

Biases and perils of MT evaluation

Martin Popel

ÚFAL (Institute of Formal and Applied Linguistics)
Charles University, Prague



November 19th 2019, Workshop on document level MT evaluation, Luxembourg

Outline

① Perils of human translation

② Types of MT evaluation

③ Perils of manual MT evaluation

WMT19

Domain & translationese

④ Perils of automatic evaluation

③B Fluency vs adequacy

Document
vs
Sentence
level

infeurope

CROSSLANG
TRANSLATION AUTOMATION

ES Team

⑤ Fun fair

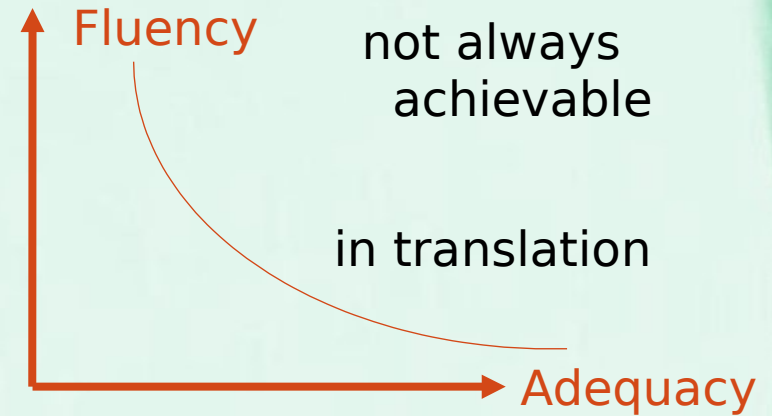


In my presentation, I focus on the following aspects of evaluations: sentence-level vs. document-level vs. document-aware, source-based vs. reference-based, direct assessment vs. comparison-based, fluency-biased vs. adequacy-biased. I also discuss the aspect of translationese and native target/source-language translators and evaluators. Finally, I will present some results from the WMT19 Metrics shared task.

Antonio Toral: Addressing claims of human parity

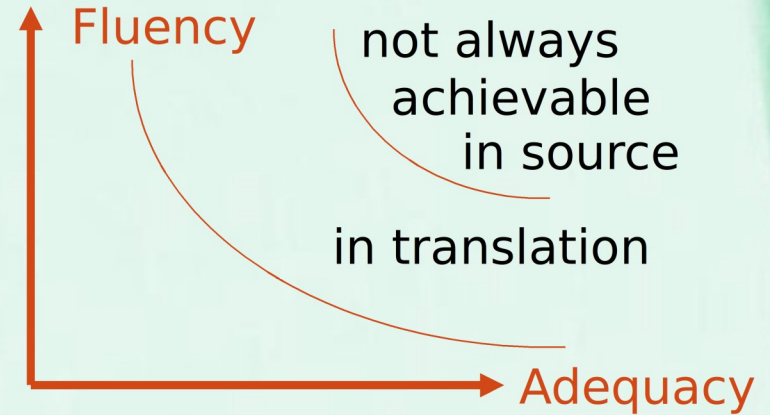
Perils of human translation

- Fluency vs. Adequacy



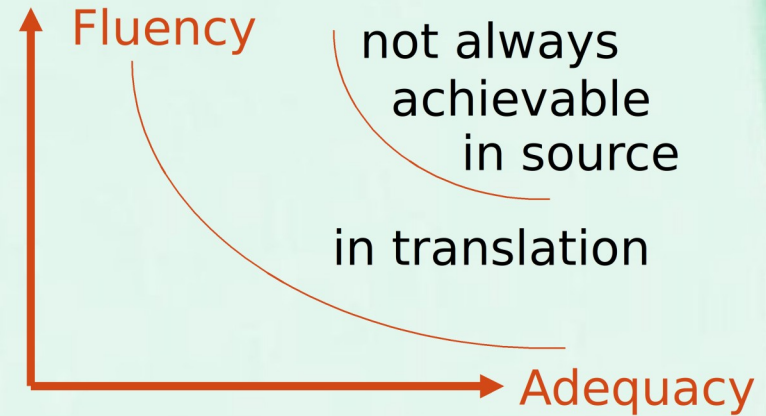
Perils of human translation

- Fluency vs. Adequacy
- Intent \Rightarrow source language \Rightarrow translation



Perils of human translation

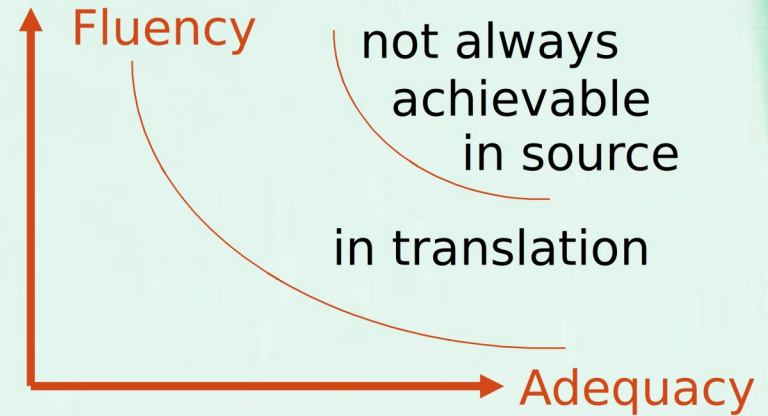
- Fluency vs. Adequacy
- Intent \Rightarrow source language \Rightarrow translation



- Semantics vs. Pragmatics. Example:
 - SRC: *I'm not going to worry too much about it.*
 - REF: *I believe everything will be OK.*

Perils of human translation

- Fluency vs. Adequacy
- Intent \Rightarrow source language \Rightarrow translation



- Semantics vs. Pragmatics. Example:
 - SRC: *I'm not going to worry too much about it.*
 - REF: *I believe everything will be OK.*

	not worry	worry
believe OK	usual	rare
don't believe OK	rare	usual

Types of manual MT evaluation

- REF-based ... show candidate and (human) reference
- SRC-based ... show candidate and source sentence
- REF&SRC-based ... show both

Types of manual MT evaluation

- REF-based ... show candidate and (human) reference
 - SRC-based ... show candidate and source sentence
 - REF&SRC-based ... show both
-
- RR = Relative Ranking ... relative, ordinal, N systems
 - DA = Direct Assessment ... “absolute”, continuous, 1 system
 - RankME = rank-based magnitude estimation ... continuous, N sys

Types of manual MT evaluation

- REF-based ... show candidate and (human) reference
- SRC-based ... show candidate and source sentence
- REF&SRC-based ... show both
- Sentence-level
- Document-level ... single score per document
- Document-aware ... show whole documents, scores per sentence
- RR = Relative Ranking ... relative, ordinal, N systems
- DA = Direct Assessment ... “absolute”, continuous, 1 system
- RankME = rank-based magnitude estimation ... continuous, N sys

Example: REF&SRC sent-level RR (WMT10–16)

"Valentino měl vždycky raději eleganci než slávu.

— Source

Valentino has always preferred elegance to notoriety.

— Reference

Best ← Rank 1 Rank 2 Rank 3 Rank 4 Rank 5 → Worst

"Valentino should always elegance rather than fame.

— Translation 1

Best ← Rank 1 Rank 2 Rank 3 Rank 4 Rank 5 → Worst

"Valentino has always rather than the elegance of glory.

— Translation 2

Best ← Rank 1 Rank 2 Rank 3 Rank 4 Rank 5 → Worst

"Valentino had always preferred elegance than glory.

— Translation 3

Best ← Rank 1 Rank 2 Rank 3 Rank 4 Rank 5 → Worst

"Valentino has always had the elegance rather than glory.

— Translation 4

Best ← Rank 1 Rank 2 Rank 3 Rank 4 Rank 5 → Worst

"Valentino has always had a rather than the elegance of the glory.

— Translation 5

Example: SRC doc-level DA (WMT19)

Document

WMT19DocSrcDA #202:Document
#independent.226349-10

English — German (deutsch)

Below are the sentences you have just rated as a single **document**. Please state how much you agree that:

The black text adequately expresses the meaning of the gray text in German (deutsch).

Russian Grand Prix: Lewis Hamilton closes in on world title after team orders hand him win over Sebastian Vettel It became clear from the moment that Valtteri Bottas qualified ahead of Lewis Hamilton on Saturday that Mercedes' team orders would play a large part in the race. From pole, Bottas got a good start and almost hung Hamilton out to dry as he defended his position in the first two turns and invited Vettel to attack his teammate. Vettel went into the pits first and left Hamilton to run into the traffic at the tail of the pack, something which should have been decisive. The Mercedes pitted a lap later and came out behind Vettel, but Hamilton went ahead after some wheel-to-wheel action that saw the Ferrari driver reluctantly leave the inside free at risk of holding out after a double-move to defend on the third corner. Max Verstappen started from the back row of the grid and was in seventh by the end of the first lap on his 21st birthday. He then led for a large part of the race as he held onto his tyres to target a quick finish and overtake Kimi Raikkonen for fourth. He eventually came into the pits on the 44th lap but was unable to increase his pace in the remaining eight laps as Raikkonen took fourth. It's a difficult day because Valtteri did a fantastic job all weekend and was a real gentleman told let me by. The team have done such an exceptional job to have a one two," said Hamilton.

— Source text

Großer Preis von Russland: Lewis Hamilton schließt auf Weltmeistertitel ein, nachdem ihm das Team den Sieg über Sebastian Vettel überlassen hat Es wurde von dem Moment an klar, dass Valtteri Bottas sich vor Lewis Hamilton am Samstag qualifiziert hatte, dass die Teamaufträge von Mercedes eine große Rolle im Rennen spielen würden. Von der Pole aus erwischte Bottas einen guten Start und ließ Hamilton fast trocken, als er seine Position in den ersten beiden Kurven verteidigte und Vettel einlud, seinen Teamkollegen anzugreifen. Vettel ging zuerst in die Gruben und verließ Hamilton, um am Rucksack in den Verkehr zu geraten, was entscheidend gewesen sein sollte. Der Mercedes drehte eine Runde später und kam hinter Vettel, aber Hamilton ging nach einigen Rad-an-Rad-Aktion, die sah, dass der Ferrari-Fahrer widerwillig verlassen die Innenseite frei in Gefahr zu halten, nach einem Doppelschlag auf der dritten Ecke zu verteidigen. Max Verstappen startete aus der hinteren Startreihe und wurde am Ende der ersten Runde an seinem 21. Geburtstag Siebter. Er führte dann für einen großen Teil des Rennens, als er auf seinen Reifen hielt, um ein schnelles Ziel zu erreichen und Kimi Räikkönen zum vierten Mal zu überholen. In der 44. Runde kam er schließlich in die Box, konnte aber sein Tempo in den verbleibenden acht Runden nicht erhöhen, da Räikkönen den vierten Platz belegte. Es ist ein schwieriger Tag, denn Valtteri hat das ganze Wochenende einen fantastischen Job gemacht und war ein echter Gentleman, der mir gesagt hat. Das Team hat so einen außergewöhnlichen Job gemacht, um ein, zwei zu haben", sagte Hamilton.

— Candidate translation

0%

100%

Reset

Submit

Example: pseudo doc-aware DA (WMT19)

Sentence pair

WMT19DocSrcDA #281:Document #reuters.218861-0

English → German (deutsch)

For the pair of **sentences** below: Read the text and state how much you agree that:

The black text adequately expresses the meaning of the gray text in German (deutsch).

North Korea says 'no way' will disarm unilaterally without trust

— Source text

Nordkorea sagt , Sprünge ohne Vertrauen entwaffnen ohne Vertrauen .

— Candidate translation



Reset

Submit

Example: pseudo doc-aware DA (WMT19)

Sentence pair

WMT19DocSrcDA #281:Document #reuters.218861-0

English → German (deutsch)

For the pair of **sentences** below: Read the text and state how much you agree that:

The black text adequately expresses the meaning of the gray text in German (deutsch).

North Korea says 'no way' will disarm unilaterally without trust

— Source text

Nordkorea sagt , Sprünge ohne Vertrauen entwaffnen ohne Vertrauen .



— Candidate translation

0%

100%

Reset

Submit

 This is the GitHub version [#wmt19dev](#) of the Appraise evaluation system.  Some rights reserved.  Developed and maintained by [Christian Federmann](#).

sentences in doc order, but one sentence per screen and no undo/back button

Example: SRC doc-aware 10-RankME

	G	H	I	J	K	L	M	N	O	P
1	Source	Translation1	T1_overall	T1_adequacy	T1_fluency	Translation2	T2_overall	T2_adequacy	T2_fluency	Optional comment
168	"And we're protecting our shareholders from employment litigation."									
169	Companies started taking ethics, values and employee engagement more seriously in 2002 after accounting firm Arthur Andersen collapsed because of ethical violations from the Enron scandal, Quinlan said.									
170	But it wasn't until "social media came into its own" that companies realized they couldn't stop their dirty laundry from going viral online.									
171	"Prior to using technology to monitor ethics, people used hope as a strategy," he said.									
172	Both Glint and Convercent offer their software as a service, charging companies recurring fees to use their products.									
173	It's a business model and opportunity that has the approval of venture capital investors, who have propped up both start-ups.	Je to obchodní model a příležitost, kterou schvalují odvážní kapitáloví investoři, jenž podpořili oba start-upy.	7	6	7	Je to obchodní model a příležitost, která má souhlas investorů rizikového kapitálu, kteří podpořili oba start-upy.	10	10	10	T1: chybný překlad termínu "venture capital"
174	Convercent raised \$10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to \$47 million.	Convercent vybral v rámci své únorové kampaně od firem jako Sapphire Ventures a Tola Capital celkově 10 milionů \$ a nakonec si odnesl kapitál ve výši 47 milionů \$.	3	4	3	Convercent získal v únoru finanční prostředky ve výši 10 milionů dolarů od firem jako Sapphire Ventures a Tola Capital, čímž se jeho celkový kapitál zvýšil na 47 milionů dolarů.	10	10	10	
175	Glint secured \$10 million in November from Bessemer Venture Partners, bringing its total funding to \$60 million.	Glint získal v listopadu 10 milionů \$ od Bessemer Venture Partners a v průběhu celé kampaně získal 60 milionů \$.	5	4	5	Glint získal v listopadu 10 milionů dolarů od společnosti Bessemer Venture Partners, čímž jeho celkové financování dosáhlo 60 milionů dolarů.	10	10	10	
176	These investments hardly come as a surprise, given the interconnected nature of companies, culture and venture capital.	Tyto investice jsou stěží překvapující vzhledem k vzájemné povaze společností, kultury a rizikovému kapitálu.	3	4	3	Tyto investice nejsou vzhledem k propojenosti společností, kultury a rizikového kapitálu žádným překvapením.	10	10	10	
177	There's a growing body of research showing today's employees expect more from their workplaces than before.	Narůstající počet výzkumů jasně potvrzuje, že dnešní zaměstnanci očekávají od svého pracoviště více než kdy dříve.	5	5	8	Roste množství výzkumů, které ukazují, že dnešní zaměstnanci očekávají od svých pracovišť více než dříve.	10	10	10	
178	In competitive markets such as Silicon Valley, high salaries and interesting projects are merely table stakes.	A na konkurenčních trzích, jakým je např. Silicon Valley, nejsou hlavní výhodou vysoké platy a zajímavé projekty.	6	5	8	Na konkurenčních trzích, jako je Silicon Valley, jsou vysoké platy a zajímavé projekty pouhými sázkami u stolu.	7	8	7	problém: význam termínu "table stakes"
179	Employees want to feel that they're accepted and valued and that they're giving their time to a company with a positive mission.	Zaměstnanci chtějí vnímat, že jsou přijímáni a oceňováni a že věnují svůj čas společnosti, která usiluje o pozitivní poslání.	9	9	9	Zaměstnanci chtějí mít pocit, že jsou přijímáni a ceněni a že věnují svůj čas společnosti s pozitivním posláním.	10	10	9	

Perils of manual MT evaluation

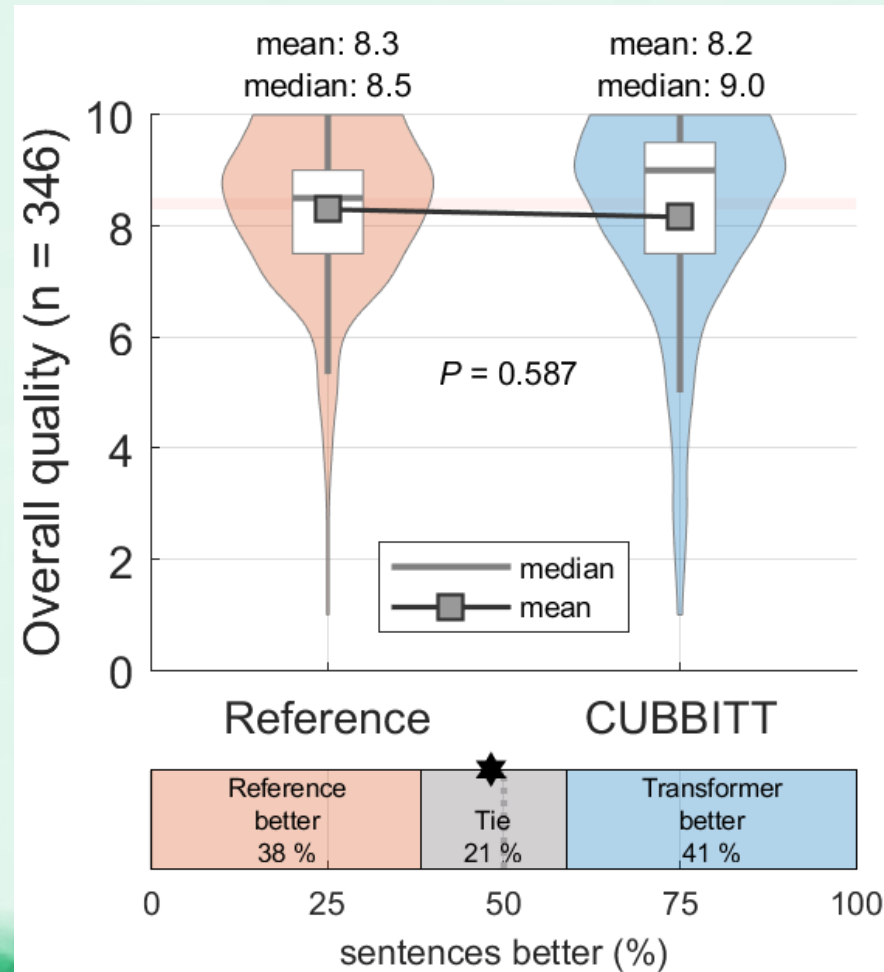
Each type of **evaluation is biased** towards some systems.

- REF-based ... similarity to human errors (or post-editing)
- SRC-based ... problems with non-professional evaluators
- REF&SRC-based ... both

- Sent-level ... false positives and false negatives (fluency+adeq.)
- Doc-level ... too coarse, psychological problems
- Doc-aware ... how to approximate doc-level? Avg, min...?

- RR ... tiny improvements/errors same as big ones
- DA ... fluency and serious adequacy errors only (but faster)
- RankME ... slower, difficult if $N > 3$

Which system is better? Median vs Mean?



Perils of automatic evaluation

BLEU (& other REF-similarity metrics) has 3 issues:

- Not enough REFs, i.e. low coverage of correct translations.
- Differences in BLEU do not correlate with Human scores (even with enough REFs).
- Human REFs may be worse than MT.

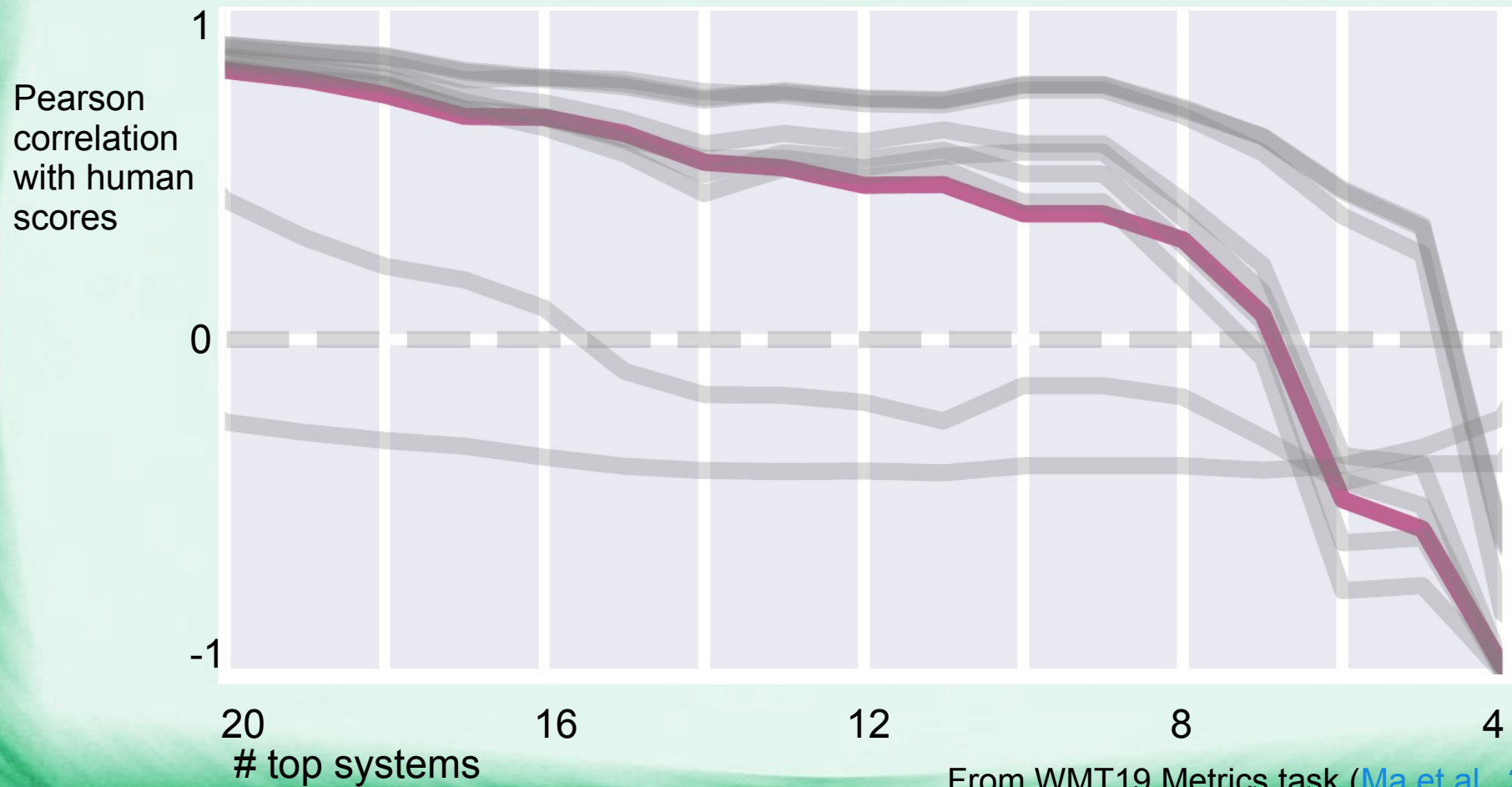
If REF is a translation:

- It may not be adequate and fluent (if non-professional translator).

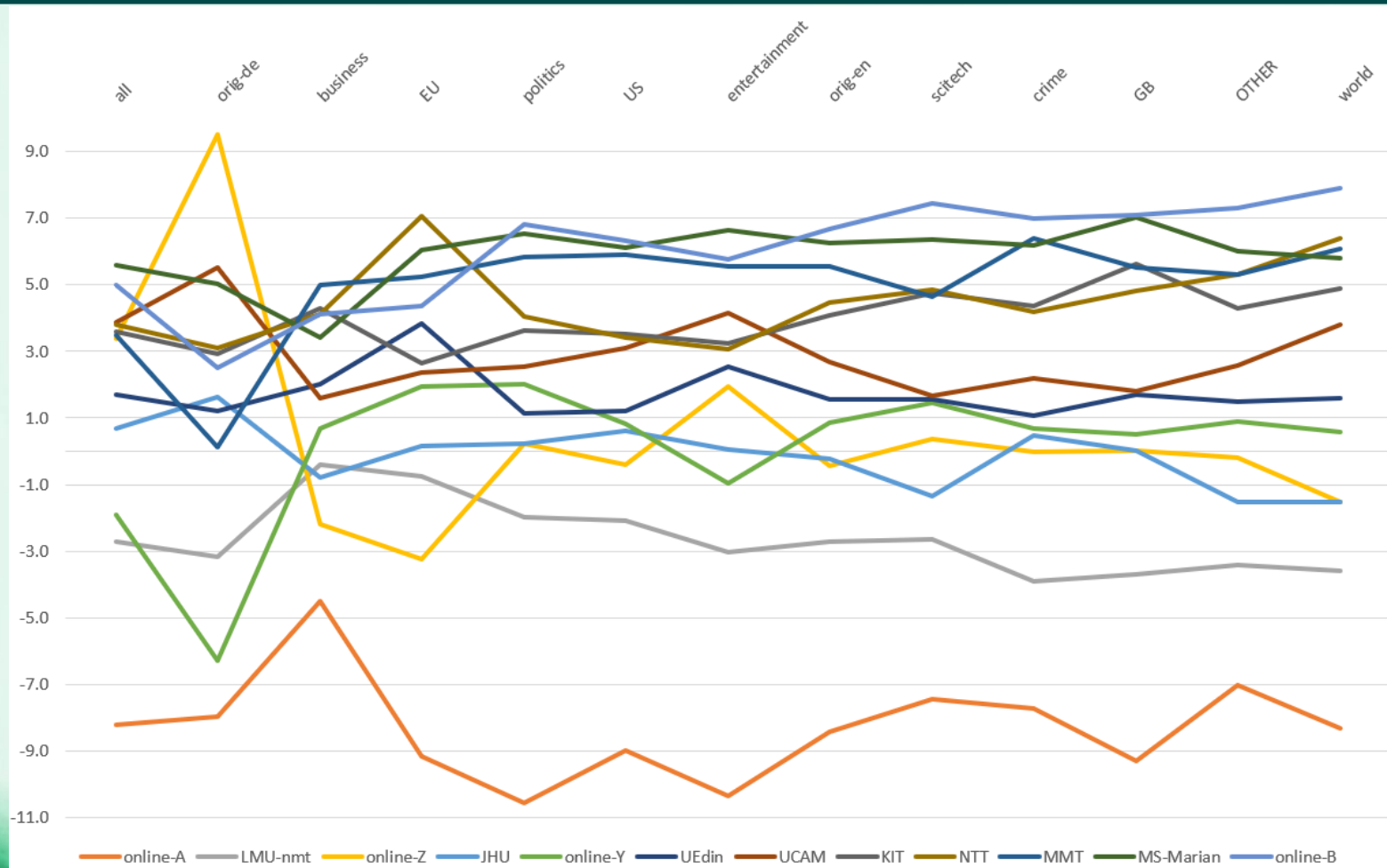
If REF is the original sentence (reverse-direction eval):

- SRC is not original, thus may not have the same meaning as REF
- SRC is likely not representative of the expected use case (domain/country & translationese)
- (tiny) risk of non-perfect adequacy+fluency of REF relative to the intent

BLEU does not correlate with humans for strong systems



Domain&orig-lang effect: BLEU WMT18 EN→DE



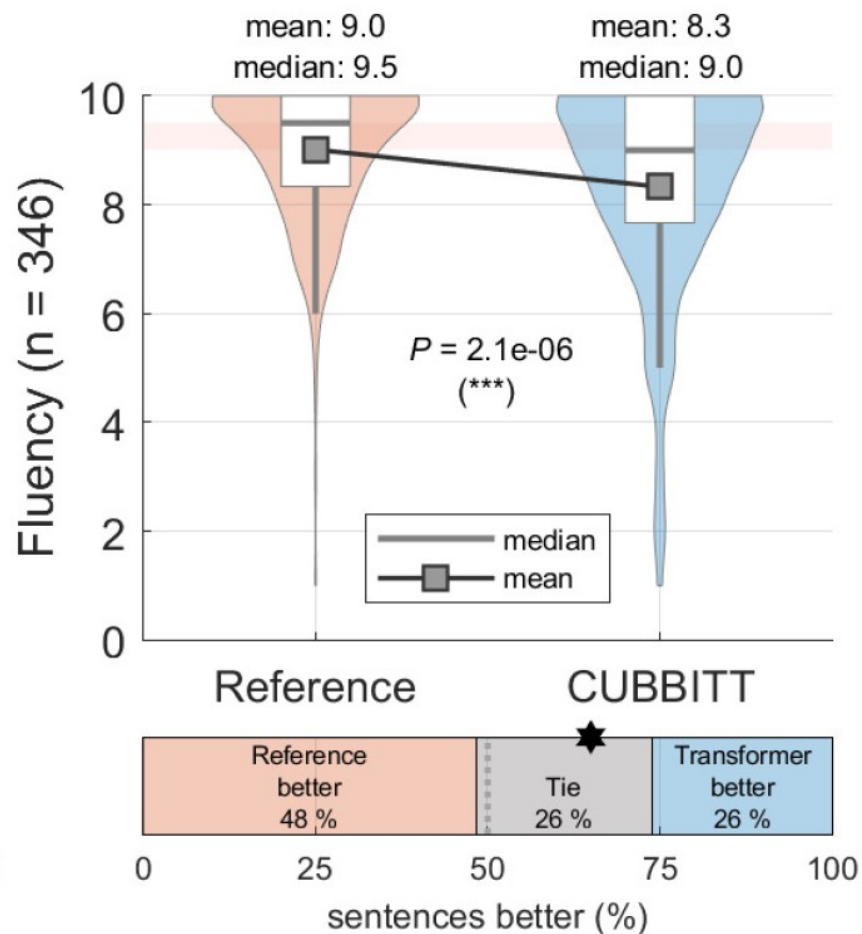
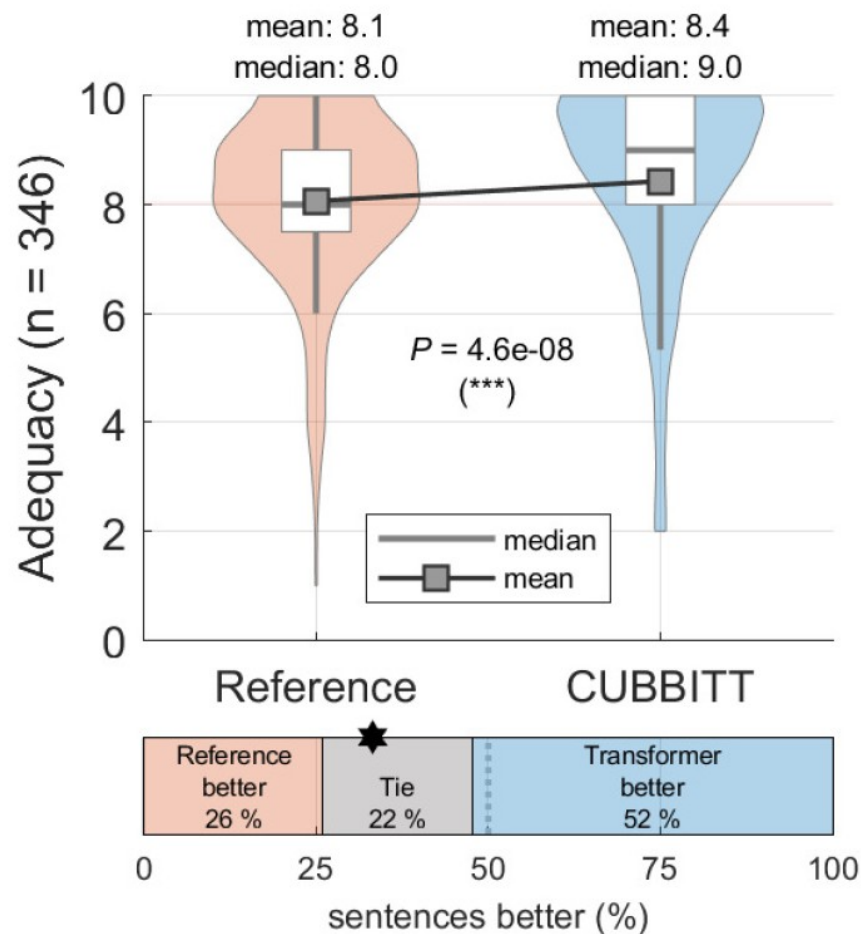
sacrebleu
--detail

Is NMT better in fluency or adequacy?

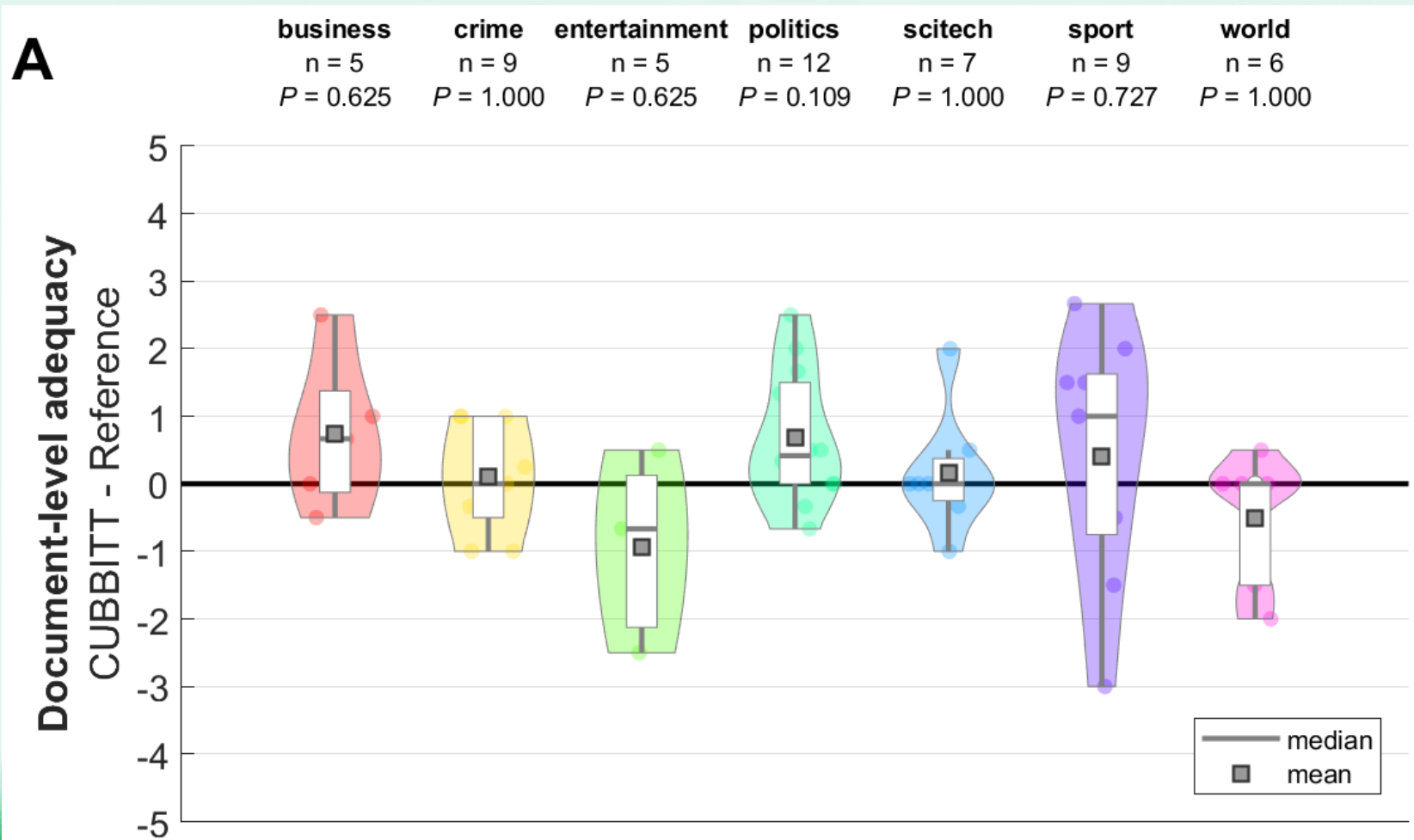
Is NMT better in fluency or adequacy?



Is NMT better in fluency or adequacy?

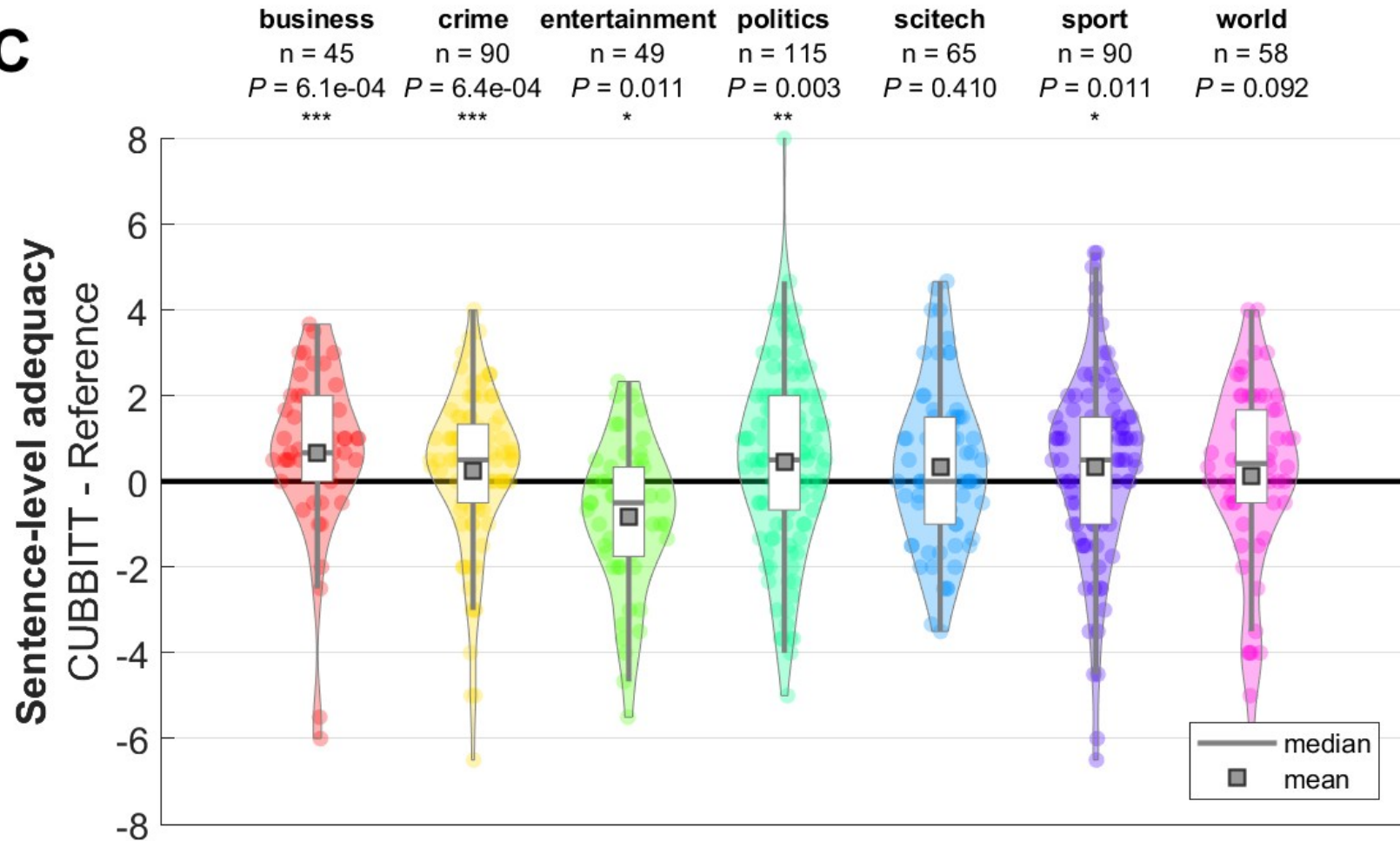


Domain effect: manual doc-level adequacy



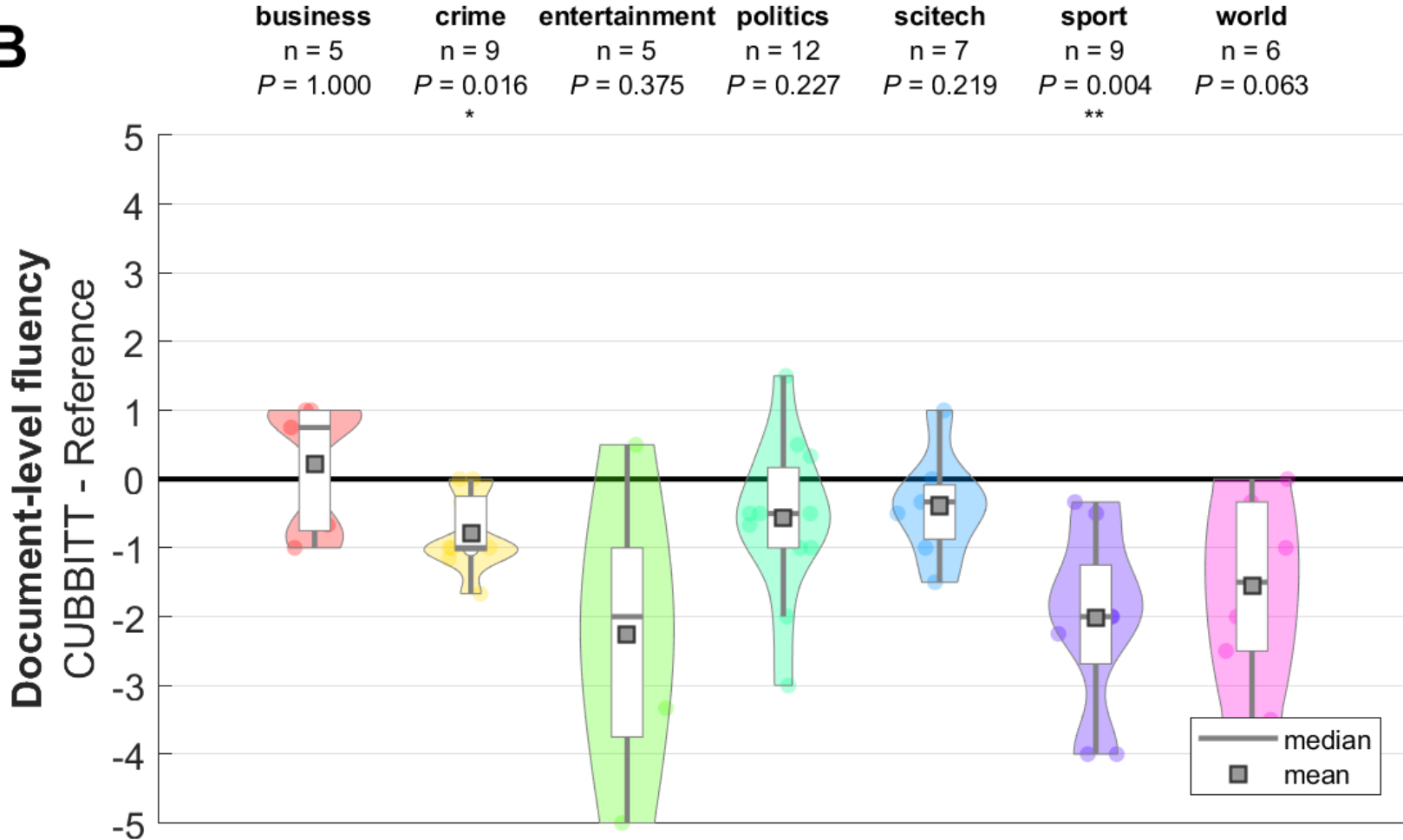
Domain effect: manual sent-level adequacy

C

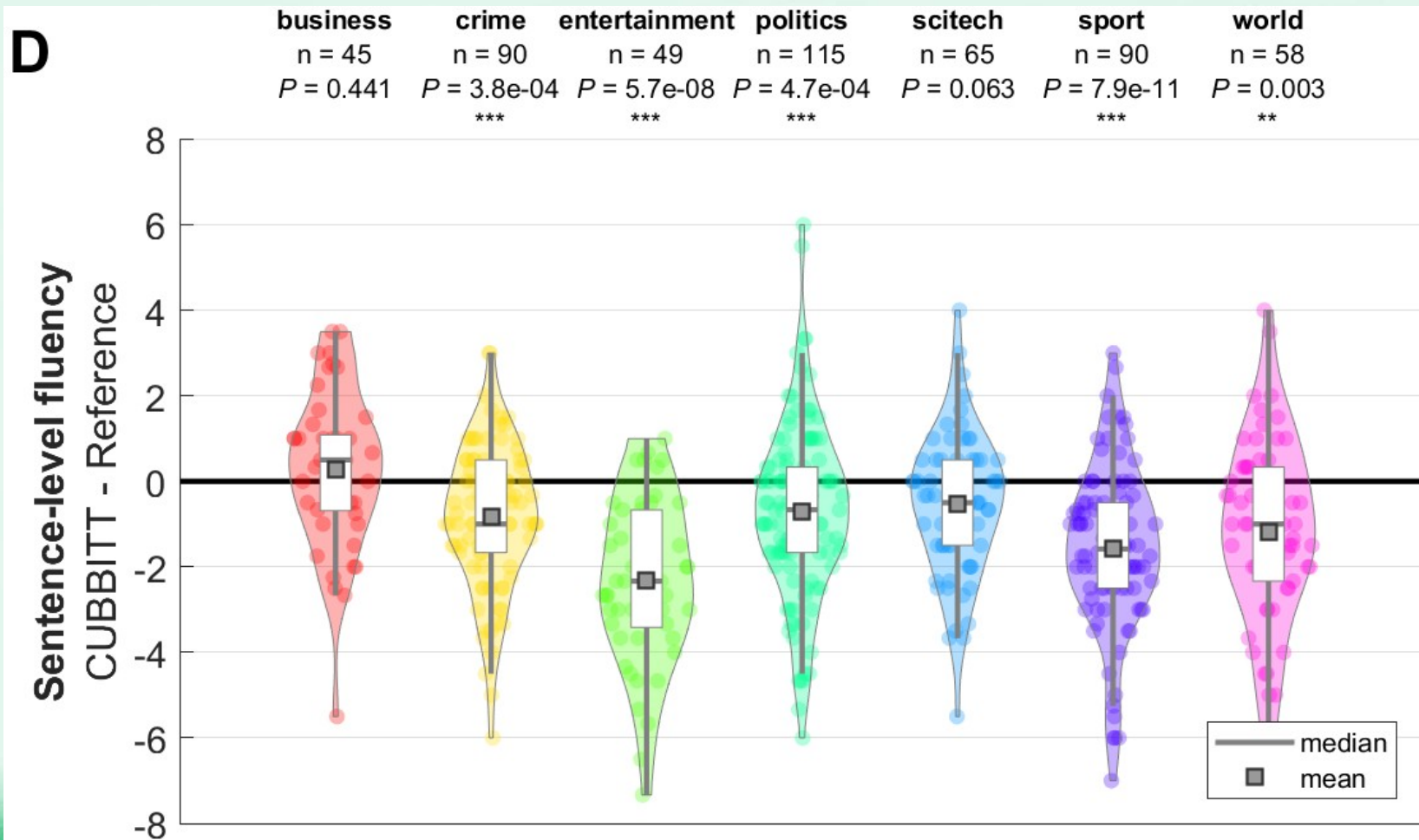


Domain effect: manual doc-level fluency

B

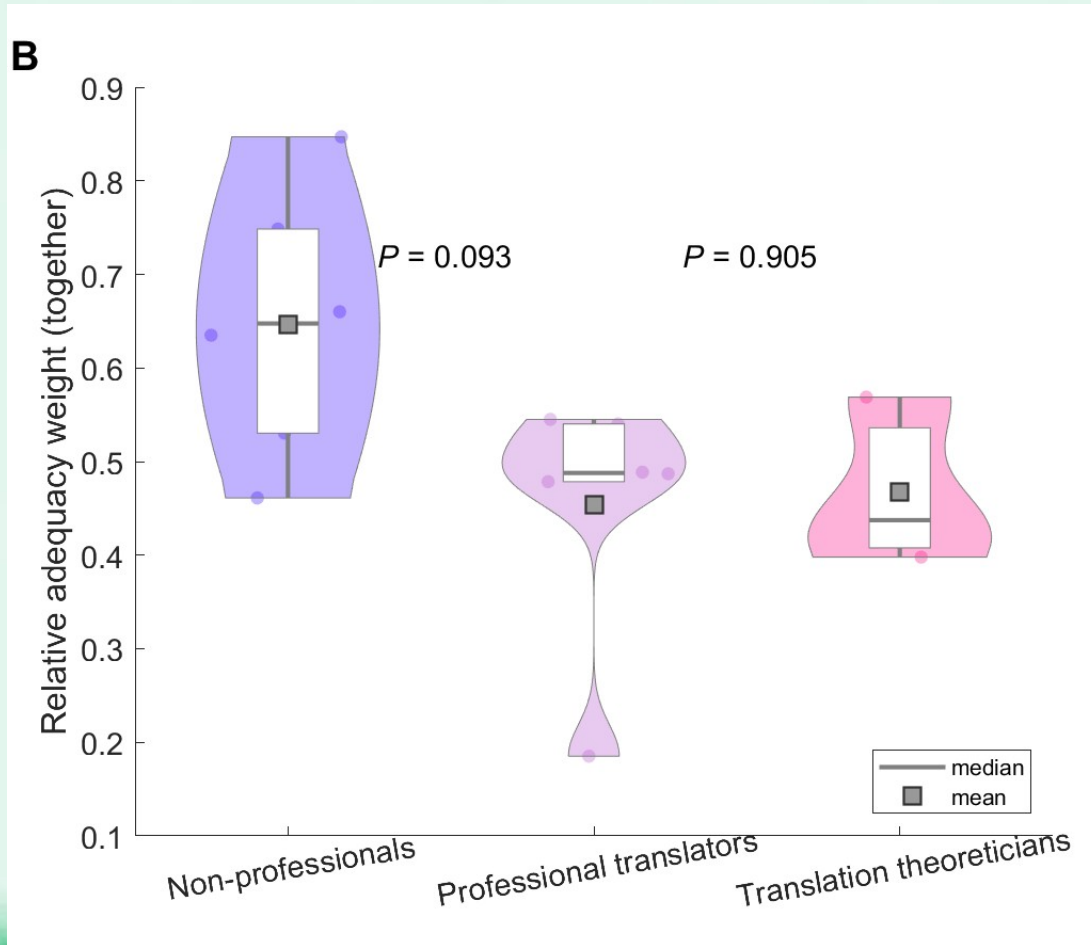


Domain effect: manual sent-level fluency

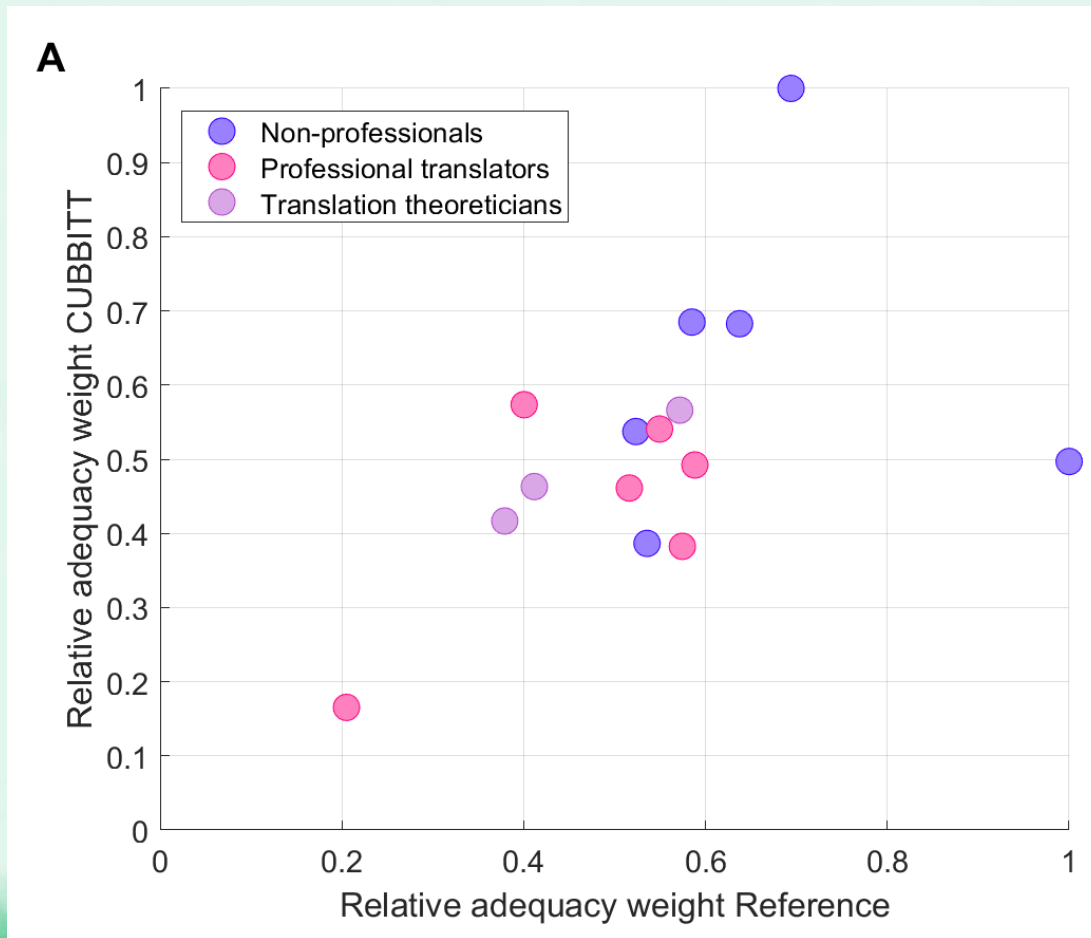


Overall quality = $x \cdot \text{adequacy} + (1-x) \cdot \text{fluency}$?

Overall quality = $x \cdot \text{adequacy} + (1-x) \cdot \text{fluency}$?



Overall quality = $x \cdot \text{adequacy} + (1-x) \cdot \text{fluency}$?



Perils of adequacy w.r.t. purpose/localization

Burning Man → *Matějská pout'* (St. Matthew's Funfair)

