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

2005
CLARIN
preparatory action(s)

2010
construction phase

2012
established

2014
operational phase
joined CLARIN ERIC

2019
operational/
preparatory phase
joined DARIAH ERIC

2022

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

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

13 partners

Terezín Initiative Institute
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

The Munich Agreement
Language services

Machine translation

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

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

Authors:
Martin Popel, Dušan Varší, Ondřej Košárko

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.

UDPipe

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

Straka, Milan and Straková, Jana, 2018, UDPipe, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-1702.

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 (UDS.gz)

Publicly Available
Language services
Machine translation

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

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

**Adequacy** (n = 512)
- Reference: mean 7.6, median 7.7
- CUBBITT: mean 7.8, median 8.3
- P = 2.8e-07 (***)

**Fluency** (n = 512)
- Reference: mean 8.3, median 8.5
- CUBBITT: mean 7.4, median 7.6
- P = 2.8e-19 (***)

Reference better: 65%
Tie: 10%
CUBBITT better: 27%
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

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, Lukasz Kaiser, Jakob Uszkoreit, Ondrej Bojar & Zdenek Zabokrtsky

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
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?

Українською

Nyní je prohlášení připraveno.

Nині є проглашений приправено.
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?

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

Я хворий. І ти?
Я хворію. І ти?
Які це наркотики?

YA khvoryy. I ty?
YA khvoriyu. I ty?
Yaki tse narkotyky?
Why direct translation?

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

Я хворий. А ти?
Я хворий. А ти?
Що це за ліки?
Why direct translation?

Microsoft Bing

Text

Prohledat web

Překladatel
Konverzace
Aplikace
Pro firmy

Čeština

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

Ukrainština

Я погано почуюся. А ви?
Я погано почуюся. А ви?
Що це за препарати?

Yo pogano pochuvayusa. a vi?
Yo pogano pochuvayusa. a vi?
Shcho tse za preparati?
Thousands of translations per week (April-September)

Translation direction

uk->cs  cs->uk
Try Czech-English: https://cubbitt.lindat.cz
Try Czech-Ukrainian: https://translator.cuni.cz

Слава Україні!