Charles Translator for Ukraine

Martin Popel
popel@ufal.mff.cuni.cz

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Charles Translator for Ukraine: Start

- March 1, 2022: *Hi folks,*
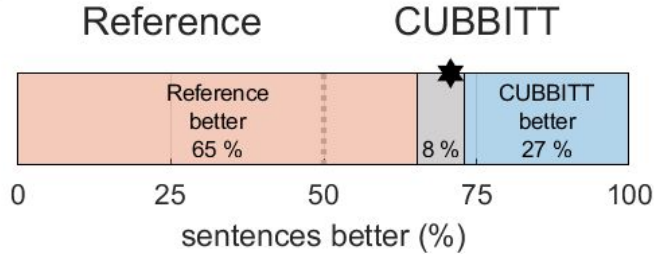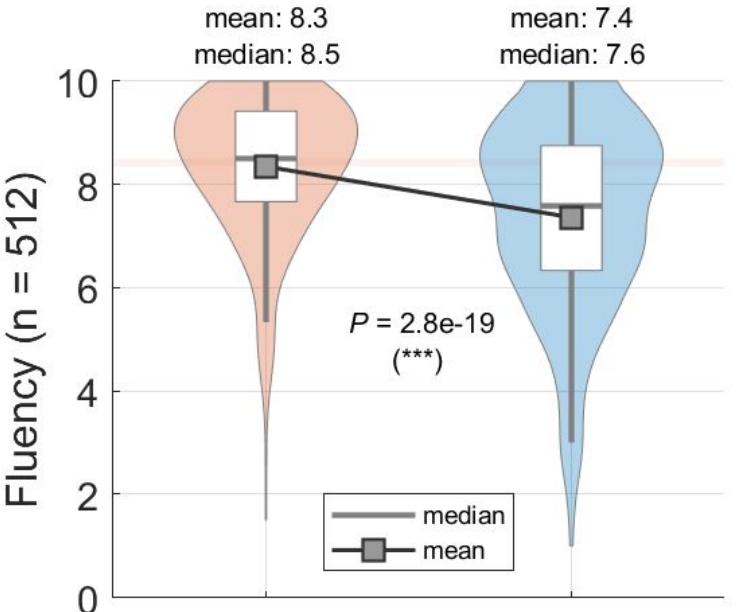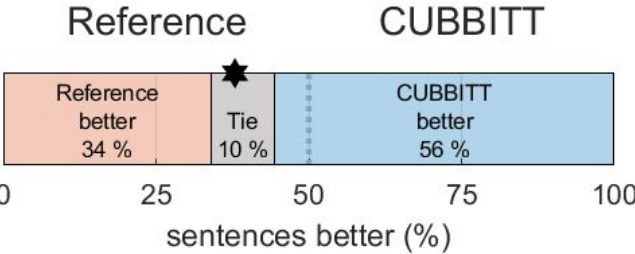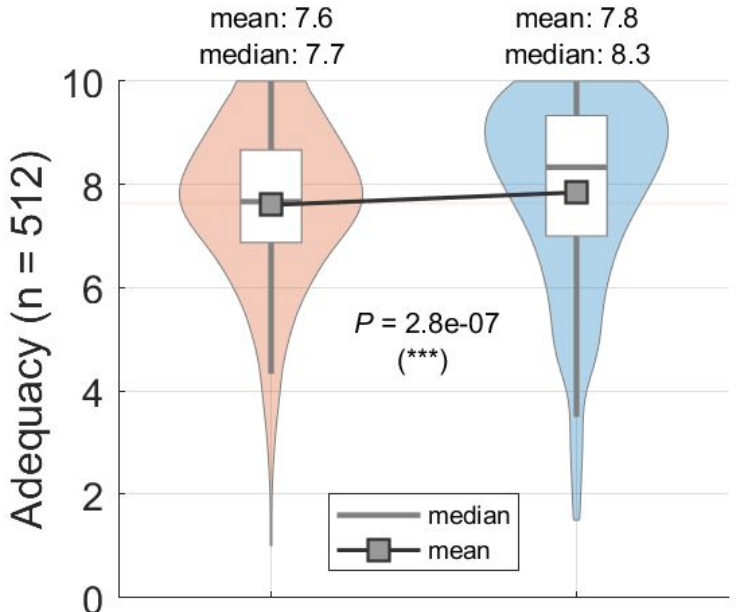 *Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?*

  Many people use Google Translate that
  pivots through English and
  a direct translator would be better.

  Best,
  Jindřich Libovický

Context-aware evaluation: 15 evaluators

Adequacy (n = 512)
- Reference — mean: 7.6, median: 7.7
- CUBBITT — mean: 7.8, median: 8.3
- P = 2.8e-07 (***)
- median
- mean
- sentences better (%): Reference better 34 %, Tie 10 %, CUBBITT better 56 %

Fluency (n = 512)
- Reference — mean: 8.3, median: 8.5
- CUBBITT — mean: 7.4, median: 7.6
- P = 2.8e-19 (***)
- median
- mean
- sentences better (%): Reference better 65 %, 8 %, CUBBITT better 27 %

3

Experience with high-quality English-Czech NMT

Context-aware evaluation: 15 evaluators

mean: 7.6
median: 7.7

mean: 7.8
median: 8.3

mean: 8.3
median: 8.5

mean: 7.4
median: 7.6

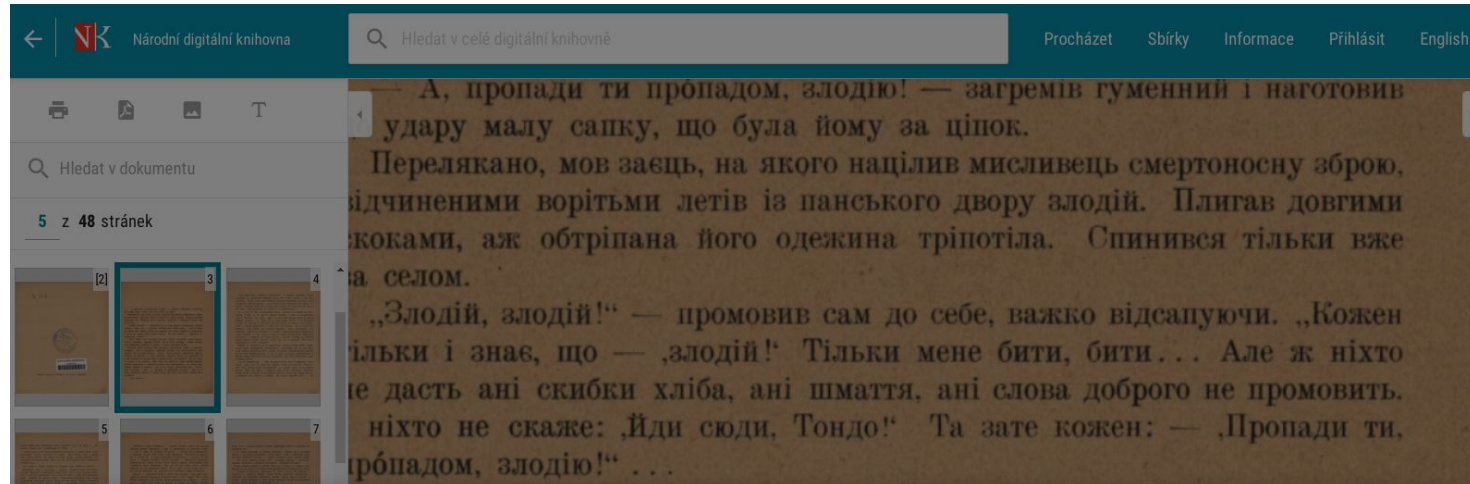Adequacy (n = 512)

10

8

6

4

2

0

Refe

Referenc
better
34 %

Charles Translator for Ukraine: Timeline

- March 1: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
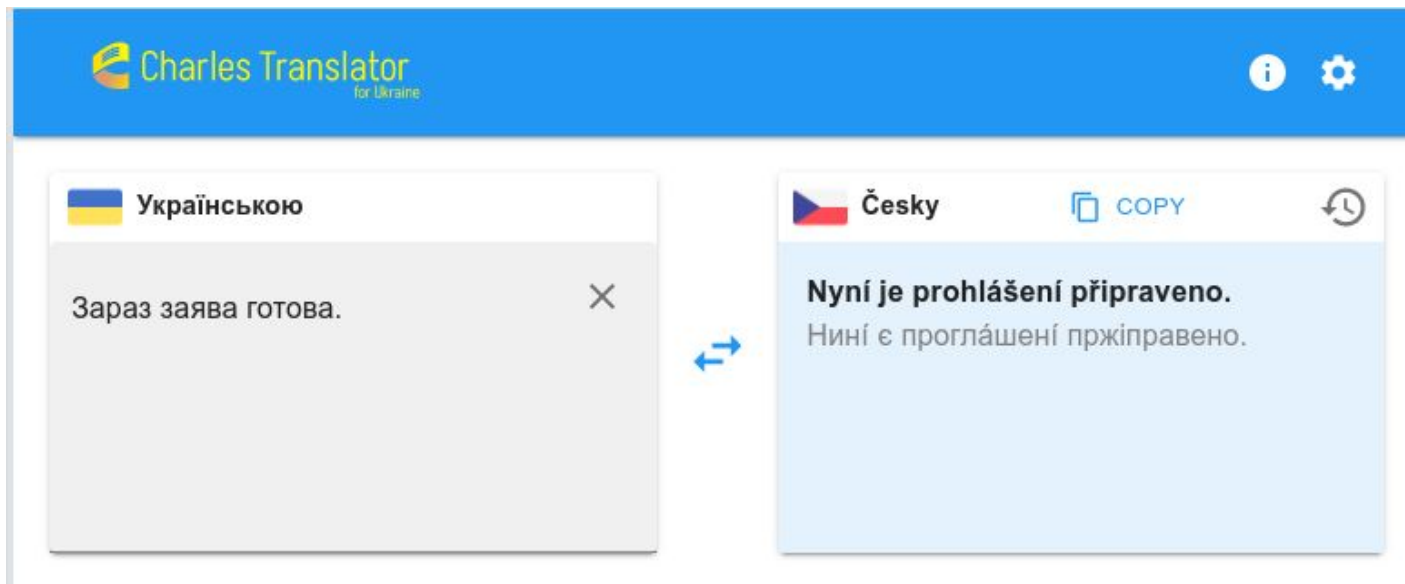- March 12: hackathon to develop a new frontend

Main work: collect and clean training data

ndk.cz has thousands of Ukrainian books, but no mapping to Czech translations and Russian OCR software:

УКРАЇНА ⇒ УКРАЇНА or УКРАША or УКРА1НА or УКРА!НA

Charles Translator for Ukraine: Timeline

- March 1: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- March 12: hackathon to develop a new frontend
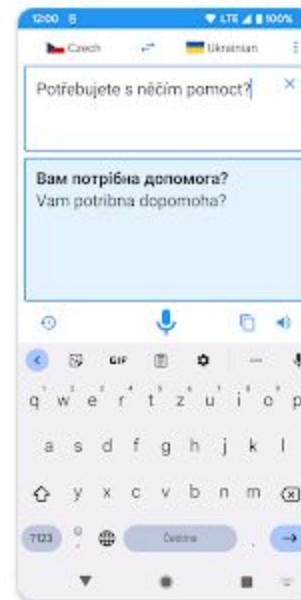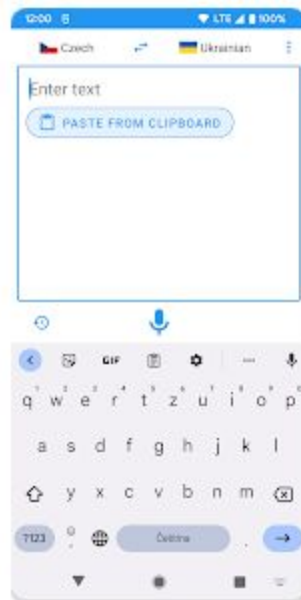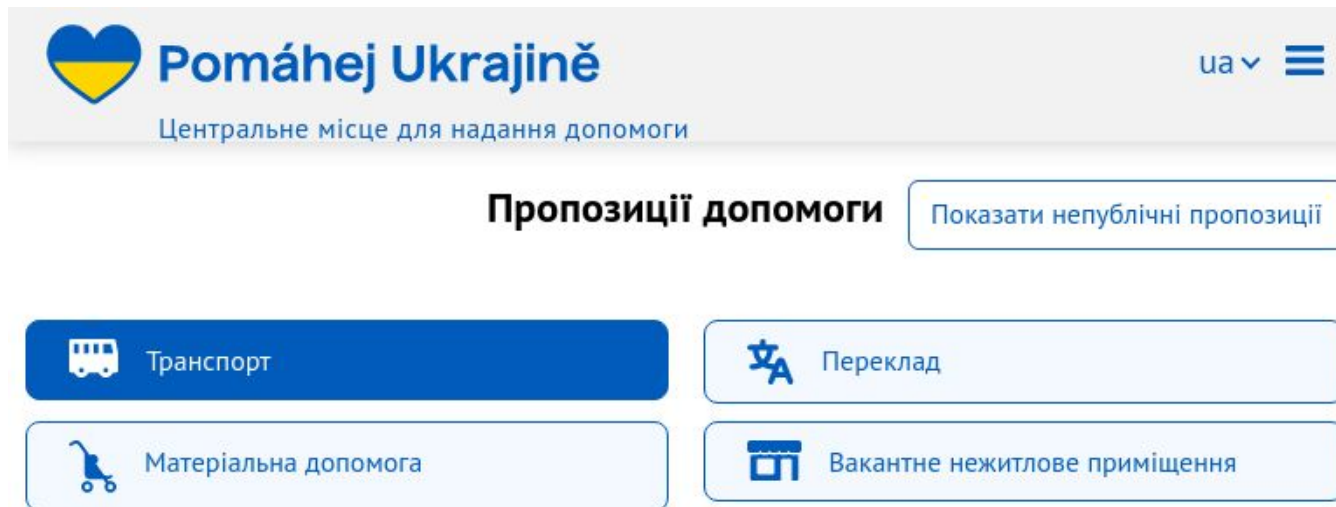- March 14: first cs ⇆ uk translator published at LINDAT/CLARIN

Charles Translator for Ukraine: Timeline

- March 1: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- March 12: hackathon to develop a new frontend
- March 14: first cs ⇆ uk translator published
- April 11: improved cs ⇆ uk models
- May 24: Android app

Why to develop our own MT?

- Quick solution until Big Tech companies improve their Ukrainian models

- Direct translation (not pivoted through English)

- Collect "in-domain" test set

- Focus on the needs of refugees and humanitarian organizations

- Offer free API (and on-premise for sensitive data)

Why direct translation?

Charles Translator
for Ukraine

🇺🇦 Українською
Зараз заява готова.

🇨🇿 Česky   COPY
Nyní je prohlášení připraveno.
Нині є прогла́шені пржіправено.

Why direct translation?

заява ⇒ application ⇒ aplikace
declaration   *computer app*

≡   **Google** Překladač   ⸬   Přihlásit se

ROZPOZNAT JAZYK   UKRAJINŠTINA   ⌄   ⇄   ČEŠTINA   ANGLIČTINA   NĚMČINA   ⌄

Зараз заява готова.   ✕   Nyní je aplikace připravena.   ☆

Zaraz zayava hotova.

🎤   🔊   19 / 5 000   Ук ▾   🔊   🗐   👍👎   ⟨

Why direct translation?

I am ill [fem]. And you [formal]?
I am ill [masc]. And you [informal]?
What kind of medicine is that?

Why direct translation?

Why direct translation?

Why direct translation?

Why direct translation?

Thousands of translations per week (April-September)

Translation direction
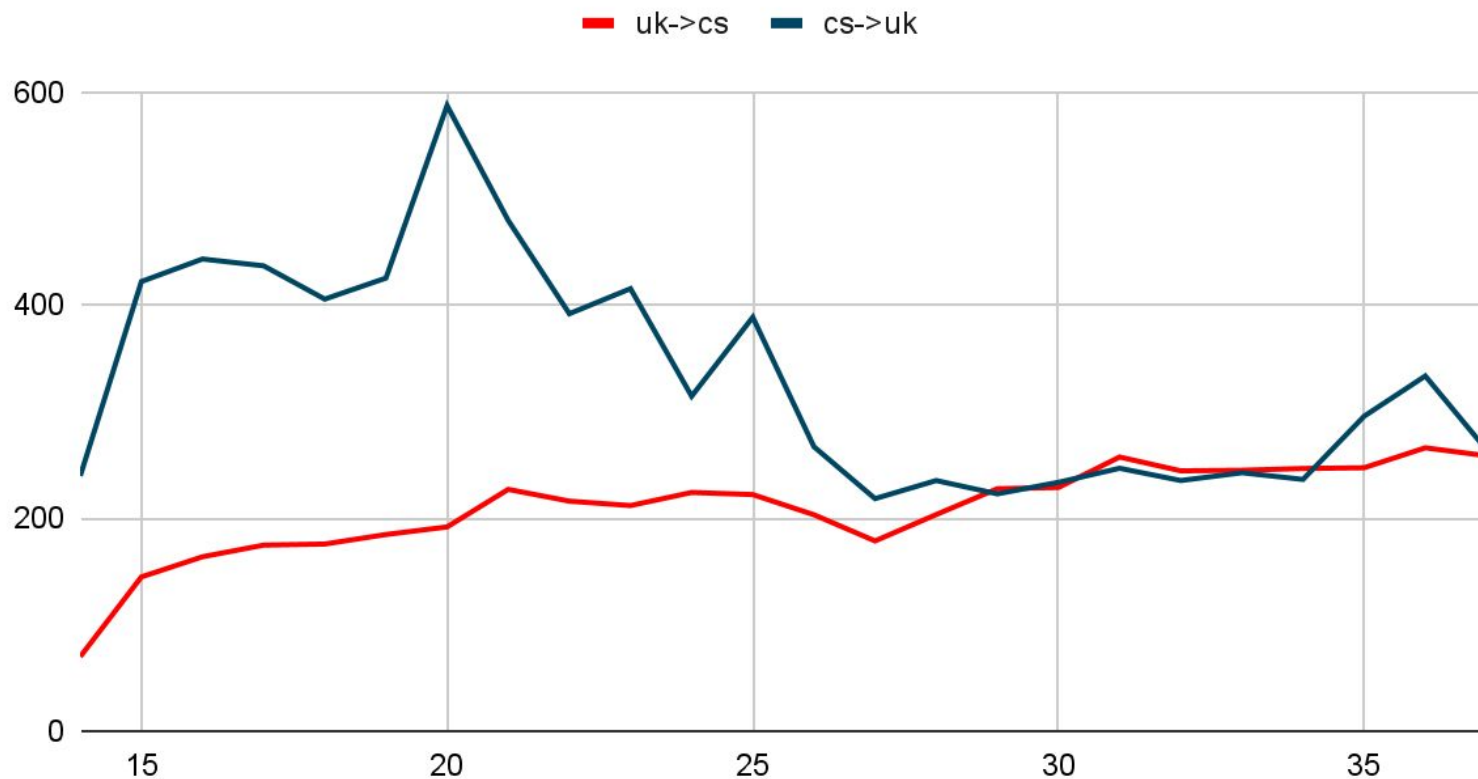
uk->cs   cs->uk

17

Charles Translator for Ukraine: Timeline

- March 1: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- March 12: hackathon to develop a new frontend
- March 14: first cs ⇆ uk translator published
- April 11: improved cs ⇆ uk models
- May 24: Android app
- since September: evaluating 11 systems in WMT22 shared task

WMT 2022 General MT shared task evaluation

WMT22 Ukrainian→Czech		
SQM	chrF	System
0.393	—	HUMAN-A
0.221	59.6	HuaweiTSC
0.181	60.7	AMU
0.137	60.4	Lan-Bridge
0.102	59.0	**Charles**
0.095	60.3	Online-B
. . .		
−0.036	57.5	Online-A
−0.157	56.3	Online-G
−0.276	55.3	Online-Y
—	53.7	eTranslation
−0.626	50.7	ALMAnaCH

WMT22 Czech→Ukrainian		
SQM	chrF	System
0.295	—	HUMAN-A
0.225	64.0	Online-B
0.151	61.5	AMU
0.125	64.0	Lan-Bridge
0.065	62.6	HuaweiTSC
0.062	61.5	**Charles**
. . .		
−0.008	60.8	Online-G
−0.075	62.2	Online-A
−0.257	59.6	Online-Y
—	57.5	eTranslation
−0.669	54.5	ALMAnaCH

- Try Czech-English   `https://cubbitt.lindat.cz`
- Try Czech-Ukrainian `https://translator.cuni.cz`