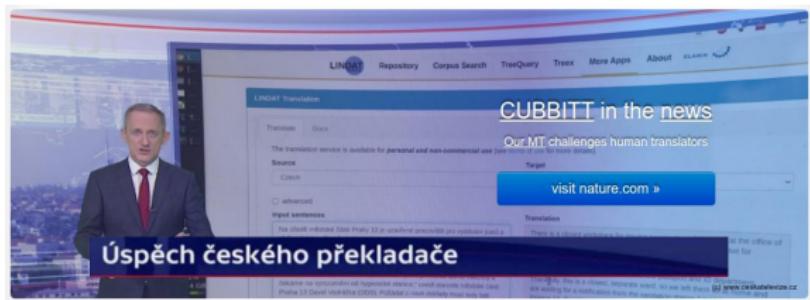
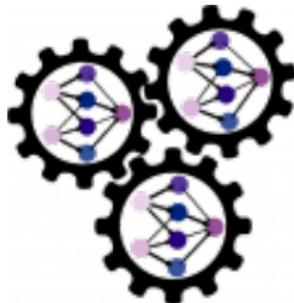


# CUBBITT

## Charles University Block-Backtranslation-Improved Transformer Translation

Mgr. Martin Popel, Ph.D.

ÚFAL (Institute of Formal and Applied Linguistics), MFF UK



<https://lindat.cz/cubbitt>

## nature communications

---

nature > nature communications > articles > article

Article | Open Access | Published: 01 September 2020

### Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals

Martin Popel , Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar & Zdeněk Žabokrtský

*Nature Communications* 11, Article number: 4381 (2020) | Cite this article

6273 Accesses | 76 Altmetric | Metrics

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders

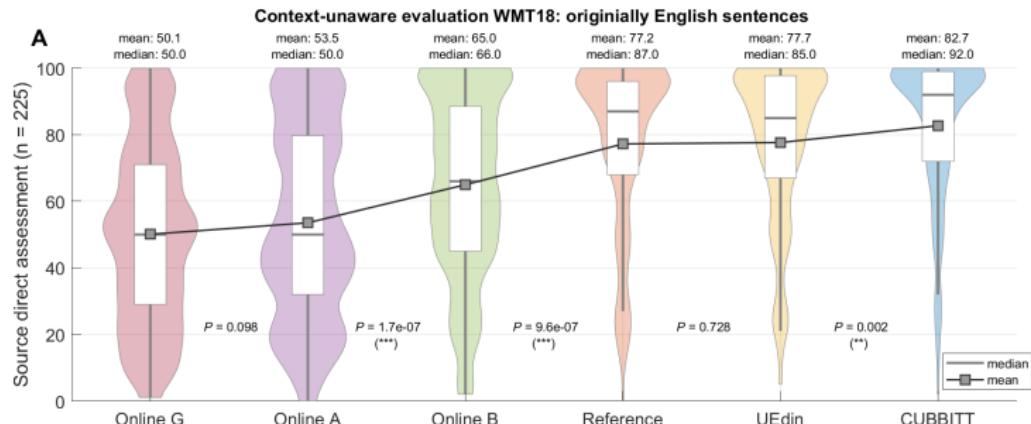
- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model

# History of the research

3

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.



- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.
- 2018/07 finished my PhD thesis

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.
- 2018/07 finished my PhD thesis
- 2018/08 new doc-aware evaluations, contacting Łukasz

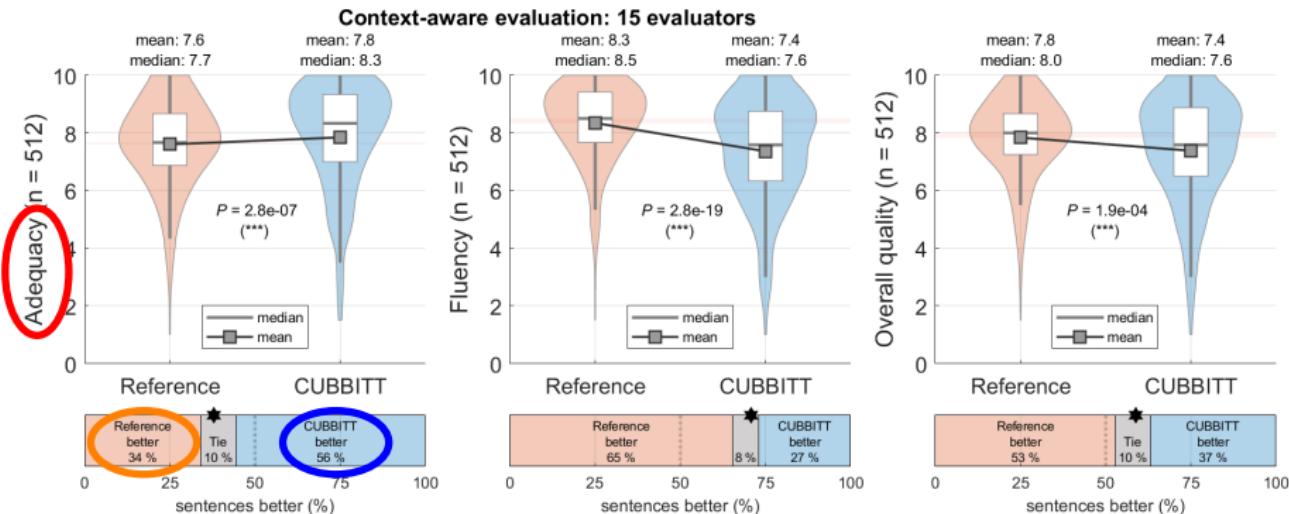
- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.
- 2018/07 finished my PhD thesis
- 2018/08 new doc-aware evaluations, contacting Łukasz
- 2018/12 submission to Science, 2019/02 reviews, /08 reject

- professors Sgall, Hajičová, Panevová, Hajič, Kuboň...  
2007 my interest in CL & NLP
- 2012/02 Markéta&Jakub use NN for t-lemma translation
- 2017/01 I use DeepNN for Doctors Without Borders
- 2017/06 Attention Is All You Need (TectoMT→Transformer)
- 2018/04 Training Tips for the Transformer Model
- 2018/07 WMT18: CUBBITT signif. better than human ref.
- 2018/07 finished my PhD thesis
- 2018/08 new doc-aware evaluations, contacting Łukasz
- 2018/12 submission to Science, 2019/02 reviews, /08 reject
- 2019/09 submission to NatCommun, 2020/02 reviews,  
experiments, /04 resubmit, 07/accepted, 09/published

# The main result of our evaluation

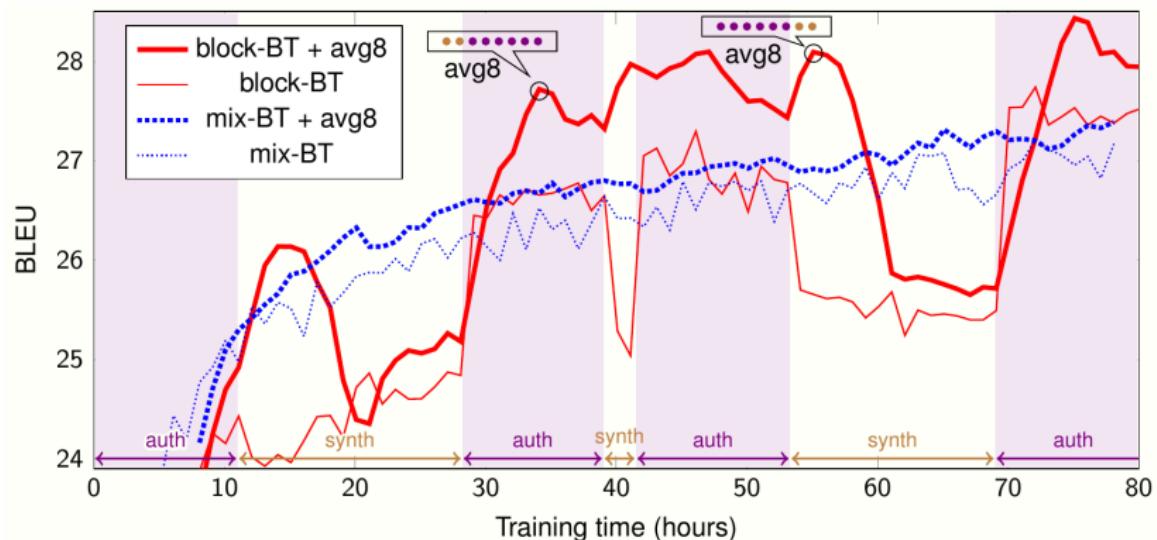
4

56 % sentences translated more adequately by CUBBITT,  
34 % sentences by a professional translation agency



# The main innovation: Block-BackTranslation

5



# Manual evaluation example

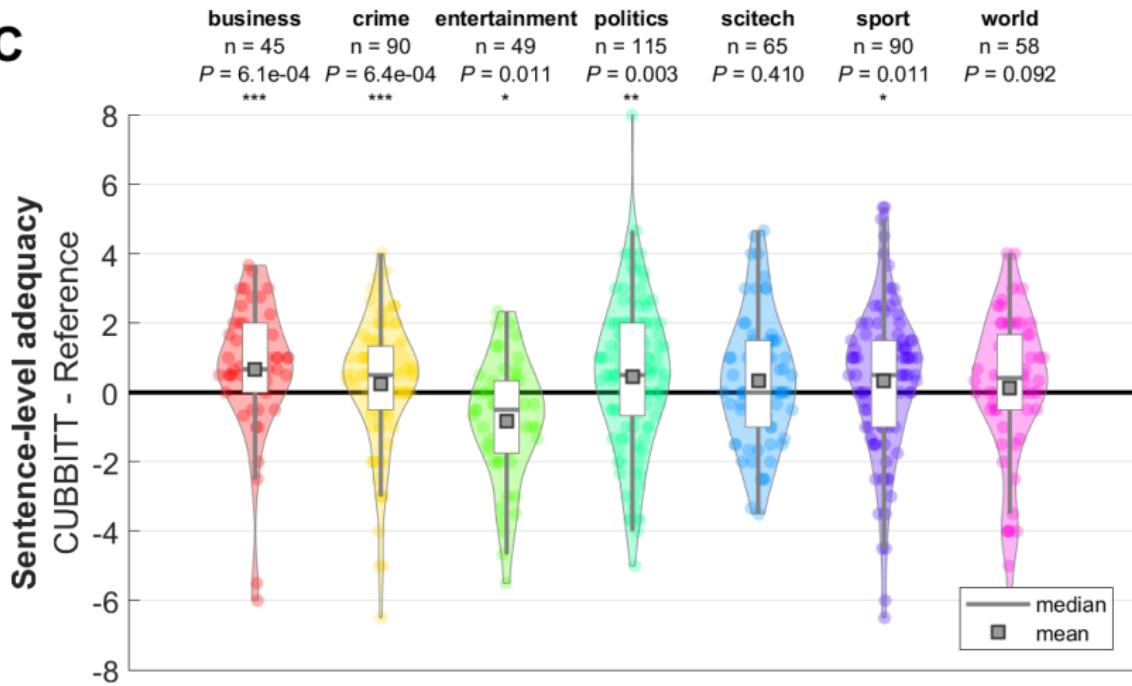
blind, sentence-level but document-aware, side-by-side (**RankME**)

	G	H	I	J	K	L	M	N	O	P
1	Source	Translation1	T1_overall	T1_adeqacy	T1_fluency	Translation2	T2_overall	T2_adeqacy	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, Quinal 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 Convergent 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 schvaluji odvážní kapitáloví investoři, jenž podporují oba startupy.	7	6	7	Je to obchodní model a příležitost, která má souhlas investorů rizikového kapitálu, kteří podporují oba start-upy.	10	10	10	T1: chybný překlad termínu "venture capital"
174	Convergent raised \$10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to \$47 million.	Convergent vybral v rámci své únorové kampaně od frenů jako Sapphire Ventures a Tola Capital celkově 10 milionů \$. A nakonec si odnesl kapitál ve výši 47 milionů \$.	3	4	3	Convergent 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 společnosti, kultury a rizikovému kapitálu.	3	4	3	Tyto investice nejsou vzhledem k propojenosti společnosti, kultury a rizikového kapitálu zádným 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	5	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, jsou 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 terminu "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	

# It depends on the input text domain

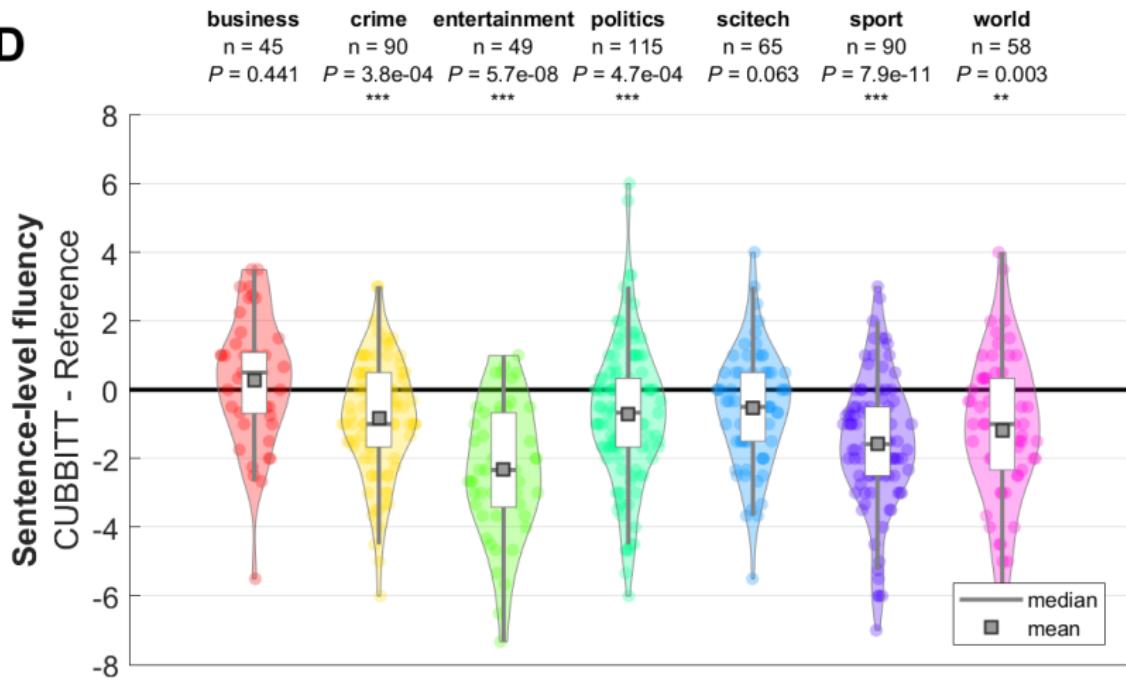
7

C



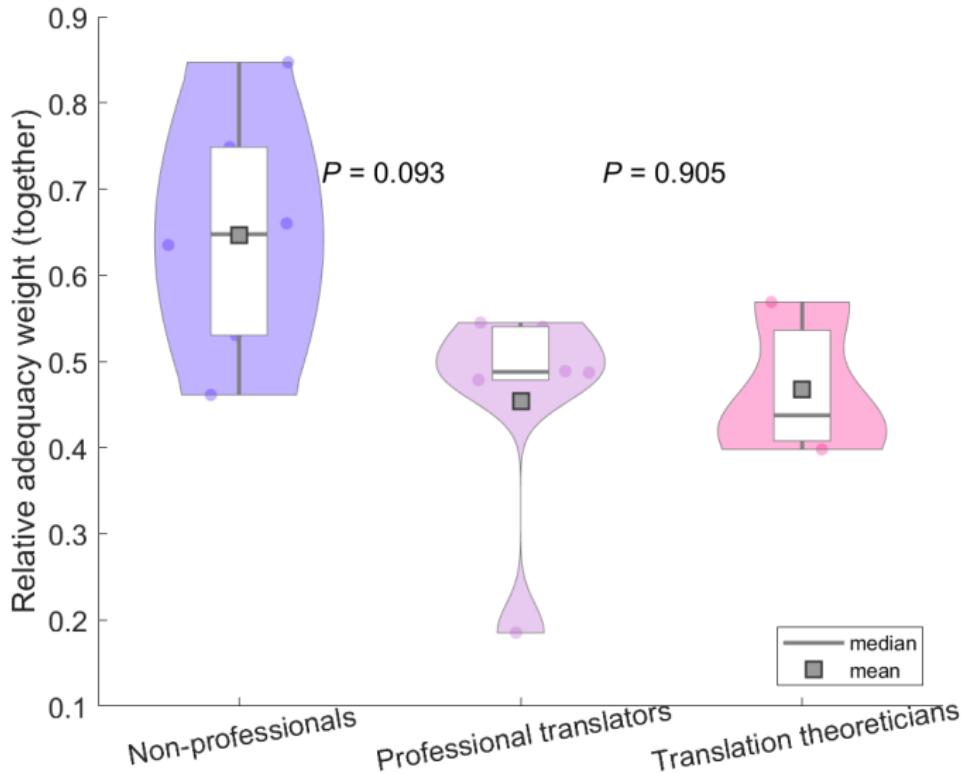
# It depends on the input text domain

7

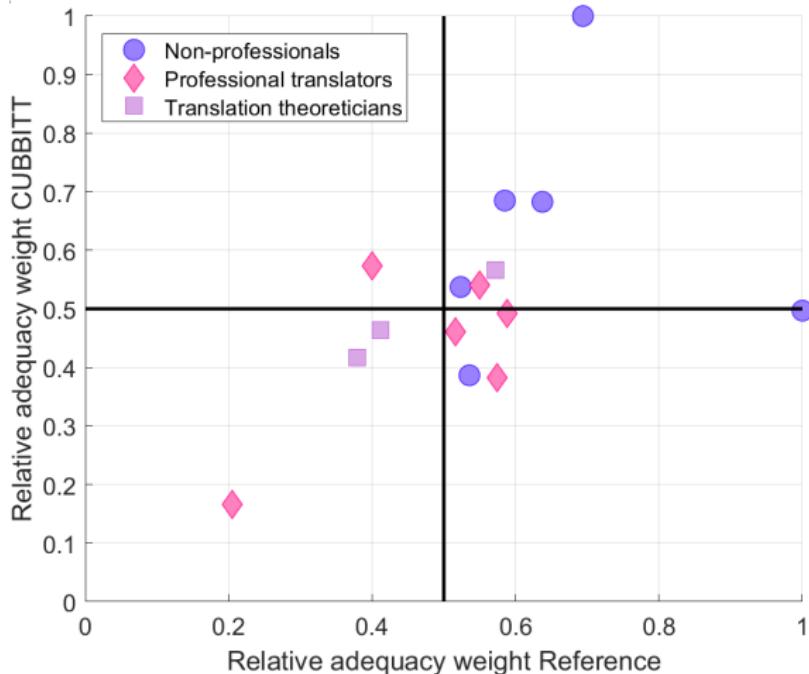
**D**

Can we predict the **overall quality**  
as a weighted average of **adequacy** and **fluency**?

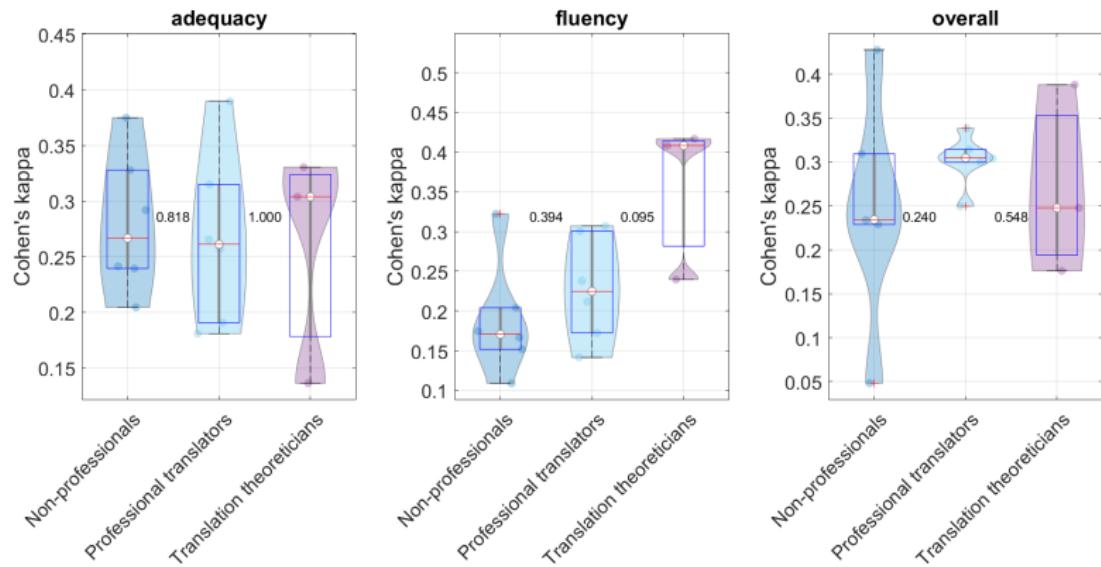
Translators put more emphasis on fluency,  
non-professionals on adequacy.



Some annotators put all the emphasis on adequacy,  
but only for one of the systems.

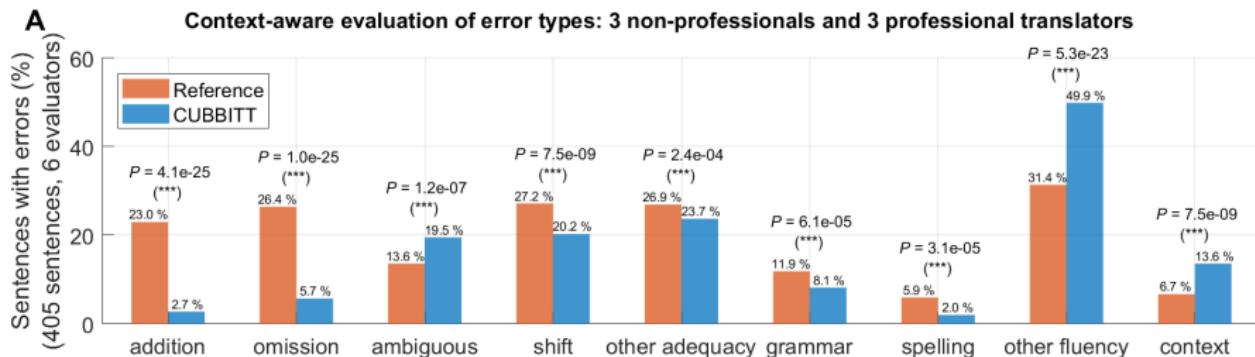


We cannot conclude that professionals are more reliable.



CUBBITT makes more errors than humans in

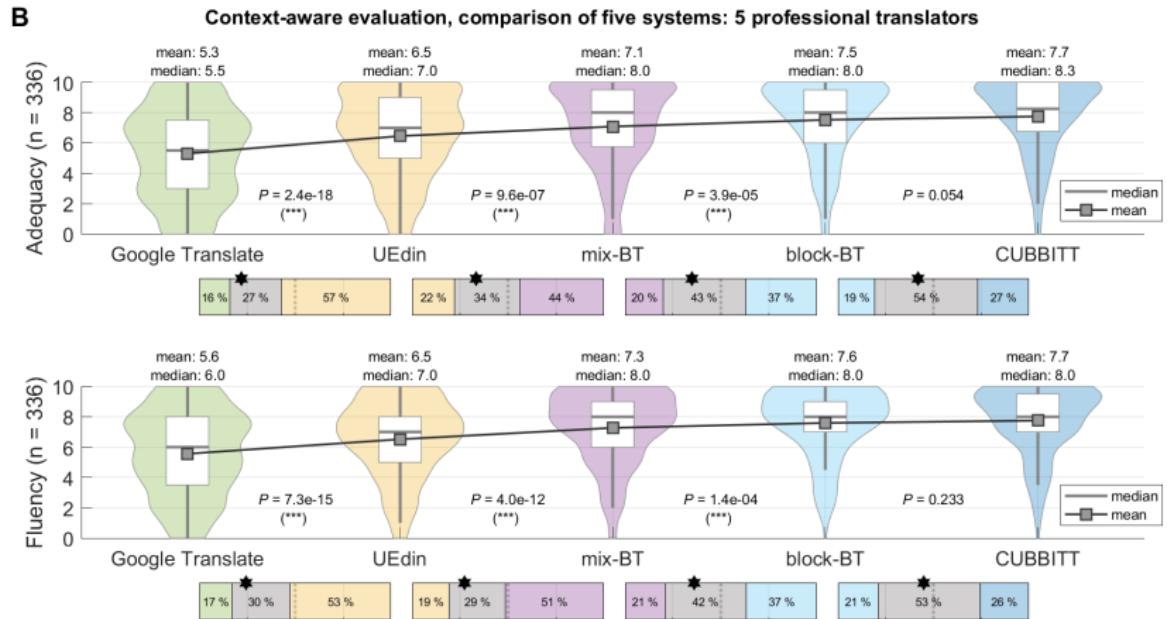
- translation of ambiguous words
- fluency (but not spelling and grammar) and
- cross-sentence context



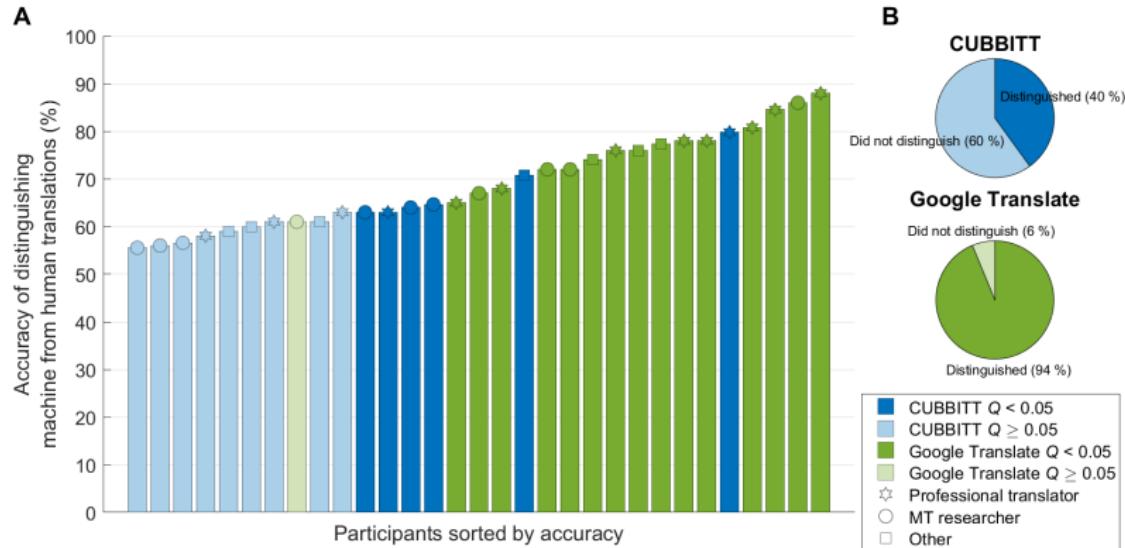
# Ablation analysis

13

BlockBT improves the quality (+0.4 adequacy, +0.3 fluency).



60 % of participants did not distinguish CUBBITT from human translations (on 100 isolated sentences)

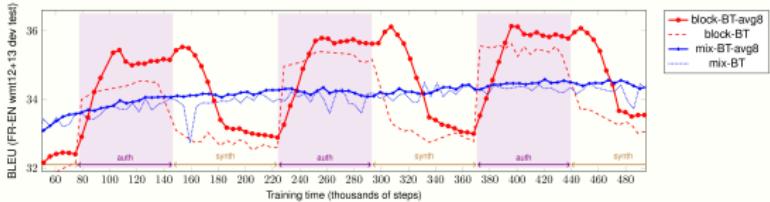


# BLEU Training curves for en-fr and en-pl

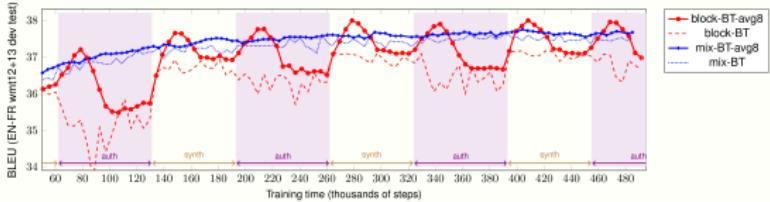
15

**A**

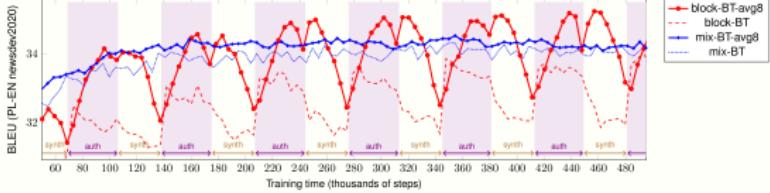
French → English

**B**

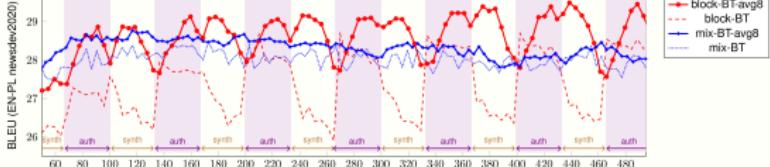
English → French

**C**

Polish → English

**D**

English → Polish



# Why Block-BT works?

16

**A**

Source sentence: "He was an original **guy** and lived life to the full" said **Gray** in a statement.



Translation of block-BT-Avg: "Byl to originální **chlap** a žil život naplno," uvedl **Gray** v prohlášení.

**Checkpoint 1 (SYNTH):** "Byl to originální **chlap** a žil život naplno" uvedl **Šedivý** v prohlášení.

**Checkpoint 2 (SYNTH):** "Byl to originální **chlap** a žil život naplno" uvedl **Šedivý** v prohlášení.

**Checkpoint 3 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 4 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 5 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 6 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 7 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 8 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**B**