratings[c]	Familiarity Ratings[d]
A	1.61	2.3 (0.50)	4.1 (1.01)	3.6 (0.67)
B	1.30	2.3 (0.82)	3.8 (0.95)	3.1 (0.93)
C	1.49	3.2 (0.75)	3.2 (0.98)	3.2 (0.78)
D	1.33	2.8 (0.75)	4.2 (0.74)	3.7 (0.66)
E	1.56	2.2 (0.83)	4.5 (0.76)	3.8 (0.52)
F	2.08	2.5 (0.93)	3.6 (1.00)	3.4 (0.76)

[a] TERp (Translation Error Rate plus) = relative to gold standard; higher = more similar

[b] 1 = not at all fluent; 5 extremely fluent (12 independent raters)

[c] 1 = very easy; 5 = very difficult

[d] 1 = very familiar; 4 = not at all familiar

Text characteristics

Proficiency self-ratings (not shown) negatively correlated with difficulty ratings, but not with familiarity ratings

Text	Inverse TERp scores[a]	Fluency Ratings[b]	Difficulty Ratings[c]	Familiarity Ratings[d]
A	1.61	2.3 (0.50)	4.1 (1.01)	3.6 (0.67)
B	1.30	2.3 (0.82)	3.8 (0.95)	3.1 (0.93)
C	1.49	3.2 (0.75)	3.2 (0.98)	3.2 (0.78)
D	1.33	2.8 (0.75)	4.2 (0.74)	3.7 (0.66)
E	1.56	2.2 (0.83)	4.5 (0.76)	3.8 (0.52)
F	2.08	2.5 (0.93)	3.6 (1.00)	3.4 (0.76)

[a] TERp (Translation Error Rate plus) = relative to gold standard; higher = more similar

[b] 1 = not at all fluent; 5 extremely fluent (12 independent raters)

[c] 1 = very easy; 5 = very difficult

[d] 1 = very familiar; 4 = not at all familiar

Sample comprehension questions

- Multiple choice: 3 questions standard across all passages

1. Select the topic **domain** to which this article belongs.
   a) Information Technology
   b) Detection Technology
   c) Nuclear Technology
   d) Medical Technology
   e) Command and Control
   f) Weapons Technology

2. Select three **keywords** that describe this passage.
   a) Hidden Information
   b) Detection Systems
   c) Cyber Security
   d) Safety Testing
   e) Communication Systems
   f) Perception & Cognition
   g) Manufacturing
   h) Robotics
   i) Imagery/Imaging

3. Select the best descriptive **title** for this passage.
   a) Detecting Malware Intrusions
   b) The Use of Behavioral Characteristics
   c) RootKit for Windows
   d) Reliable Windows Technology

Sample comprehension questions

- Multiple choice: 2 questions unique to each passage
  - On which of the following is the detection technique based?
    a) Hook System Call
    b) Hard System Call
    c) Root System Call
    d) Lock System Call
- True or false
  - According to the author, this detection method is completely reliable. (T/F)
- Short answer (expected response = a few words)
  - What kind of attack is the technique designed to detect?
- Open-ended (expected response = a sentence)
  - How does the detection technique work?

Demographics

- *n* = 51 (23 males, 28 females)
- Average age = 23.5 years (range = 18-58)
- 45 reported at least some college education
- Recruitment targeted advanced undergraduate learners of Chinese
  - All native English speakers
  - 4 also native Chinese (heritage speakers)

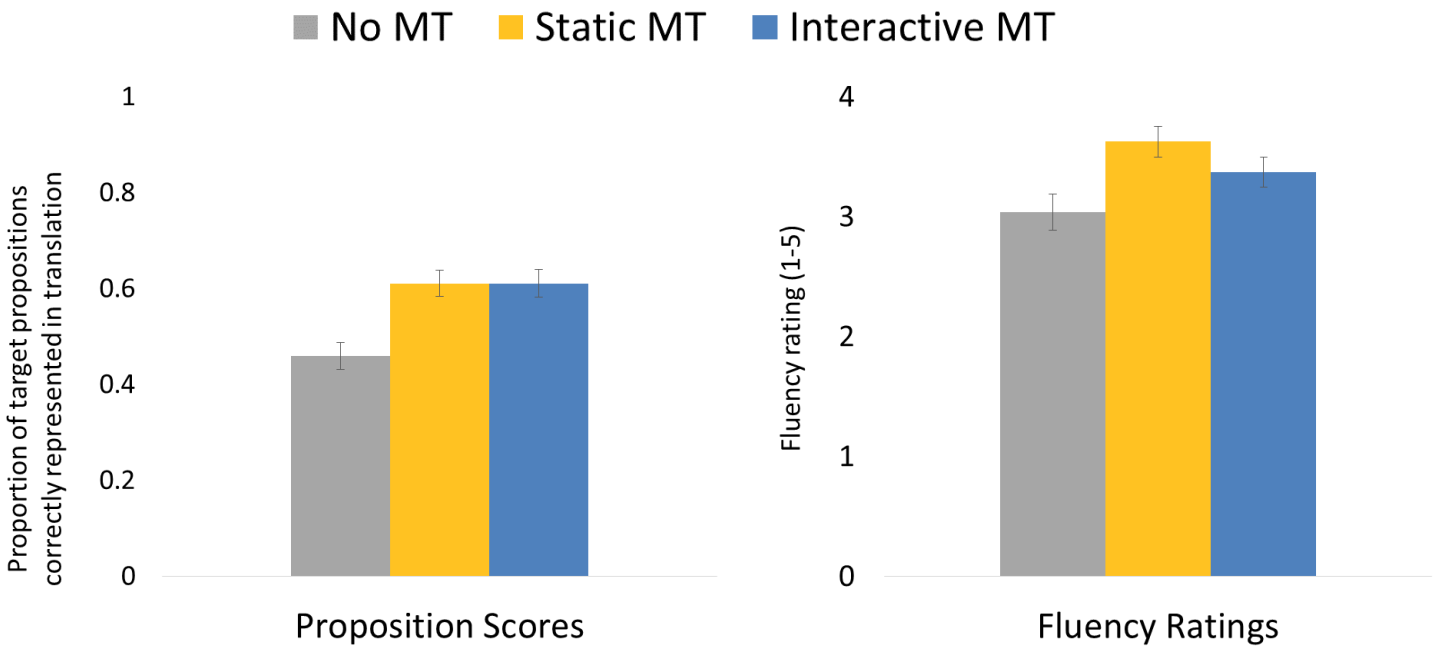
Chinese experience

- Average months of study = 43.2 (range = 6-168)
- Most had taken at least one course at the 300-level or higher
- Immersion experience: *n* = 43
  - Most in teens or 20's
  - Most for 1-5 months

Translation outcome measures

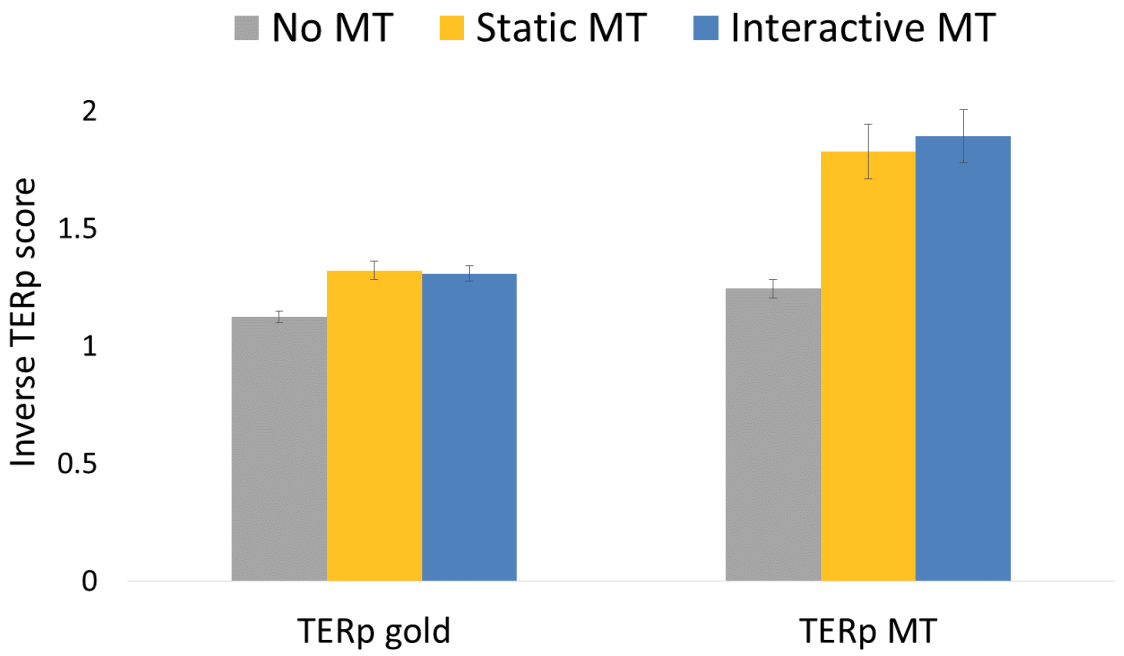
- Proposition score (accuracy)
  - Scored by research team; counted presence or absence of pre-determined set of “propositions”
- Fluency ratings
  - Judged by 3 members of the research team
- Inverse TERp scores (higher = more similar)
  - Distance from gold standard
  - Distance from MT output
- Time on task

Translation quality



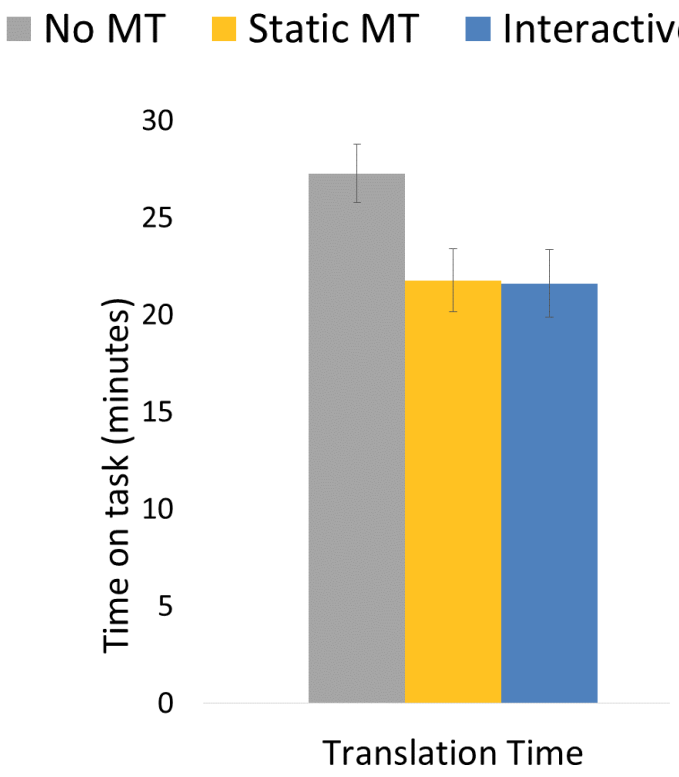
Proposition scores and fluency ratings were significantly lower in the No MT condition than in the two MT conditions, which did not differ significantly from each other.

TERp scores



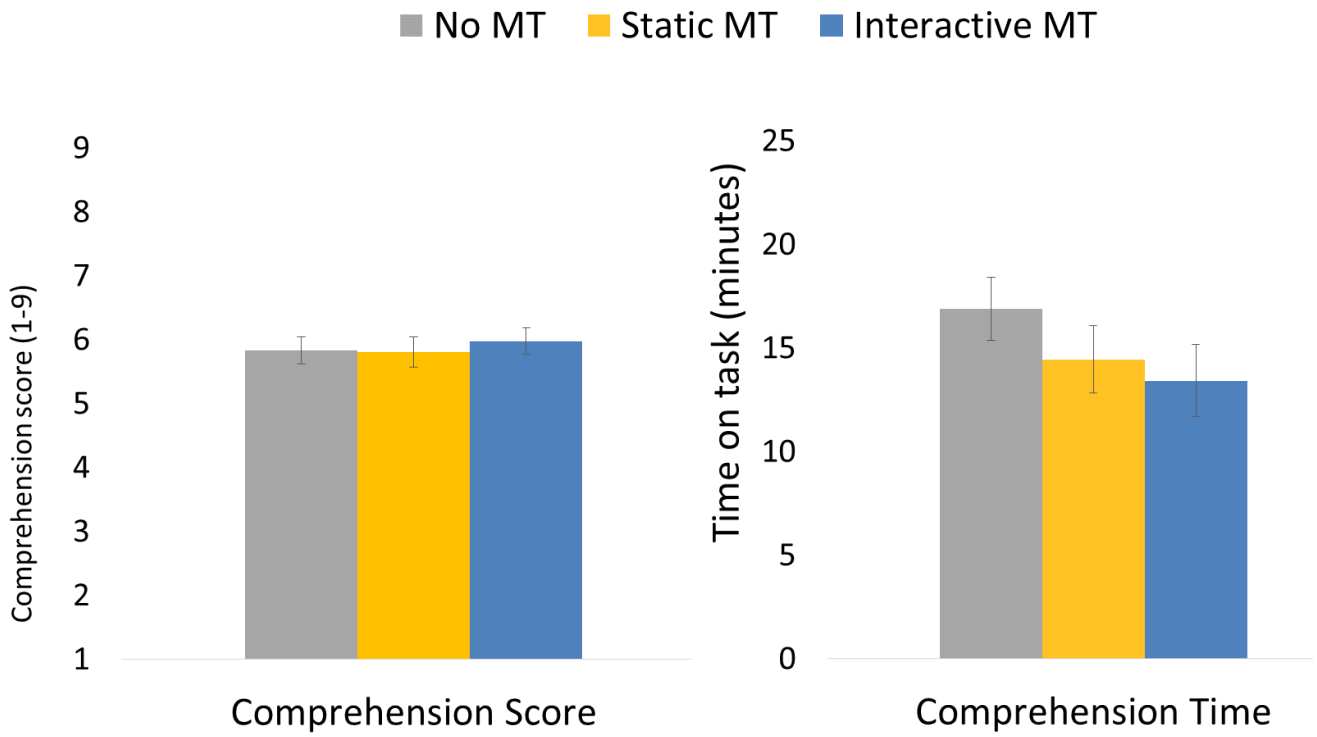
Translations produced in the No MT condition were significantly more distant from both the gold standard and the MT output than were translations produced in the two MT conditions, which did not differ significantly from each other.

Translation time on task



Participants took significantly longer to generate their translations in the No MT condition than in the two MT conditions, which did not differ significantly from each other.

Comprehension outcomes



- Comprehension scores did not differ significantly across conditions.
- Participants took significantly longer to answer the comprehension questions in the No MT condition than in the two MT conditions, which did not differ significantly from each other.

Comprehension vs. Translation

- How did you approach the task differently than you might have if you were writing a translation?
  - *Tried to get a "gist" of what was being said.*
  - *Focused on keywords / concepts rather than a cohesive translation.*
  - *Spent less time worrying about grammar.*
  - *I just trusted the provided MT output.*
  - *A lot of questions were related to very specific aspects of the text.*
  - *I had to really understand the meanings of technical terms, as they were needed to respond to the questions.*

Reasons for not interacting with MT

- Translation condition
  - *I decided to stick with using the online dictionary. I did check with the MT output to correlate with what I looked up.*
  - *I have had better experiences using the online dictionary.*
  - *I felt the MT output, dictionary, and my own knowledge and experience were sufficient.*
- Comprehension condition
  - *I used the dictionary more.*
  - *MT output sufficient.*
  - *Because I felt that it would be the same.*

Summary and conclusions

- Differential benefits of MT for translation and comprehension tasks
  - Notable that MT was sometimes helpful even when the quality of the output was relatively poor
  - For comprehension, MT may have helped participants get the gist more quickly, but helped less with details
- Interacting with MT did not help
  - Lack of use; perception that MT would not change
  - Translators may need training on how to use MT most effectively

Future directions

- Other languages and/or tasks
- Effects of proficiency
  - Lower: need more help
  - Higher: better able to discern when MT is misleading
  - Any benefit for native speakers?
- Other domains and genres
  - e.g., colloquial material without complete, grammatical sentences
- Combinations of tools

CENTER FOR ADVANCED
STUDY OF LANGUAGE

Questions on Part 2?

CENTER FOR ADVANCED
STUDY OF LANGUAGE

Utility of translation memory (TM) in an operational context

Motivation

- An USG organization with access to some CAT tools (MT, concordance search, on-line dictionaries)
- TM resources limited
  - Termbase sharing is manual
  - Labor-intensive to import termbases
- CAT resources have a history of performing differently in an USG context

Materials

- Provided by client
- Mixed genre
  - Transcribed/translated conversations
  - Translated written correspondence
  - Translated documents
  - Targeted summaries (discarded)
- Alignment
  - Often at the paragraph level

Approach

- Worked with 3 experienced analysts to align text
  - Source sentence as maximum length
  - Often could get to the level of an English phrase
- Formatted aligned translation units (TUs) as TMX

Choosing a comparison group

- "Ideal" TM scenario
  - Phone manuals
  - Within-text matches
  - Between text matches
- Scientific abstracts (very narrow domain)

Choosing a metric

- Is TM providing as many matches in the USG context?
  - N-gram matches
  - TU matches

Results

- The USG material had substantially fewer n-gram matches than any of the comparison groups
- No comparison for TU matches
- 30 matches within 1000 TUs

Unique applications

- Event-driven terminology
  - Пятое кольцо
  - “Fifth ring”

Unique applications

- Event-driven terminology
  - Пятое кольцо
  - “Fifth ring”

  - 2014 Sochi Olympics
  - Meaning: inept performance/inept performer or failure to complete a job

Unique applications

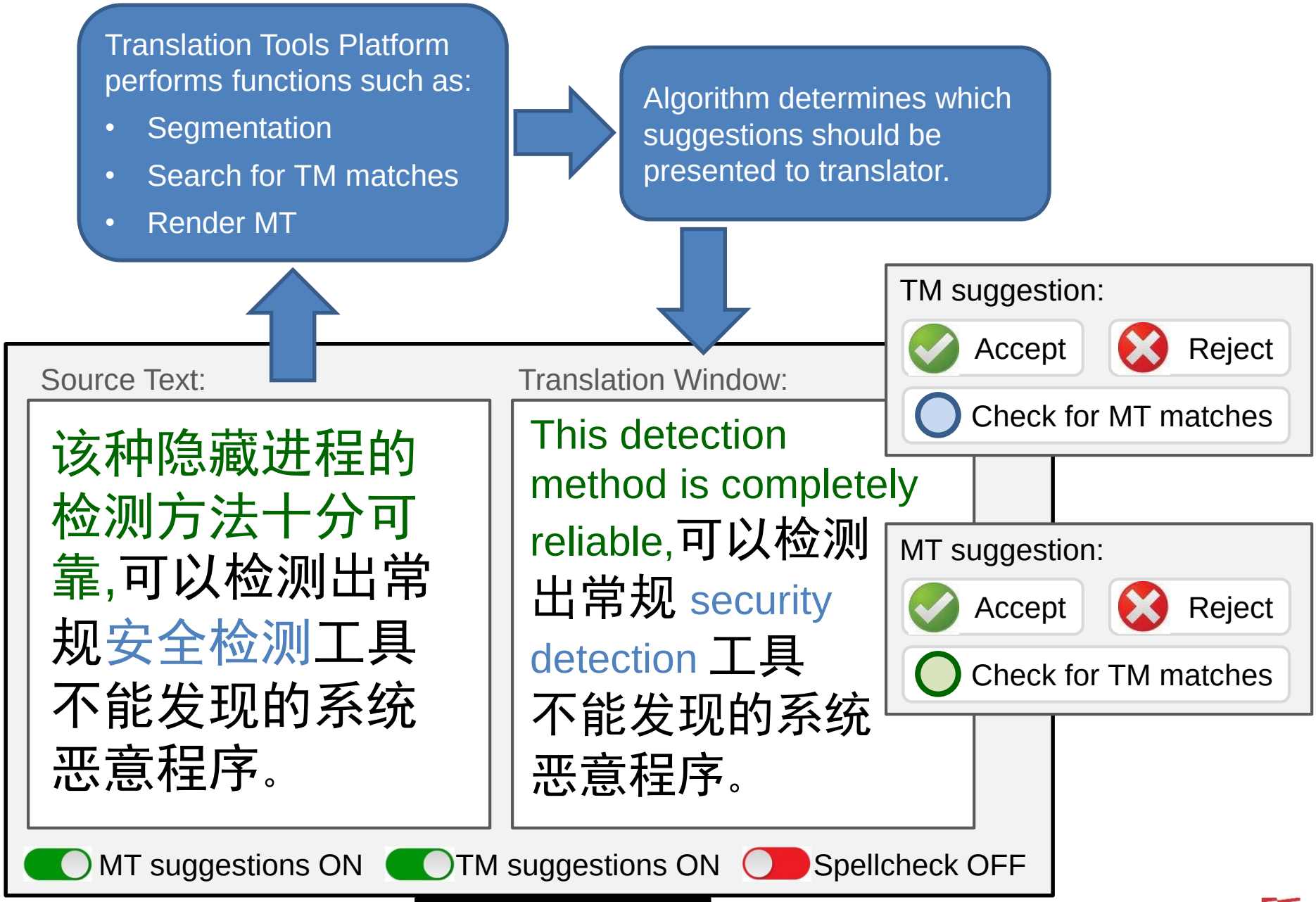
- Connecting the dots
  - Evolving terminology used to “talk around” topics or refer to individuals
  - Individuals using the same terminology may be connected in some way
  - Alerting analysts to areas of mission overlap and areas for potential collaboration

Unique applications

- Entity/event disambiguation
  - Less than exact matches may prompt analysts to investigate whether individuals, organizations, or events with similar names are in fact the same individual / organization / event
  - Access to prior reports allows quick access to data that may assist with disambiguation

Unique applications

- Entity/event disambiguation
  - International Council of Scientific Unions (1931-1998)
  - Changed to International Council for Science (1998-present)


Translation Tools Platform performs functions such as:
• Segmentation
• Search for TM matches
• Render MT
Algorithm determines which suggestions should be presented to translator.
TM suggestion:
Accept
Reject
Check for MT matches
Source Text:
该种隐藏进程的检测方法十分可靠,可以检测出常规安全检测工具不能发现的系统恶意程序。
Translation Window:
This detection method is completely reliable,可以检测出常规 security detection 工具不能发现的系统恶意程序。
MT suggestion:
Accept
Reject
Check for TM matches
MT suggestions ON
TM suggestions ON
Spellcheck OFF

Obstacles to adoption

- At the level of the organization
  - Monetary/development investment
  - Need to invest in data
- At the level of the individual
  - Want seamless access/input
  - Resistant to additional burden
  - Want to be able to opt-in

Desired features

- Supported linking to:
  - Databases
  - Prior reporting
  - Prior authors
- Glossary creation
- Domain tagging
- Support for multiple languages

Summary and conclusions

- TM may provide fewer matches from memory compared to TM used in an "ideal" scenario
- May improve at scale
  - Will never approach the size of termbases of outside information
- Additional applications for TM may make the investment uniquely beneficial

CENTER FOR ADVANCED
STUDY OF LANGUAGE

Questions about Part 3?

Overall conclusions

- Important to consider the task/use case
- Even big changes in BLEU score might not be reflected in comprehension
- Both automatic evaluations and human comparisons are abstractions compared to human comprehension
- Any CAT tools:
  - may face adoption hurdles
  - may offer unanticipated benefits