Presentation by local National Coordinator

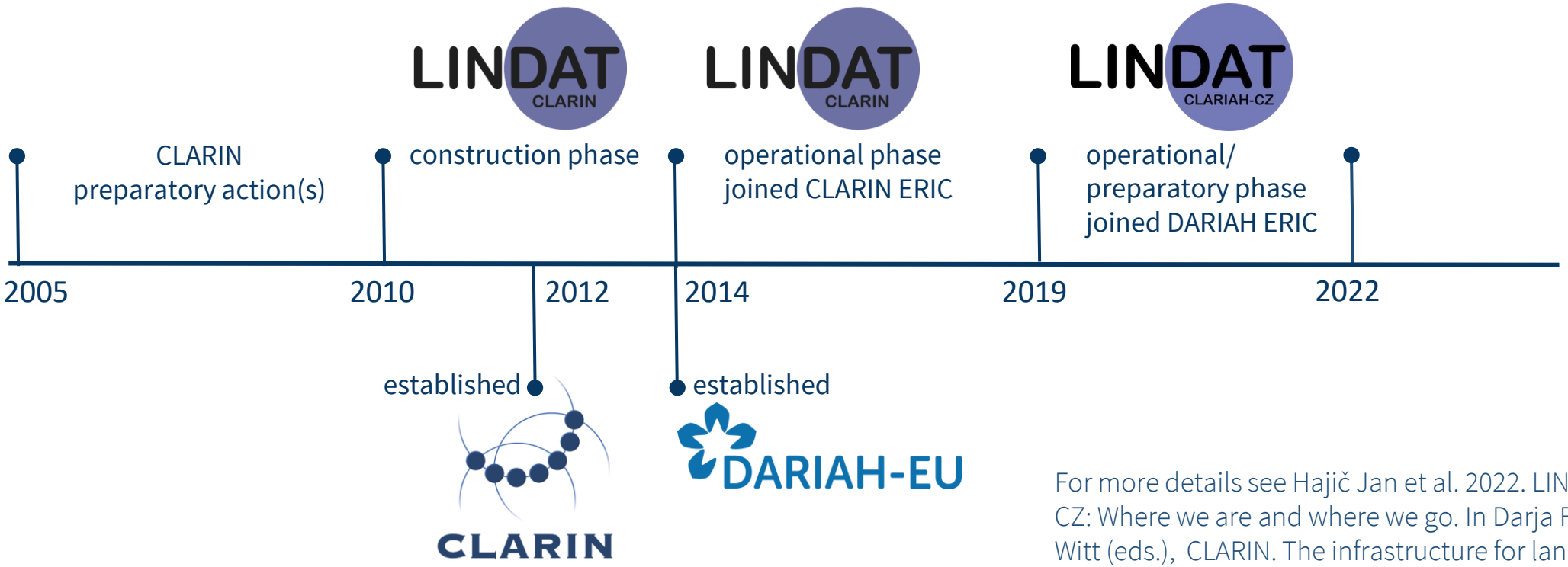
CLARIN Annual Conference
October 10-12, 2022
Prague, Czech Republic

Eva Hajičová, Barbora Hladká, Martin Popel

Welcome to Prague

About LINDAT/CLARIAH-CZ (http://lindat.cz)

Digital Research Infrastructure for the Language Technologies, Arts and Humanities

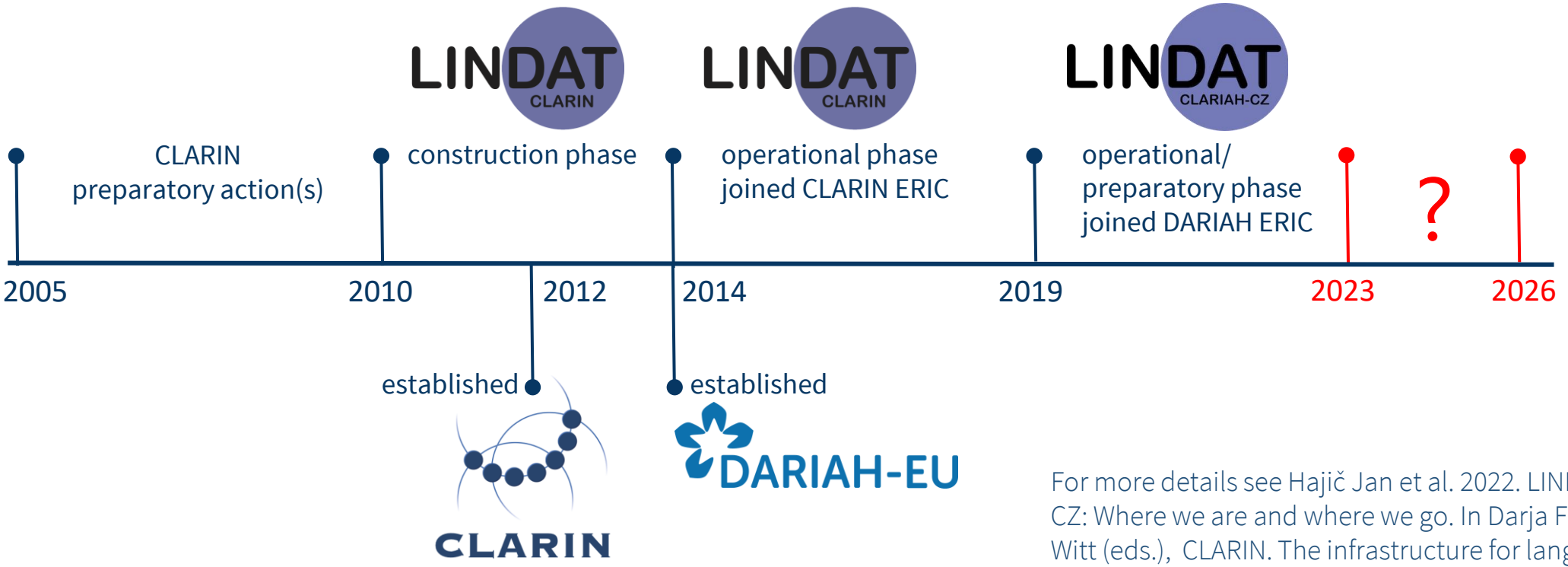


For more details see Hajič Jan et al. 2022. LINDAT/CLARIAH-CZ: Where we are and where we go. In Darja Fišer & Andreas Witt (eds.), CLARIN. The infrastructure for language resources. Berlin: deGruyter.

The LINDAT/CLARIAH-CZ PROJECT (LM2018101) is fully supported by the Ministry of Education, Sports and Youth of the Czech Republic under the programme LM of "Large infrastructures".

About LINDAT/CLARIAH-CZ (http://lindat.cz)

Digital Research Infrastructure for the Language Technologies, Arts and Humanities



For more details see Hajič Jan et al. 2022. LINDAT/CLARIAH-CZ: Where we are and where we go. In Darja Fišer & Andreas Witt (eds.), CLARIN. The infrastructure for language resources. Berlin: deGruyter.

The LINDAT/CLARIAH-CZ PROJECT (LM2018101) is fully supported by the Ministry of Education, Sports and Youth of the Czech Republic under the programme LM of "Large infrastructures".

About LINDAT/CLARIAH-CZ (http://lindat.cz)

Faculty of Mathematics and Physics, Charles University *

Faculty of Arts, Charles University

Faculty of Arts, Masaryk University

Faculty of Informatics, Masaryk University

Faculty of Applied Sciences, University of West Bohemia in Pilsen

Library of the Czech Academy of Sciences

Moravian Library in Brno

National Library of the Czech Republic

Národní filmový archiv

National Gallery Prague

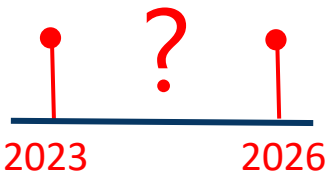
Czech Language Institute of the Czech Academy of Sciences

Institute of History of the Czech Academy of Sciences

Institute of Philosophy of the Czech Academy of Sciences

13 partners

* Center for Visual History Malach provides access to foreign oral history resources in the Holocaust and genocide research domains



operational/construction phase
join EHRI ERIC

- 4 new partners from European Holocaust Research Infrastructure (EHRI-CZ)
- oral + written resources

MASARYK INSTITUTE AND ARCHIVES OF THE CAS, v.v.i.

Terezín Initiative Institute

PAMÁTNÍK TEREZÍN
MALÁ PEVNOST
MUZEUM GHETTA
národní kulturní památka

Národní archiv

Data repository

As of Oct 1, 2022

- # items (metada + data) 1,395
- data volume 4.4 TB
- # unique downloads by unregistered users for fully open data 540k
- # downloads of items that require signing a license 1,200
- # registered users (only needed to submit data or sign a license) 1,218

Data repository

Non-linguistic content as well, e.g., collections created by the National Film Archive

Bohumil Veselý's Gallery of Famous Men and Women

This item is Publicly Available and licensed under:
Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

Name	3901475_12_hd-master.mov
Size	1.07 GB
Format	QuickTime video
Description	best quality
MD5	250b4401c3a2d9b98616321ff8c27e6f

Download file

Name	3901475_12_NAHLED.mp4
Size	54.9 MB
Format	MPEG-4 video
Description	lower quality
MD5	bf1b44128d276cea357bef6ec125038a

Download file

The Munich Agreement

This item is Publicly Available and licensed under:
Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

Name	ZF_0439_3901684_sot1_hd-master.mov
Size	1.07 GB
Format	QuickTime video
Description	best quality
MD5	47826f13ba8d8f7746f5be14518c243e

Download file

Name	ZF_0439_3901684_sot1_NAHLED.mp4
Size	68.58 MB
Format	MPEG-4 video
Description	lower quality
MD5	3823be78de2394de300c12cb74bcfd61

Download file

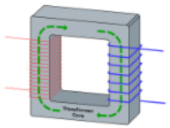
Language services

Machine translation

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Košarko, Ondřej; Variš, Dušan and Popel, Martin, 2019, *LINDAT Translation service*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2922.

Share:



Authors:
Martin Popel, Dušan Variš, Ondřej Košarko

Description:
A neural networks based translation service provides a simple UI and API that lets you use Transformer models trained by our experts. Five models are currently provided with more to come.

Project home Run

UDPipe

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Straka, Milan and Straková, Jana, 2016, *UDPipe*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-1702.

Terms of Use ▶

Most popular ▼

- Machine translation
- UDPipe
- KonText
- UWebASR

Text processing ▶

Search services ▶

Lexical resources ▶

Automatic speech recognition ▶

ToolService LINDAT / CLARIAH-CZ

Universal Dependencies 2.10 models for UDPipe 2 (2022-07-11)

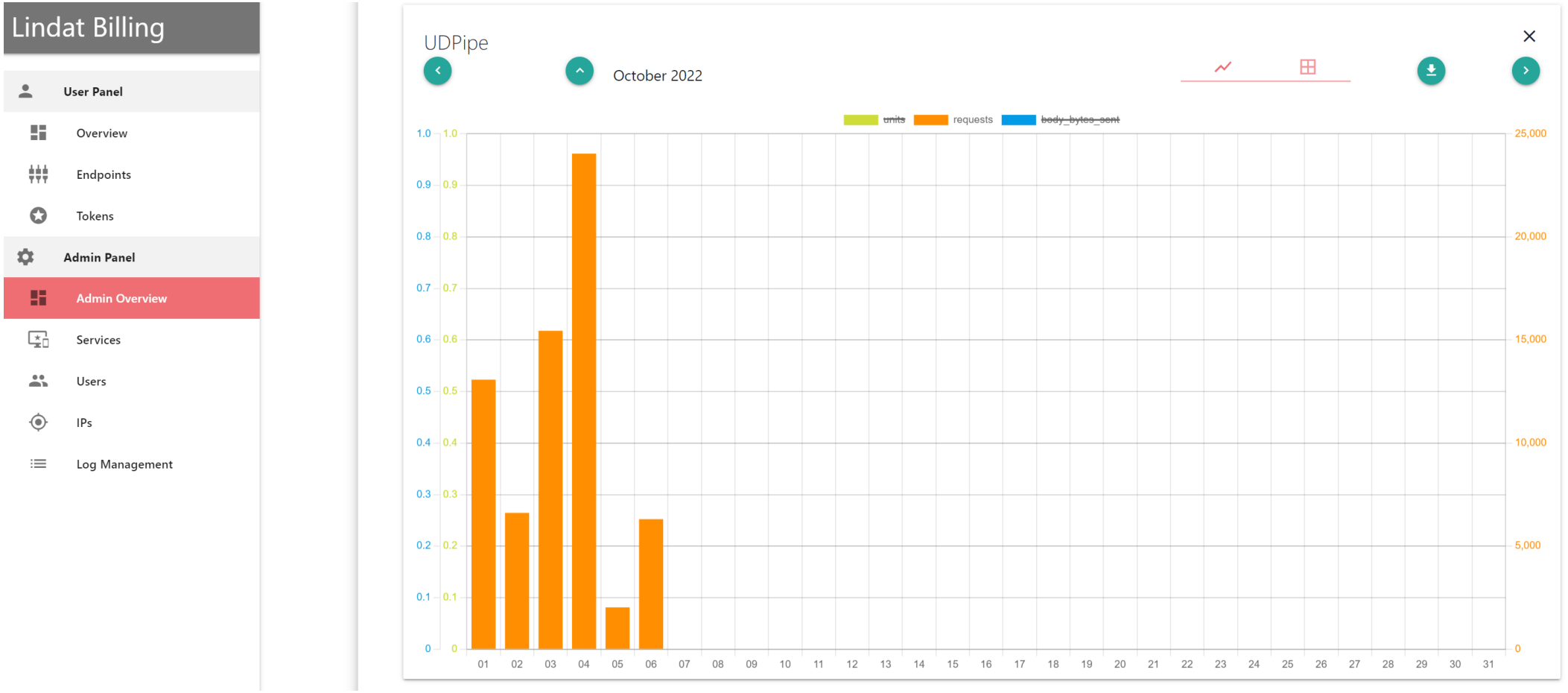
(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2022-07-11)

Author(s):
Straka, Milan

This item contains 1 file (6.88 GB).

Publicly Available

Language services



Machine translation

LINDAT/CLARIAH-CZ ∈ Institute of Formal and Applied Linguistics ∈ Charles University

Charles Translator

Charles Translator for Ukraine

Congratulations to Martin Popel and his u4u team on winning the Steven Krauwer Award

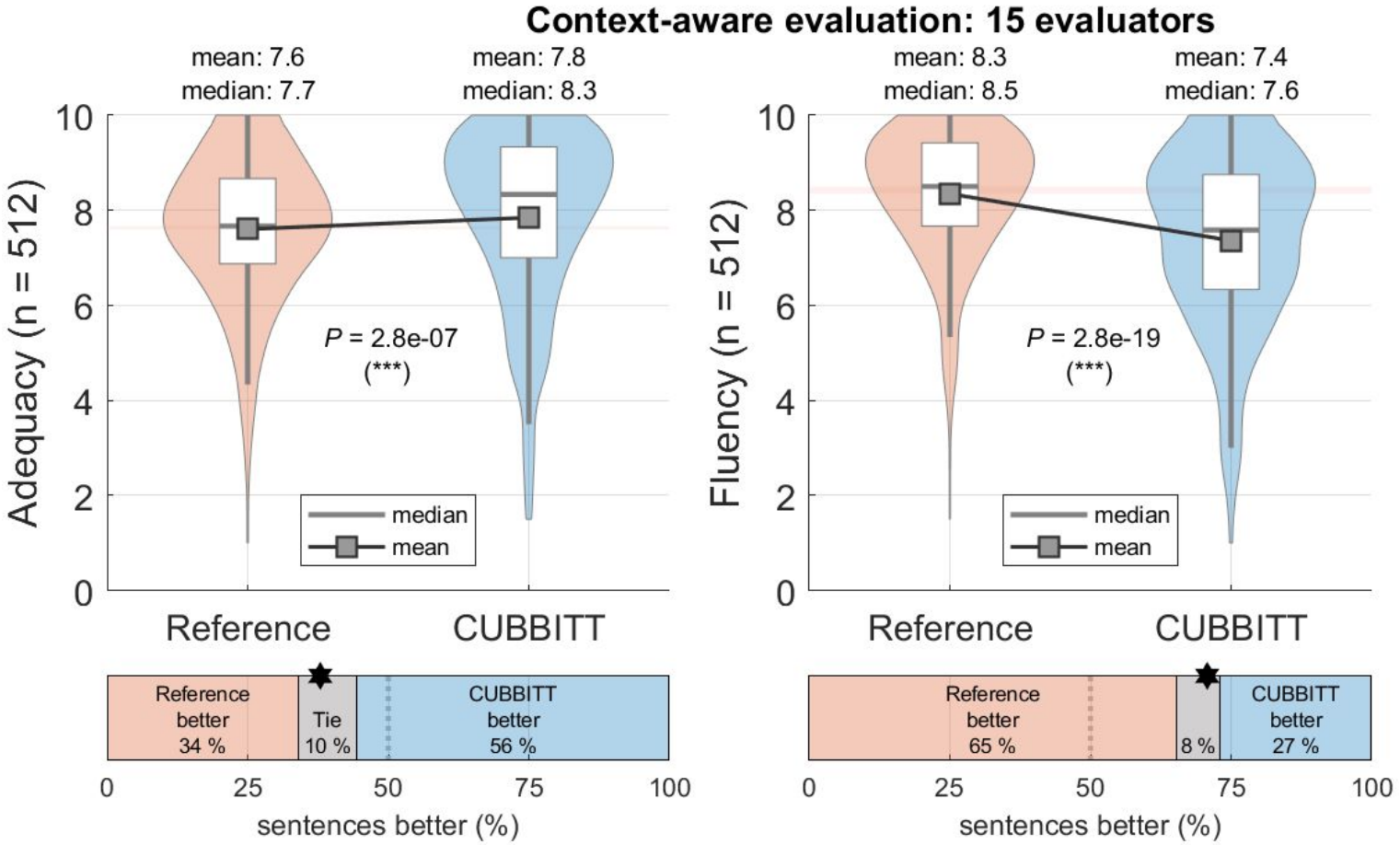
U4U (UFAL for Ukraine) Timeline

- 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

Experience with high-quality English-Czech NMT



Experience with high-quality English-Czech 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

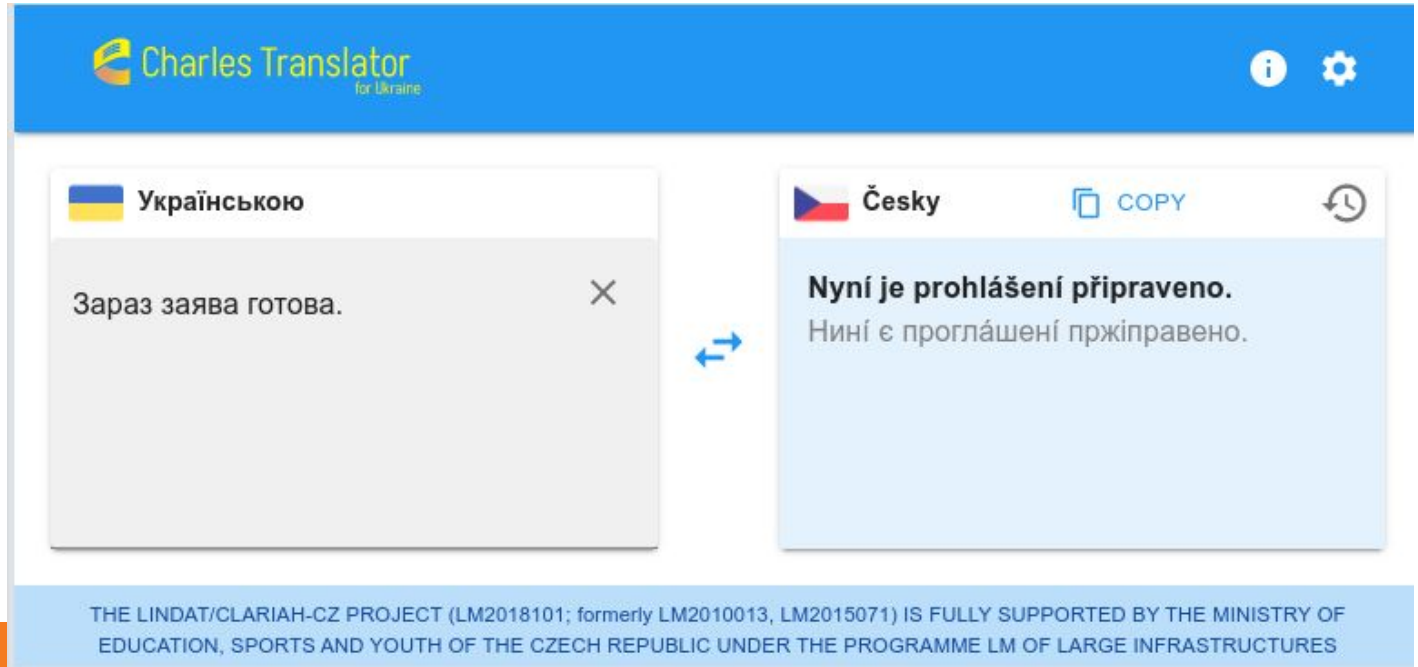
6273 Accesses | **76** Altmetric | Metrics

U4U (UFAL 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

U4U (UFAL for Ukraine) Timeline

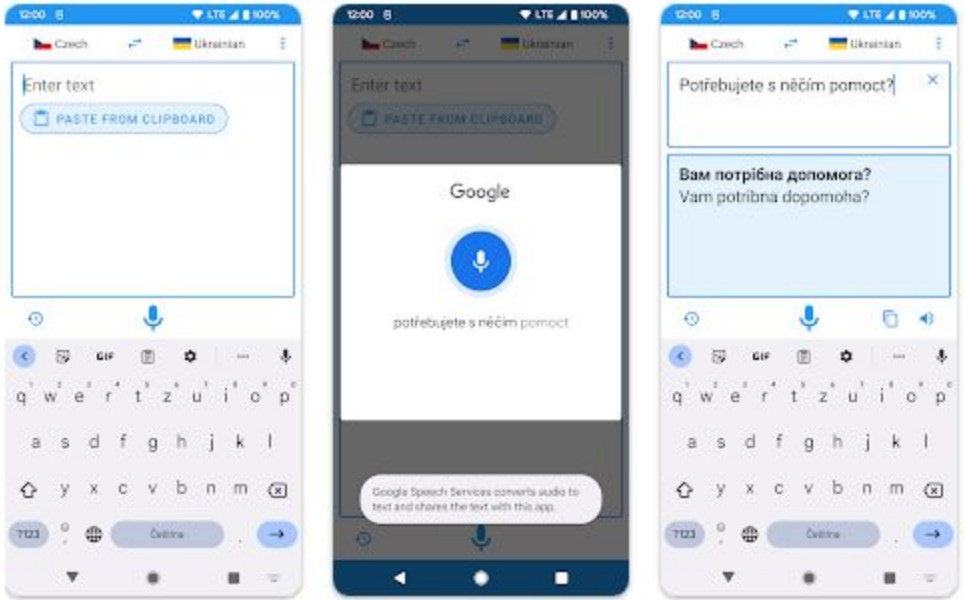
- March 1: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- March 12: hackathon to develop a new frontend
- March 14: first cs ⇆ uk translator published at LINDAT/CLARIN



U4U (UFAL 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

GET IT ON Google Play

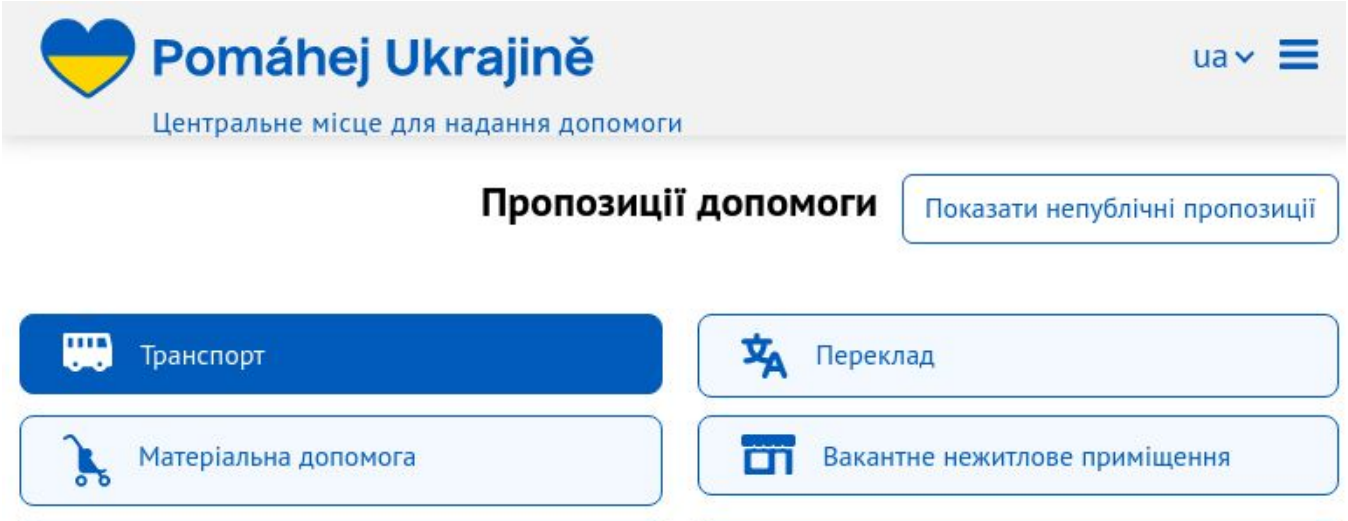


U4U (UFAL for Ukraine) Timeline

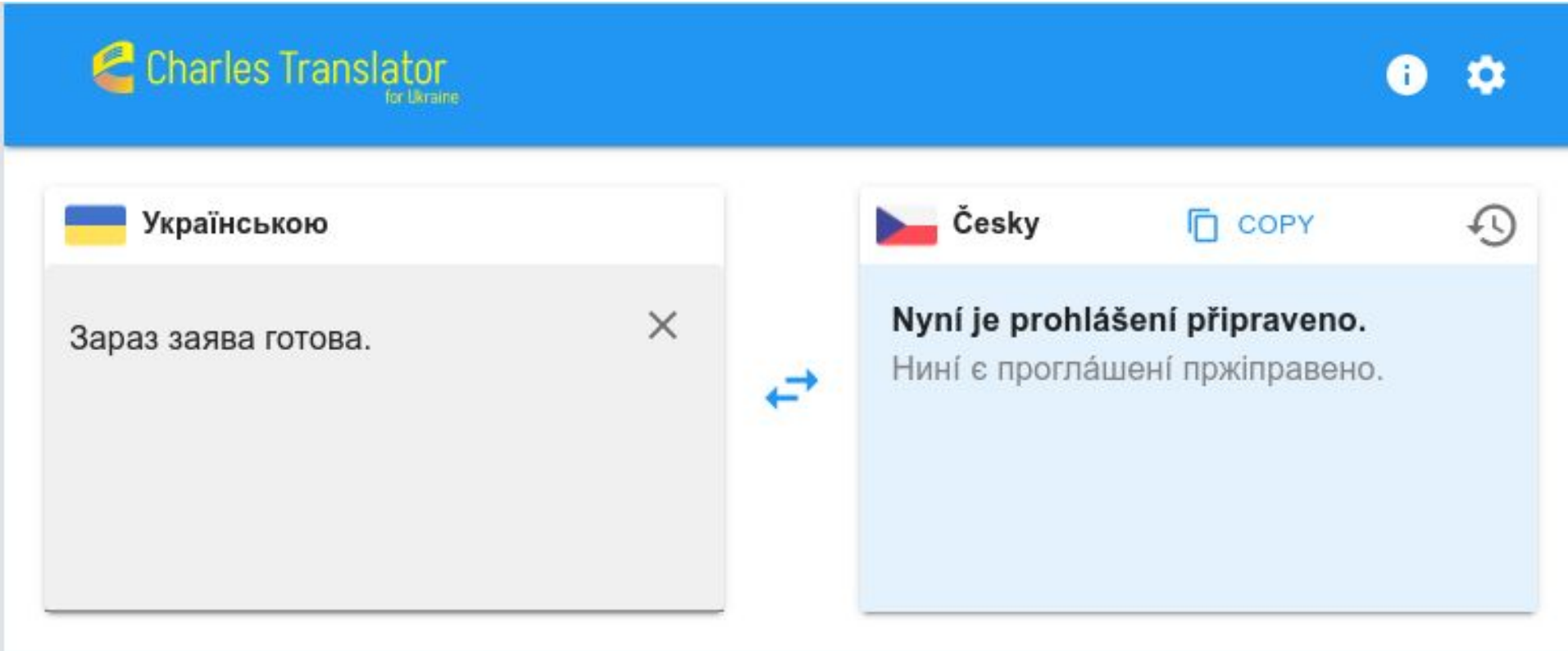
- March 1: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- March 12: hackathon to develop a new frontend
- March 14: first cs ⇆ uk translator published
- April 11: improved cs ⇆ uk models
- May 24: Android app
- since September: evaluating 11 systems in WMT22 shared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

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?

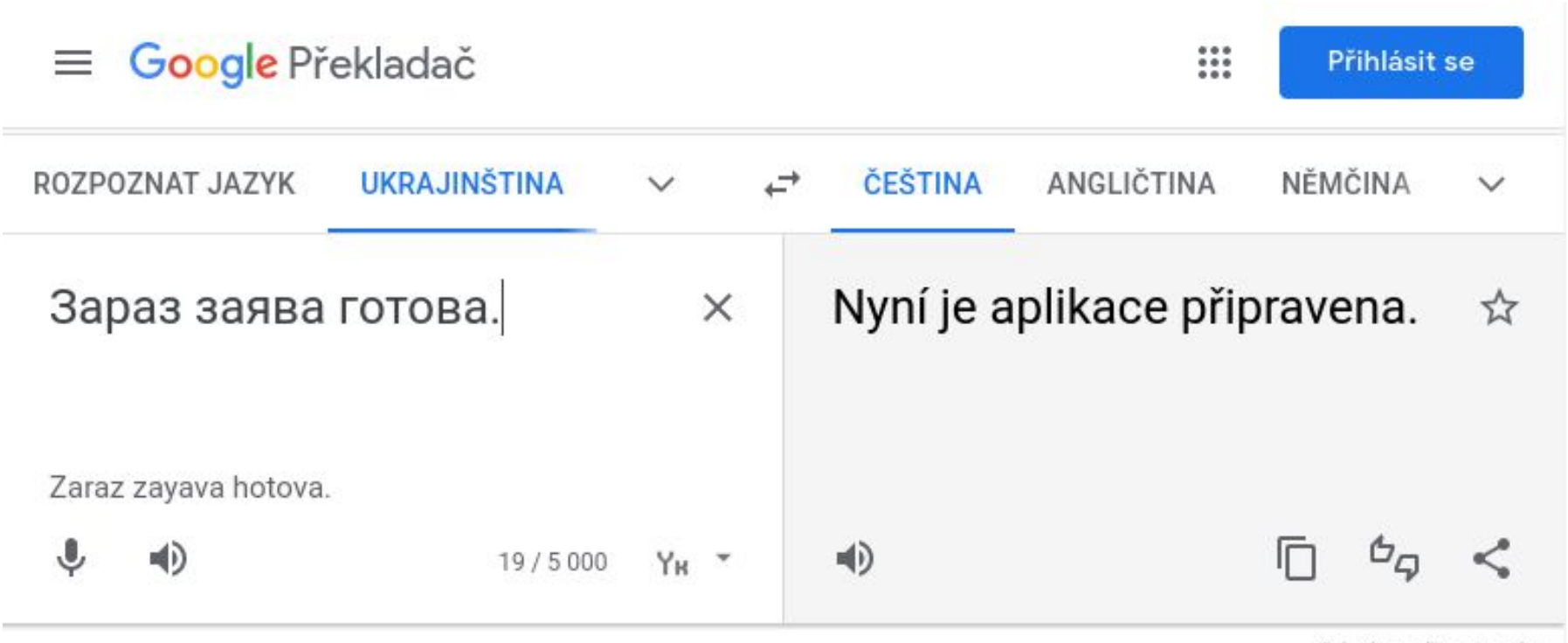


THE LINDAT/CLARIAH-CZ PROJECT (LM2018101; formerly LM2010013, LM2015071) IS FULLY SUPPORTED BY THE MINISTRY OF EDUCATION, SPORTS AND YOUTH OF THE CZECH REPUBLIC UNDER THE PROGRAMME LM OF LARGE INFRASTRUCTURES

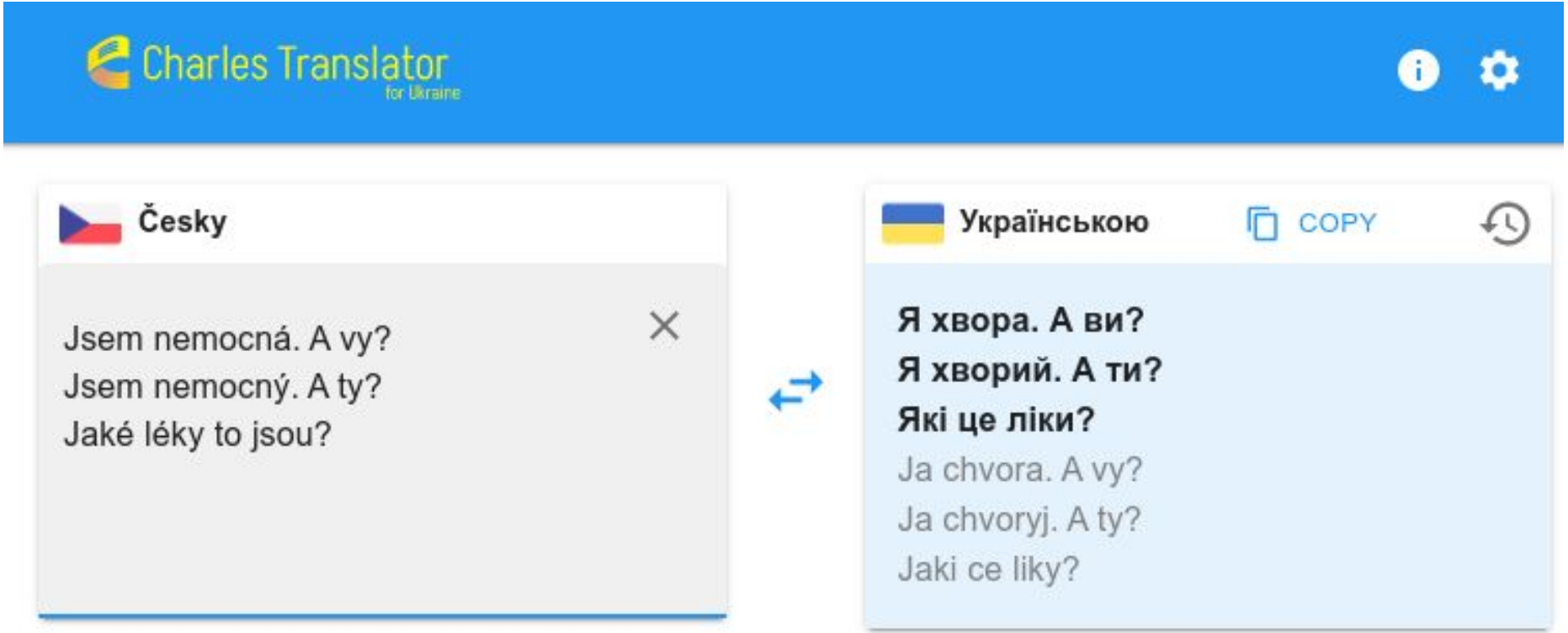
Why direct translation?

заява ⇒ application ⇒ aplikace

declaration *computer app*

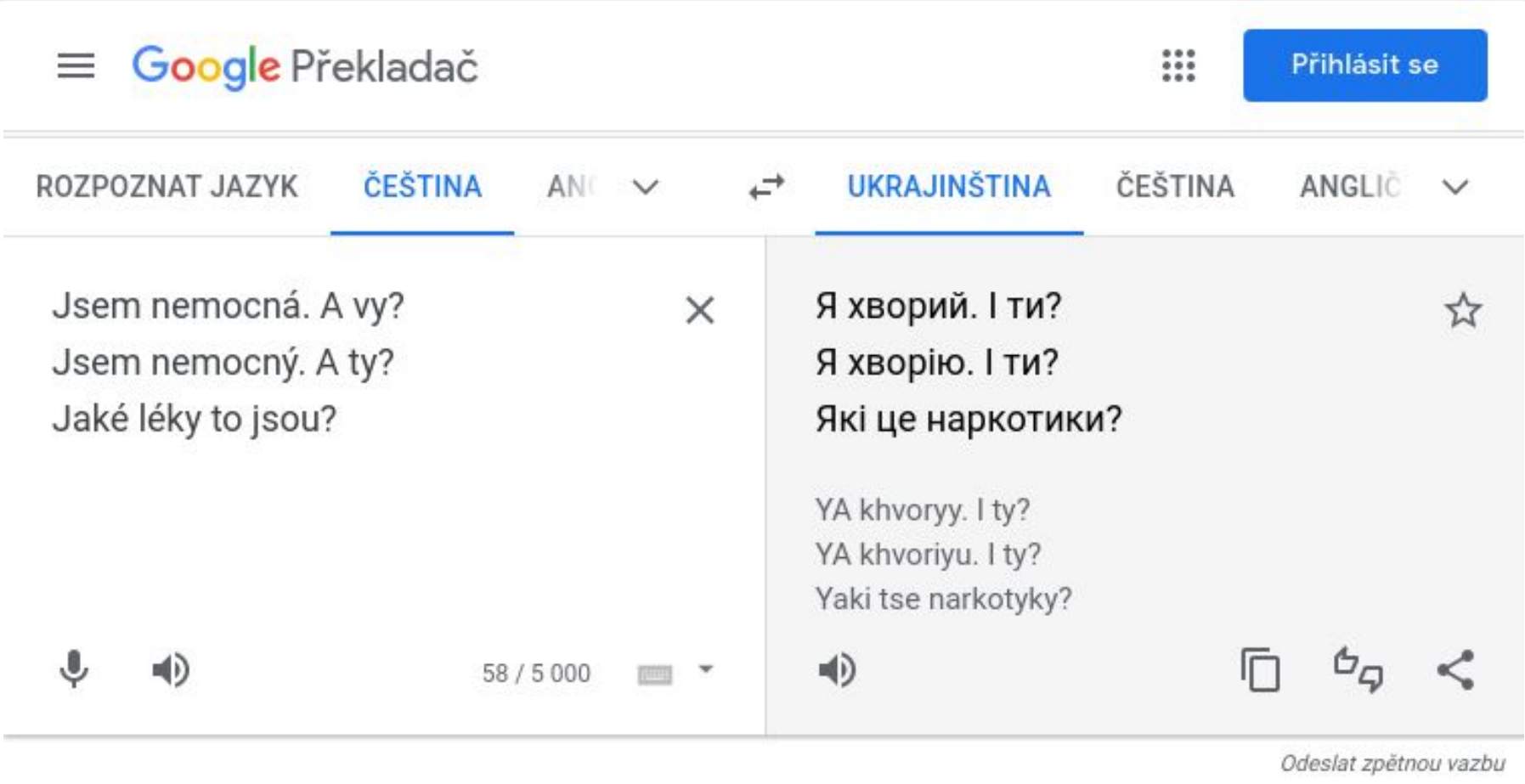


Why direct translation?

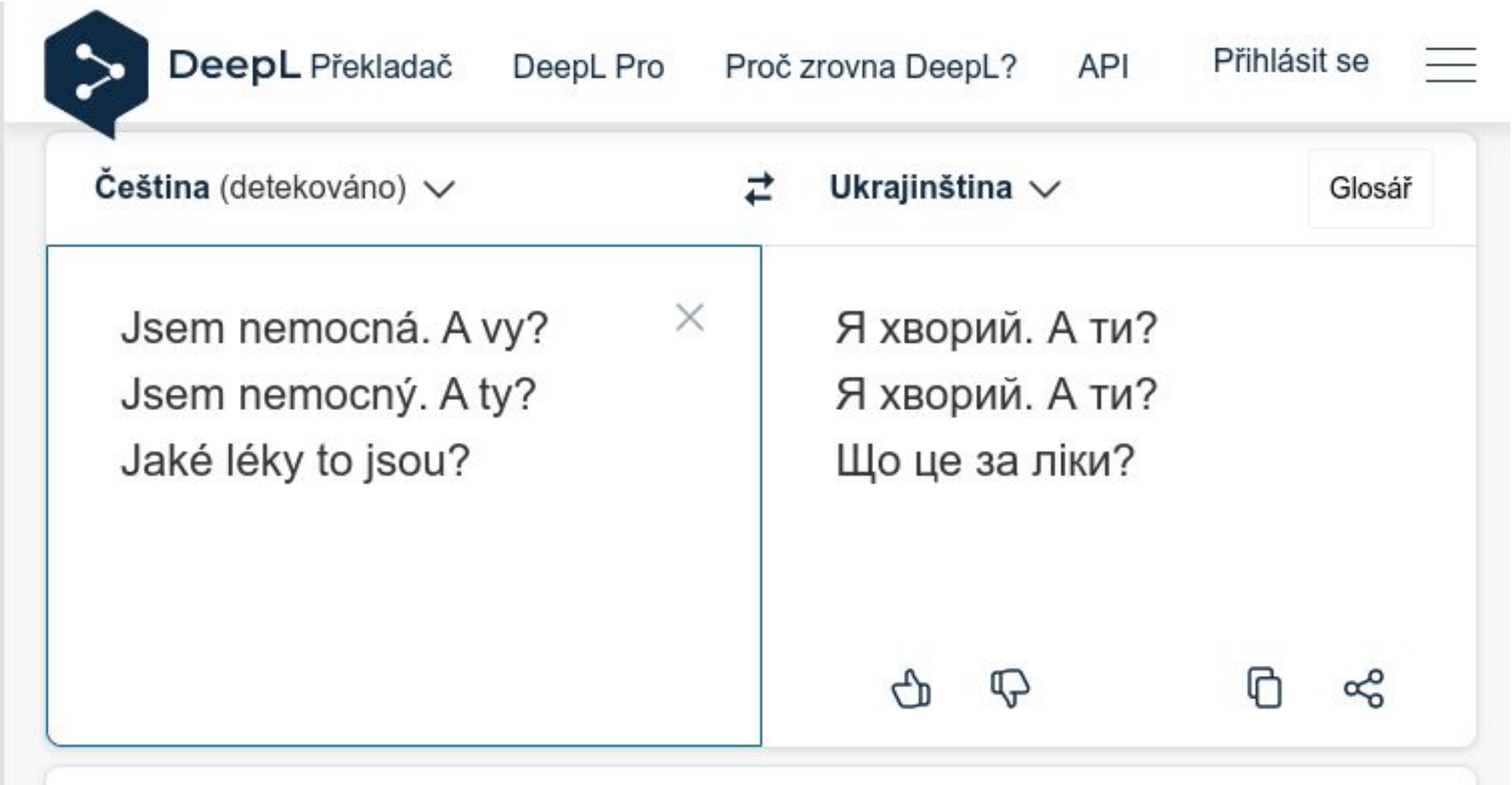


I am ill [fem]. And you [formal]?
I am ill [masc]. And you [informal]?
What kind of medicine is that?

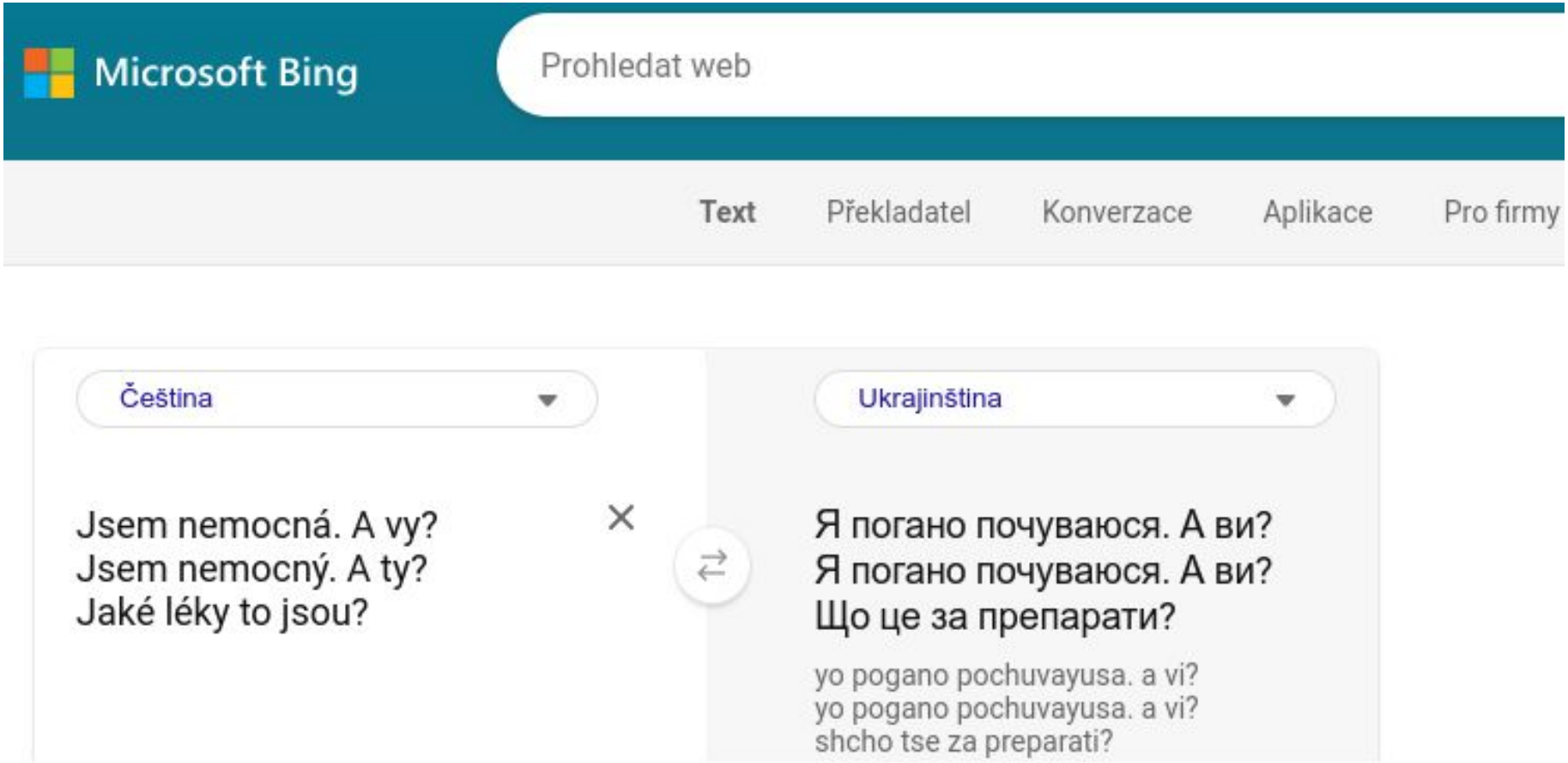
Why direct translation?



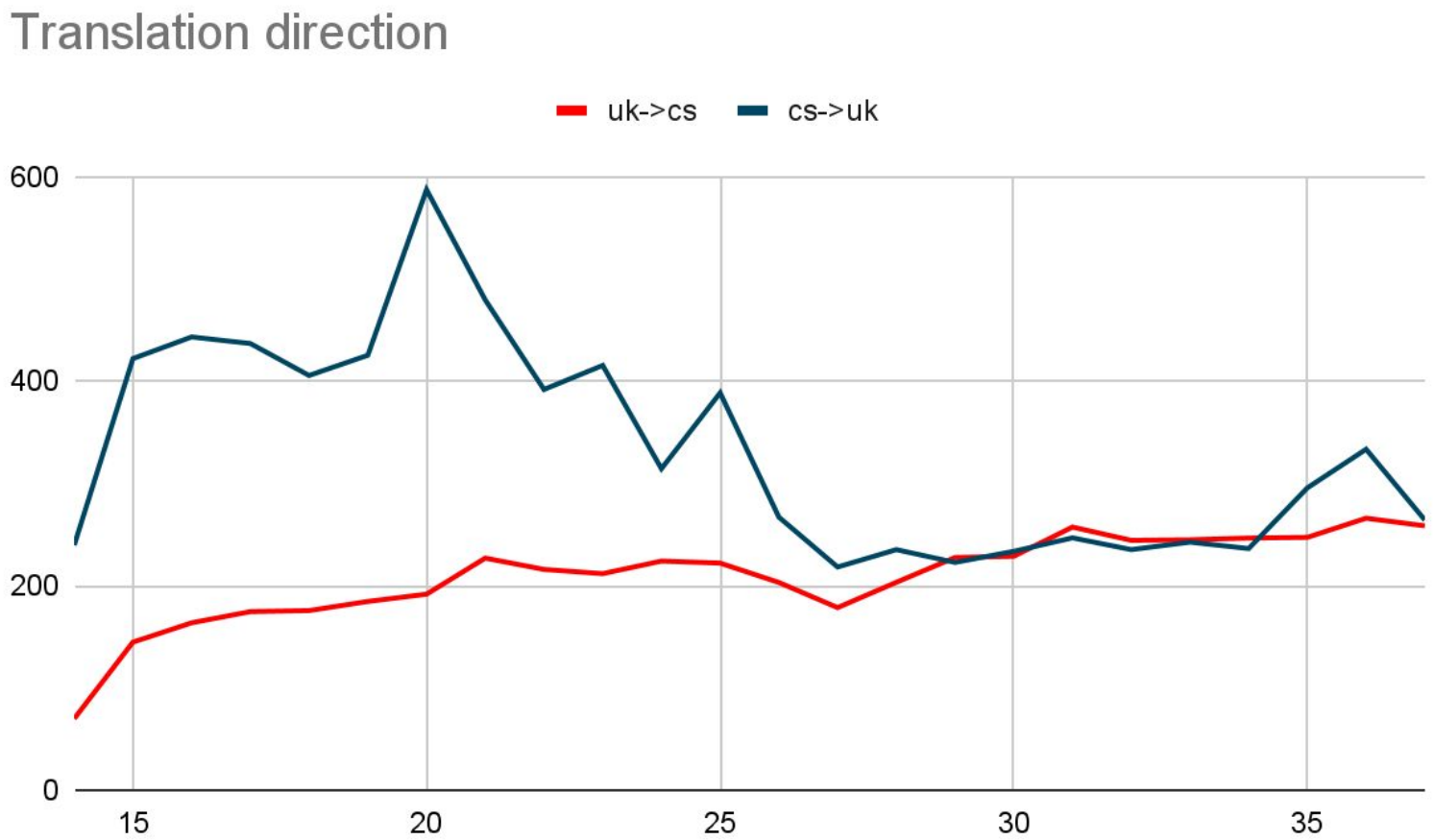
Why direct translation?



Why direct translation?



Thousands of translations per week (April-September)



Слава Україні!

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