Charles Translator:

A Machine Translation System between Ukrainian and Czech

Martin Popel, Lucie Poláková, Michal Novák, Jindřich Helcl, Jindřich Libovický, Pavel Straňák, Tomáš Krabač, Jaroslava Hlaváčová, Mariia Anisimova, Tereza Chlaňová

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

Introduction

Story of Charles Translator in timeline

- Feb 24, 2022: start of Russia's full-scale invasion of Ukraine
- Mar 1: *Hi folks, shouldn't we build a Czech-Ukrainian translator?*

Story of Charles Translator in timeline

- Feb 24, 2022: start of the Russian full-scale invasion of Ukraine
- Mar 1: *Hi folks, shouldn't we build a Czech-Ukrainian translator?*
- Mar 12: hackathon to develop a new web app
- Mar 14: first cs ⇆ uk translator published

Story of Charles Translator in timeline

- Feb 24, 2022: start of the Russian full-scale invasion of Ukraine
- Mar 1: *Hi folks, shouldn't we build a Czech-Ukrainian translator?*
- Mar 12: hackathon to develop a new web app
- Mar 14: first cs ⇆ uk translator published
- Apr 11: improved cs ⇆ uk models
- May 24: Android app

Story of Charles Translator in timeline

- Feb 24, 2022: start of the Russian full-scale invasion of Ukraine
- Mar 1: *Hi folks, shouldn't we build a Czech-Ukrainian translator?*
- Mar 12: hackathon to develop a new web app
- Mar 14: first cs ⇆ uk translator published
- Apr 11: improved cs ⇆ uk models
- May 24: Android app
- since Jun 2022:
  - support of audio inputs/outputs
  - improved frontends
  - WMT test sets

Why did we need a new translator?

- in the first days of invasion, more than half a million Ukrainian refugees granted temporary protection in Czechia
  - need for a fast and effective means of communication

Why did we need a new translator?

- in the first days of invasion, more than **half a million Ukrainian refugees** granted temporary protection in Czechia
  - need for a fast and effective means of communication

March 2022: over 200 000 Ukrainian refugees in Czechia

Photo: René Volfík, irozhlas.cz

Why did we need a new translator?

- in the first days of invasion, more than **half a million Ukrainian refugees** granted temporary protection in Czechia
    - need for a fast and effective means of communication

Why did we need a new translator?

- in the first days of invasion, more than half a million Ukrainian refugees granted temporary protection in Czechia
  - need for a fast and effective means of communication
- publicly available translation systems at that time were not sufficient
  - no direct translation between uk and cs: pivoting through English
  - no specific focus on the needs of refugees and humanitarian organizations
  - no free API

Why did we need a new translator?

- in the first days of invasion, more than half a million Ukrainian refugees granted temporary protection in Czechia
  - need for a fast and effective means of communication
- publicly available translation systems at that time were not sufficient
  - no **direct translation** between uk and cs: pivoting through English
  - no specific focus on the needs of refugees and humanitarian organizations
  - no free API

Why direct translation?

Photo: René Volfík, irozhlas.cz

Why direct translation?

заява ⇒ application ⇒ aplikace
declaration   *computer app*

Why direct translation?

Charles Translator
for Ukraine

Українською

Зараз заява готова.

Česky   COPY

Nyní je prohlášení připraveno.
Нині є проглашені пржіправено.

THE LINDAT/CLARIAH-CZ PROJECT (LM2018101; formerly LM2010013, LM2015071) IS FULLY SUPPORTED BY THE MINISTRY OF EDUCATION, SPORTS AND YOUTH OF THE CZECH REPUBLIC UNDER THE PROGRAMME LM OF LARGE INFRASTRUCTURES

Why direct translation?

Google Překladač

⚙ Přihlásit se

ROZPOZNAT JAZYK | ČEŠTINA | ANG ⌄ | ⇄ | UKRAJINŠTINA | ČEŠTINA | ANGLIČ ⌄

Jsem nemocná. A vy?
Jsem nemocný. A ty?
Jaké léky to jsou?
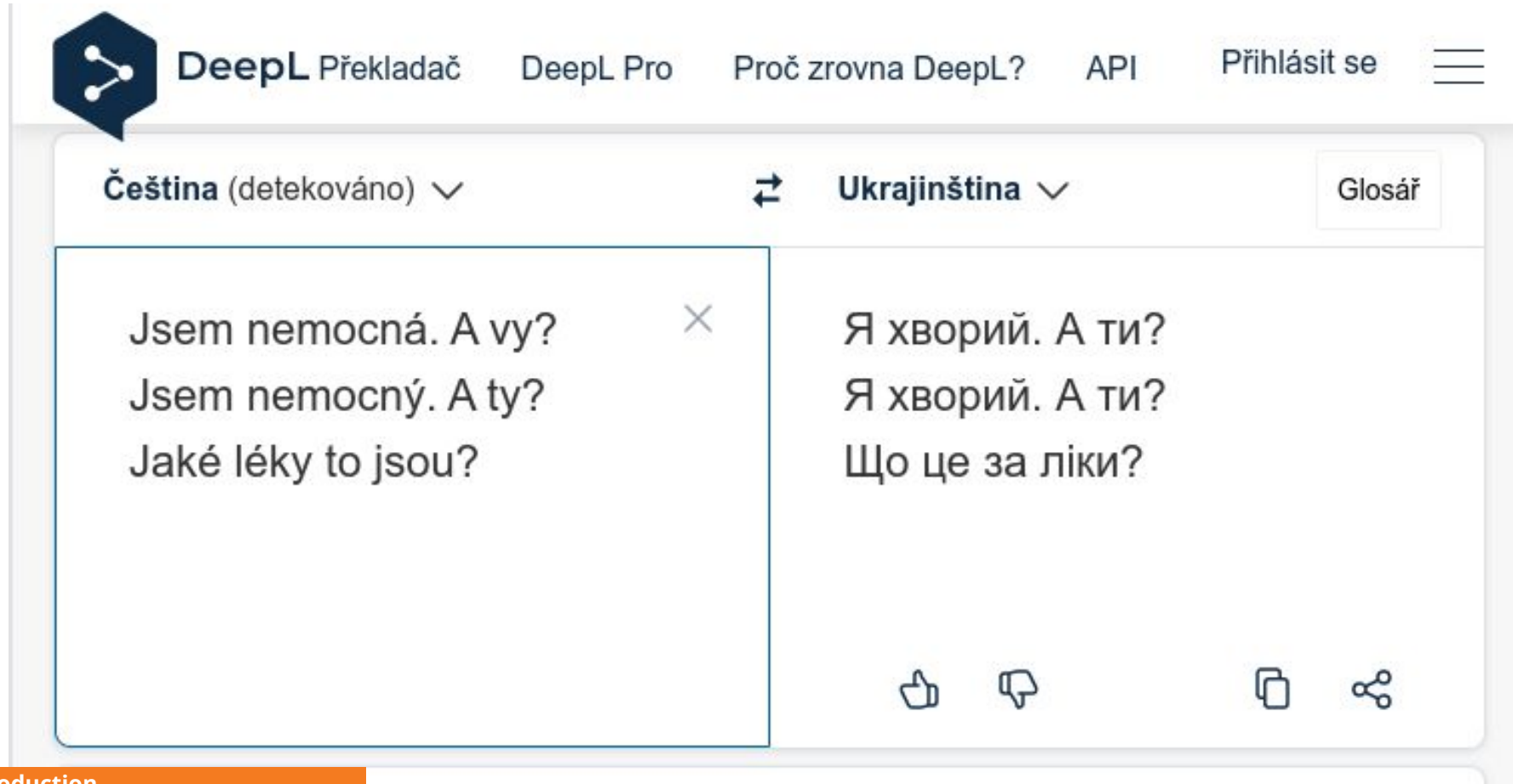
Я хворий. I ти?
Я хворію. I ти?
Які це наркотики?

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

YA khvoryy. I ty?
YA khvoriyu. I ty?
Yaki tse narkotyky?

58 / 5 000

Odeslat zpětnou vazbu

Why direct translation?

Why direct translation?

Digital Europe
eTranslation

About | Help | Liability

Translate text | Translate documents | My translation requests | My settings | English

Jsem nemocná. A vy?
Jsem nemocný. A ty?
Jaké léky to jsou?

Я хворий. А як щодо тебе?
Я хворий. А ти?
Що це за ліки?

58 / 2500

From Czech ⟷ To Ukrair

Why direct translation?

Why direct translation?

Components of Charles Translator

Components of Charles Translator

Interfaces

```
                           ┌──────────┐
                        ┌─▶│ REST API │──┐
                        │  └──────────┘  │
┌──────────────┐        │  ┌──────────┐  │      👤
│ Translation  │───────────▶│ Web app │────────▶
│   service    │        │  └──────────┘  │
└──────────────┘        │  ┌──────────┐  │
                        └─▶│ Android  │──┘
                           │   app    │
                           └──────────┘
```

Translation service

- method
    - Transformer architecture with iterated block backtranslation

Translation service

- method
  - Transformer architecture with **iterated block backtranslation**

Experience with high-quality Czech-English NMT

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

Translation service

- method
  - Transformer architecture with iterated block backtranslation

Translation service

- method
  - Transformer architecture with iterated block backtranslation
- training data
  - collected in a short period of time with the help of volunteer subjects
  - parallel data:
    - translators, translation agencies
    - the InterCorp project of Czech National Corpus
    - the OPUS repository
  - monolingual data:
    - cs: CzEng 2.0
    - uk: WMT News Crawl, Leipzig Corpora, etc.
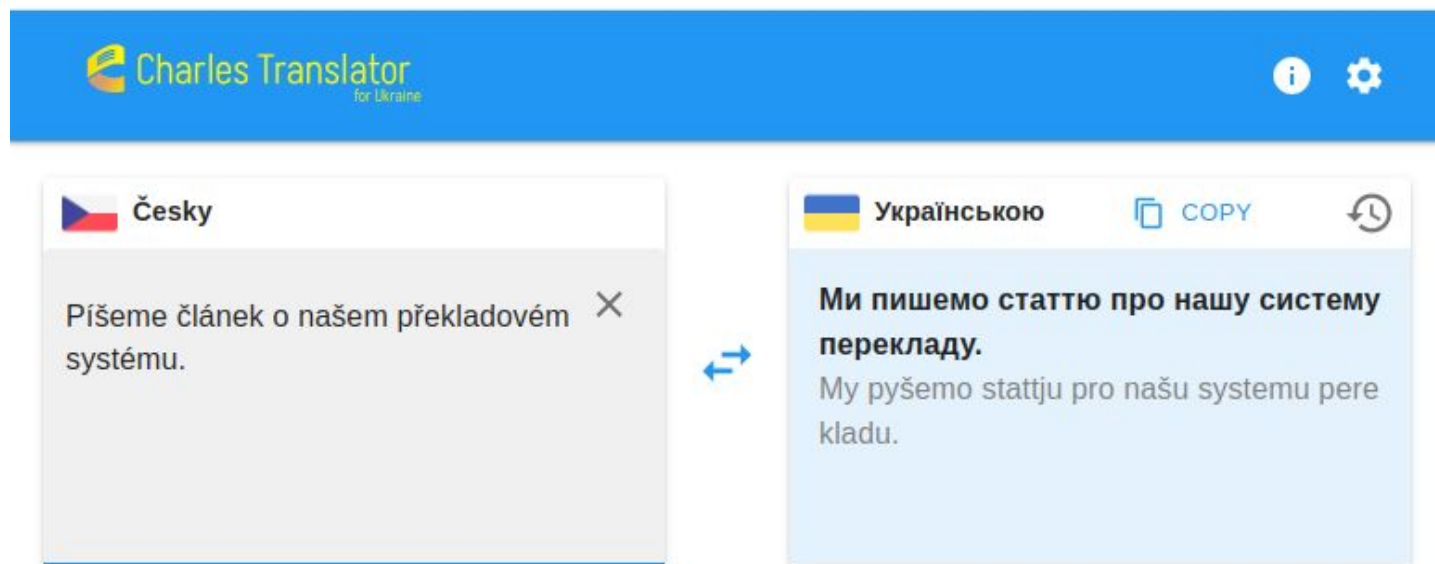
Translation service

- method
  - Transformer architecture with iterated block backtranslation
- training data
  - collected in a short period of time with the help of volunteer subjects
  - parallel data:
    - translators, translation agencies
    - the InterCorp project of Czech National Corpus
    - the OPUS repository
  - monolingual data:
    - cs: CzEng 2.0
    - uk: WMT News Crawl, Leipzig Corpora, etc.
- test data
  - pseudonymized and manually translated system logs from ChT + news crawls
  - annotated with user type and topic
  - published as WMT22 UK→CS and WMT23 CS→UK test sets

Translation service

- method
  - Transformer architecture with iterated block backtranslation
- training data
  - collected in a short period of time with the help of volunteer subjects
  - parallel data:
    - translators, translation agencies
    - the InterCorp project of Czech National Corpus
    - the OPUS repository
  - monolingual data:
    - cs: CzEng 2.0
    - uk: WMT News Crawl, Leipzig Corpora, etc.
- test data
  - pseudonymized and manually translated system logs from ChT + news crawls
  - annotated with user type and topic
  - published as WMT22 UK→CS and WMT23 CS→UK test sets
- engine
  - LINDAT/CLARIN infrastructure

Interfaces

- REST API
  - for non-commercial use available upon request

Interfaces

- REST API
  - for non-commercial use available upon request
- web app
  - transliteration/transcription
  - extended to other language pairs
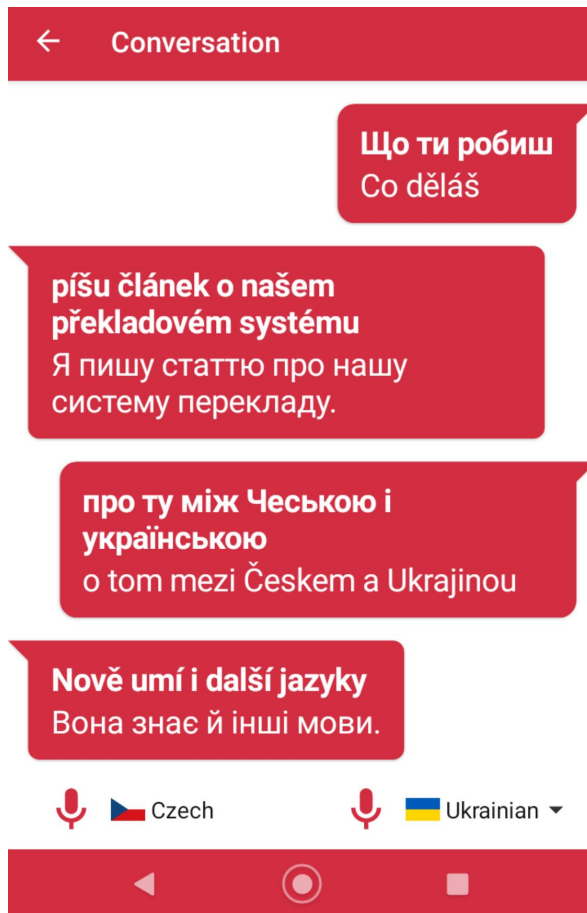 - React/Node.js
  - https://translator.cuni.cz/

Web app

Interfaces

- REST API
  - for non-commercial use available upon request
- web app
  - transliteration/transcription
  - extended to other language pairs
  - React/Node.js
  - https://translator.cuni.cz/
- Android app
  - web app features
  - audio input/output
  - conversation mode
  - Kotlin using Jetpack Compose
  - installed by around 2000 people
  -

Android app

Quality of translations

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

WMT 2022 General MT shared task evaluation

Czech→Ukrainian

Range	Ave.	Ave. z	System
1	85.6	0.295	HUMAN-A
2-5	84.6	0.225	Online-B
2-3	84.1	0.151	AMU
3-6	82.5	0.125	Lan-Bridge
3-6	81.1	0.065	HuaweiTSC
4-8	81.9	0.062	CharlesTranslator
6-8	80.2	0.026	CUNI-JL-JH
6-8	80.2	−0.002	CUNI-Transf.
9-10	79.8	−0.008	Online-G
9-10	79.2	−0.075	Online-A
11	76.0	−0.257	Online-Y
12	68.4	−0.669	ALMAnaCH-Inria

Ukrainian→Czech

Range	Ave.	Ave. z	System
1	89.6	0.417	HUMAN-A
2-3	85.6	0.182	AMU
2-4	83.5	0.148	HuaweiTSC
4-8	83.5	0.127	Lan-Bridge
3-8	82.0	0.110	CUNI-Transf.
4-8	82.5	0.082	CharlesTranslator
4-8	81.4	0.052	CUNI-JL-JH
4-8	81.9	0.042	Online-B
9-10	80.0	-0.101	Online-A
9-10	77.5	-0.138	Online-G
11	73.9	-0.351	Online-Y
12	69.2	-0.617	ALMAnaCH-Inria

WMT 2023 General MT shared task evaluation

Czech→Ukrainian

Rank	Ave.	System
1-3	83.7	ONLINE-B
1-3	83.6	GPT4-5shot
1-3	83.2	Human-refA
4-8	82.8	ONLINE-W
4-8	82.4	CUNI-GA
4-8	81.8	Charles Translator
4-8	81.3	GTCOM_DLUT
4-8	80.6	ONLINE-A
9-11	79.5	ONLINE-G
9-13	78.7	ONLINE-Y
9-13	78.7	MUNI-NLP
10-13	77.4	Lan-BridgeMT
10-13	76.9	NLLB_MBR_BLEU
14	76.7	NLLB_Greedy

Use cases

Use cases

- actively asking organizations that assisted Ukrainian refugees about their translation needs
 - shared our REST API with them

Use cases

- actively asking organizations that assisted Ukrainian refugees about their translation needs
  - shared our REST API with them

- Pomáhej Ukrajině web platform

Pomáhej Ukrajině uses Charles Translator
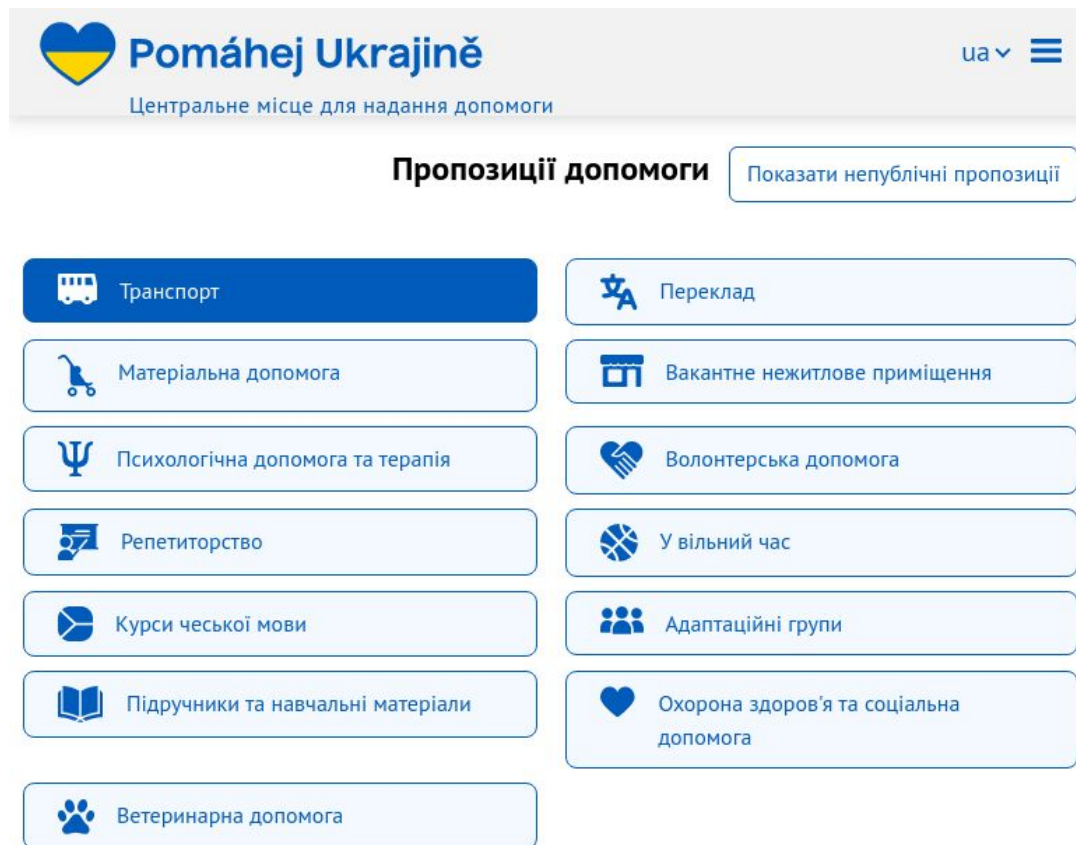

- connects public offers of assistance with the needs of refugees and aiding organizations
- translating segments of the offers

Use cases

- actively asking organizations that assisted Ukrainian refugees about their translation needs
  - shared our REST API with them

- Pomáhej Ukrajině web platform

Use cases

- actively asking organizations that assisted Ukrainian refugees about their translation needs
  - shared our REST API with them

- Pomáhej Ukrajině web platform

- Police of the Czech Republic

Czech Police uses Charles Translator

Czech Police uses Charles Translator

- prosecution of
  war crimes in Ukraine
- translating testimonies of witnesses
- Ukrainian ⇆ Czech, Russian ⇆ Czech
- on-premise installation (sensitive data)

Use cases

- actively asking organizations that assisted Ukrainian refugees about their translation needs
  - shared our REST API with them

- Pomáhej Ukrajině web platform

- Police of the Czech Republic

Service usage

- Apr-Sep 2022
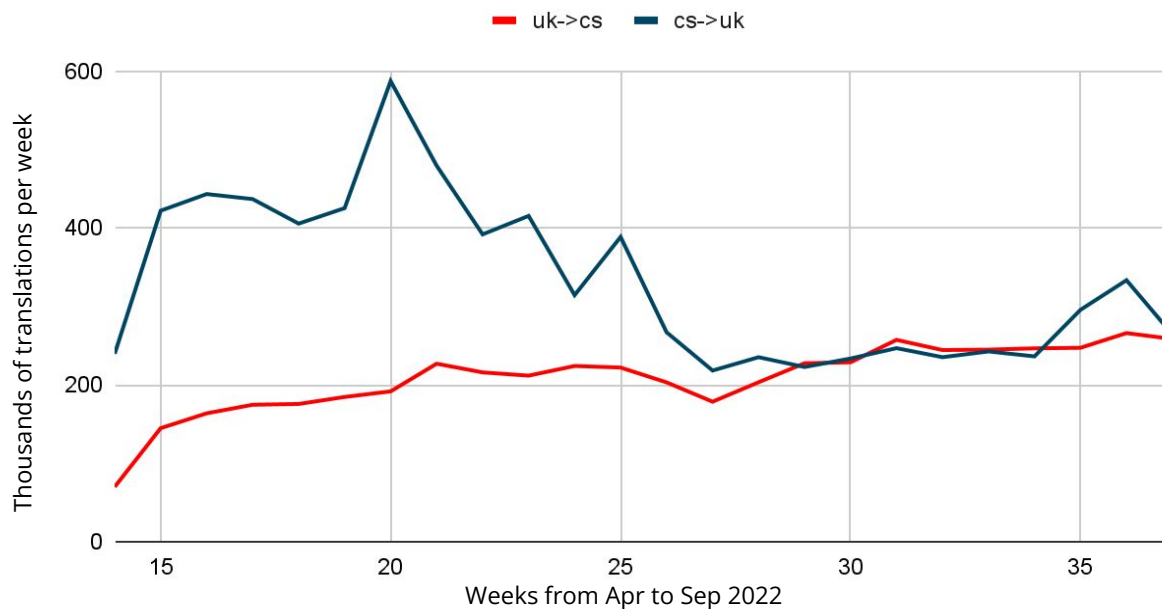
Service usage

- Apr-Sep 2022
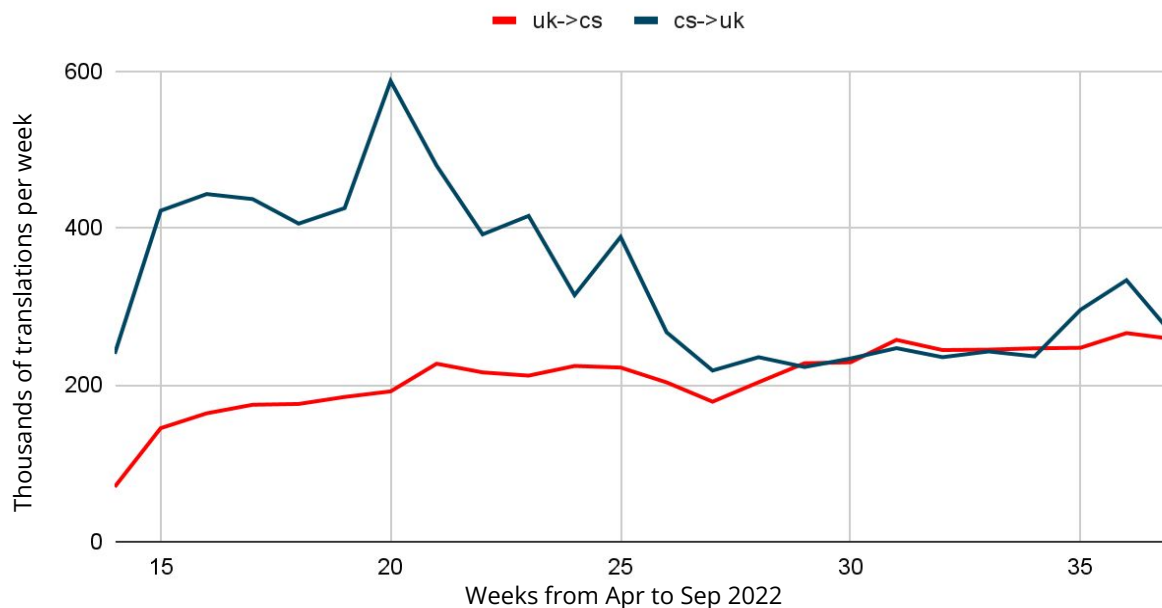
- Sep 2023
  - UK→CS: 30k translation requests per day
  - CS→UK: 12k translation requests per day

Try it: **https://translator.cuni.cz**

Thank you for your attention!

Acknowledgements

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

Main work: collect and clean training data
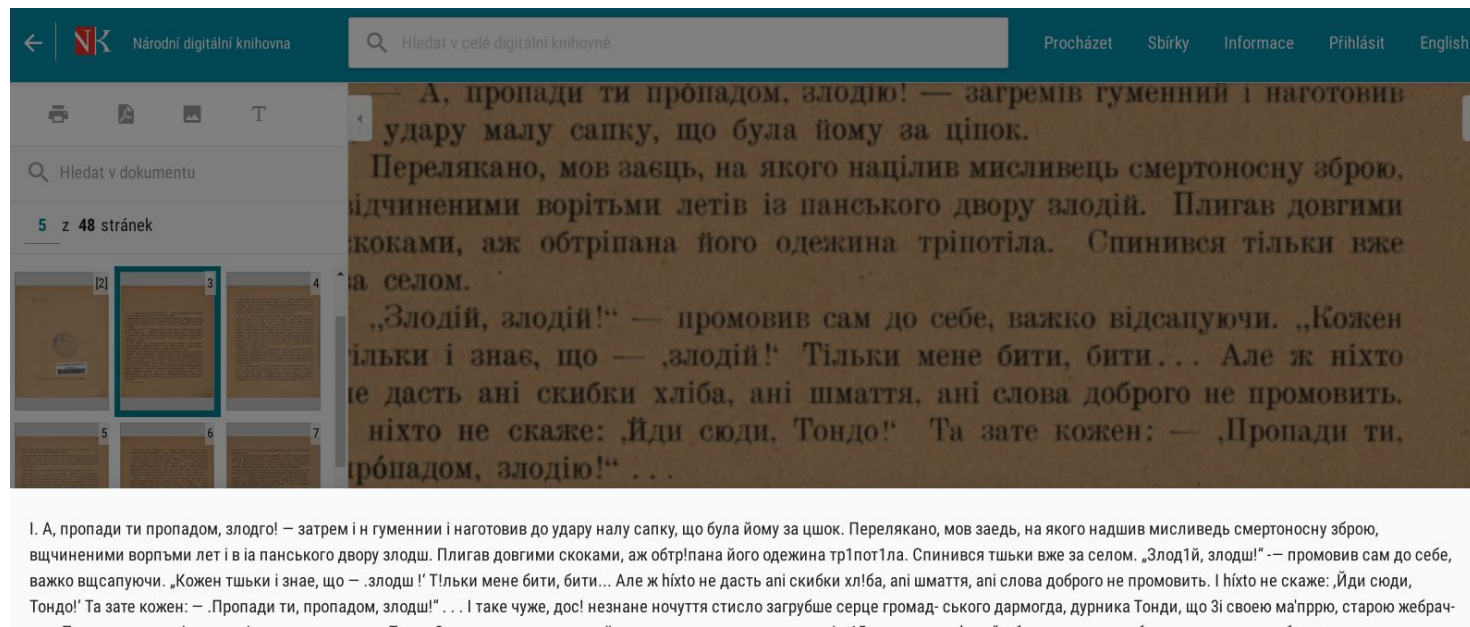
ndk.cz has thousands of Ukrainian books, but no mapping to Czech translations and Russian OCR software:

УКРАЇНА ⇒ УКРАÌНА or УКРАША or УКРА1НА or УКРА!НА

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
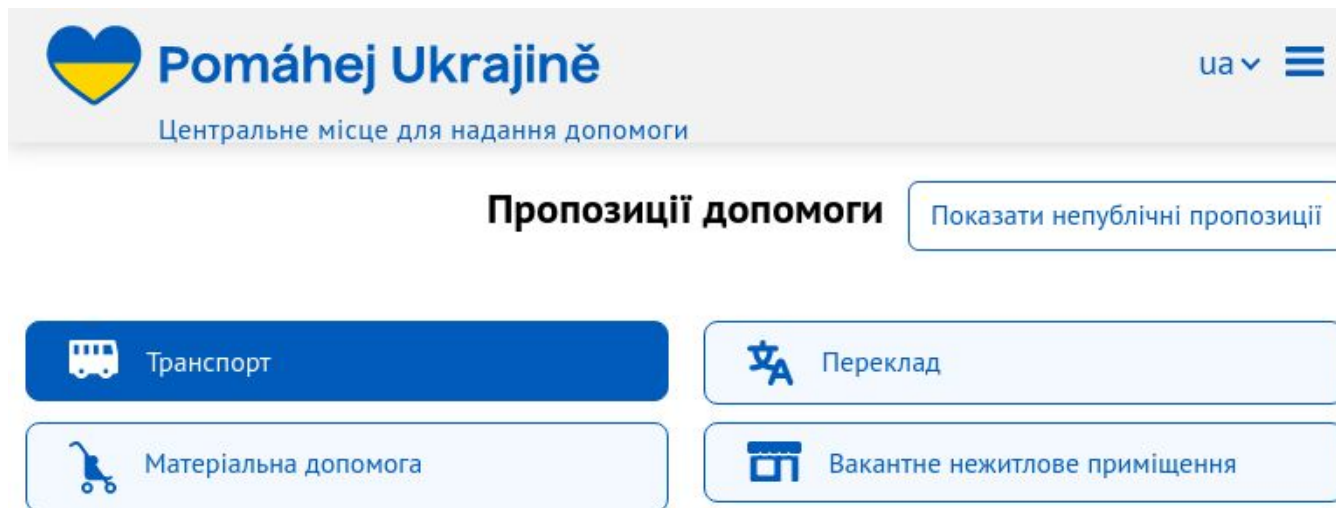- since September: evaluating 11 systems in WMT22 shared task

Charles Translator for Ukraine: Timeline

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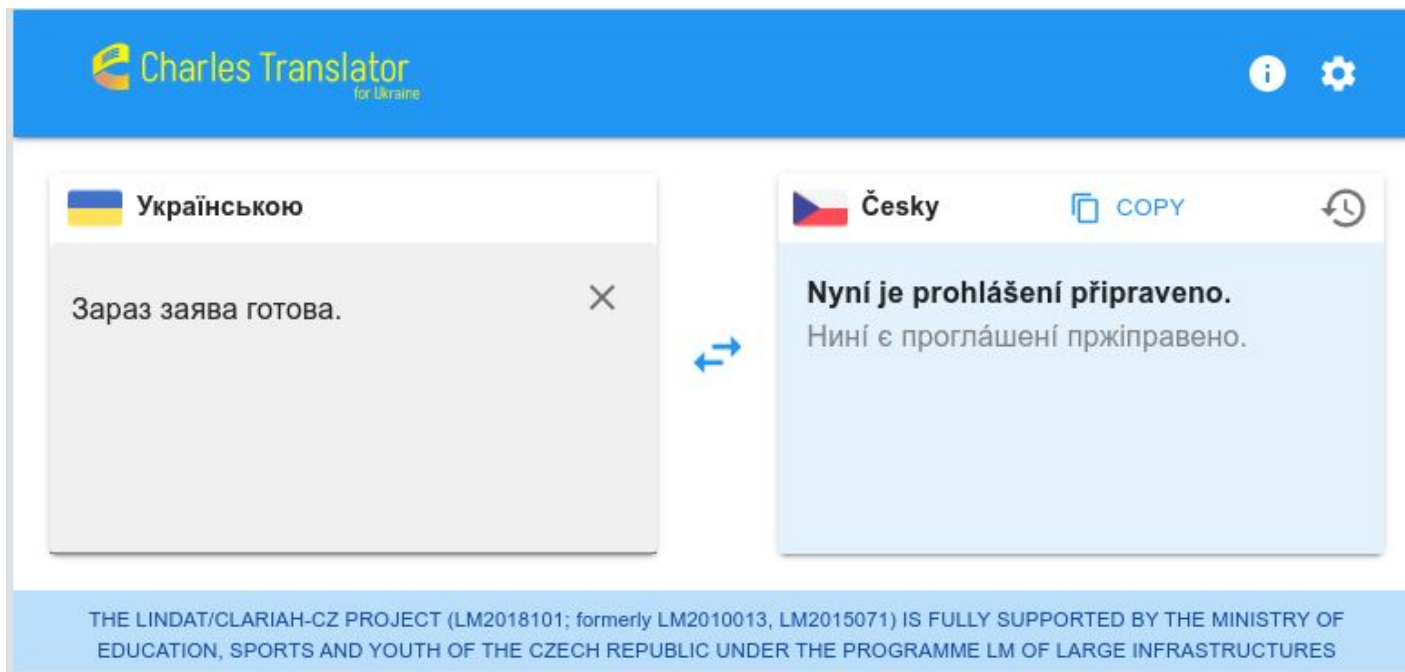- March 12: hackathon to develop a new frontend

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

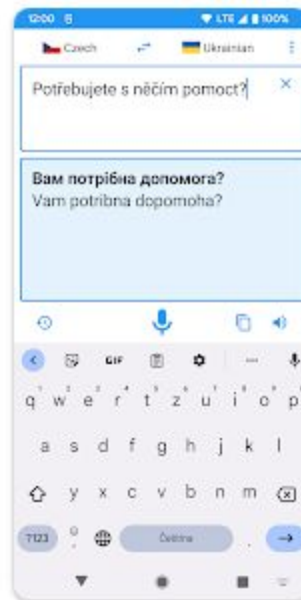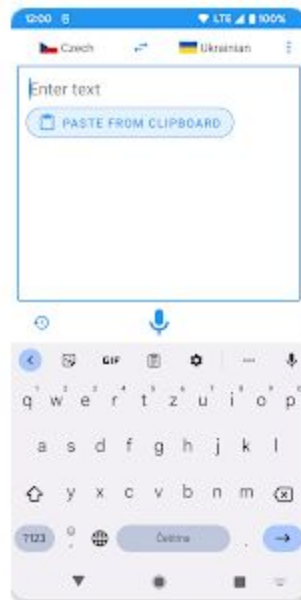- Offer free API (and on-premise for sensitive data)

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

Story of Charles Translator in timeline

- March 1: *Hi folks, shouldn't we build a Czech-Ukrainian translator?*
- 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

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