



LINDAT/CLARIAH-CZ

Large Research Infrastructure

Digital Research Infrastructure for Language Technology, Arts and Humanities

„Digitální výzkumná infrastruktura pro jazykové technologie, umění a humanitní vědy“

Funded by:



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

Jan Hajič

*LINDAT/CLARIAH-CZ coordinator
Institute of Formal and Applied Linguistics
Computer Science School
Faculty of Mathematics and Physics
Charles University, Prague, Czechia*

Christopher Cieri 1963-2023

Chris Cieri was a native Philadelphian who grew up among the diverse population in the neighborhoods of South Philadelphia. He had a lifelong affinity for the Italian language, and it was no surprise that when he entered the University of Pennsylvania as an undergraduate, he chose Linguistics for his major. In fact, he concluded his Masters' degree concurrently with his Bachelor studies in 1985-1986; his thesis topic was *Italian Lexical Items in the English Speech of Italo-Americans*. Under the direction of Bill Labov, he received his PhD at Penn as well. His dissertation, *Modeling Phonological Variation in Multidialectal Italy*, was based on fieldwork in L'Aquila, Italy, of which he had fond memories and spoke about often.



Linguistic Data Consortium, director

Long-term member of LINDAT IAB



In memory of the many victims.

@Faculty of Arts, Dec. 21, 2023

LINDAT/CLARIAH-CZ

- LINDAT = CLARIN & DARIAH & **EHRI** in Czechia
 - European networks for supporting research (in SSH)
 - CLARIN: Language resources and technology
 - DARIAH: Digital humanities and arts
 - **EHRI: European Holocaust Research Infrastructure**
 - LINDAT/CLARIAH-CZ: 2023-2026
 - Follows previous periods: 2010-2014, 2015-2019, 2019-2022
 - Combines membership in EU CLARIN, DARIAH and **EHRI** networks
 - **15** institutions in the Czech combined network (was: 4 in 2010, 11 in 2022)
 - UK, MU, Academy of Sciences, National and Moravian Libraries, NG, NFA
 - **New in 2023:**
 - » **National Archives**
 - » **Masaryk Institute and Archives CAS**
 - » **Terezín Memorial**
 - » **Terezín Initiative Institute**
 - » **complements the already included Center for Visual History Malach**

- LINDAT = CLARIN & DARIAH & **EHRI** in Czechia

- European network


- CLARIN: Language
- DARIAH: Digital
- **EHRI: European Holocaust Research Infrastructure**

- LINDAT/CLARIAH-CZ

- Follows previous
- Combines research
- **15** institutions
- UK, MU, Acad

- New in 2023

- » New
- » M
- » Te
- » IT
- » CC


CS | EN

EHRI Czech National Node

Top-quality research on the Holocaust is a prerequisite for informed discussion about Czech, European and world modern history and for understanding the risks and mechanisms of racism and genocide in their various forms. The **European Holocaust Research Infrastructure** (EHRI) connects collections and sources divided by borders and languages, promotes digital methods and supports researchers. The Czech EHRI national node is a gateway to EHRI services and community and a signpost for Holocaust research in the Czech Republic.

Research infrastructure

The EHRI Czech national node has been part of the research infrastructure **LINDAT/CLARIAH-CZ** supported by the Ministry of Education, Youth and Sports since 2023.


EHRI is funded by the European Commission under FP7, Horizon 2020 and Horizon Europe.

Since 2018, EHRI has been on the roadmap of European research infrastructures and is currently transforming into a permanent organisation - **European Research Infrastructure Consortium** (ERIC). The Czech Republic supports this process through the Ministry of Education, Youth and Sports and is represented in the EHRI Interim General Assembly.

Main goals and services

EHRI-CZ provides the following types of services:

- Makes EHRI data and services available, including EHRI Portal, EHRI Document Blog, EHRI Editions, EHRI Geospatial Repository, Conny Kristel Scholarship, and others.
- Creates data on sources on the history of the Holocaust in the Czech lands and uploads them to the EHRI Portal.
- Links and enhances the victim databases and methodologically supports their further development.
- Applies digital methods to the digitised sources, including automatic text and speech recognition, identification of places, historical actors, keywords, etc.
- Supports research using spatial methods and interactive maps, including the MemoMAP application.



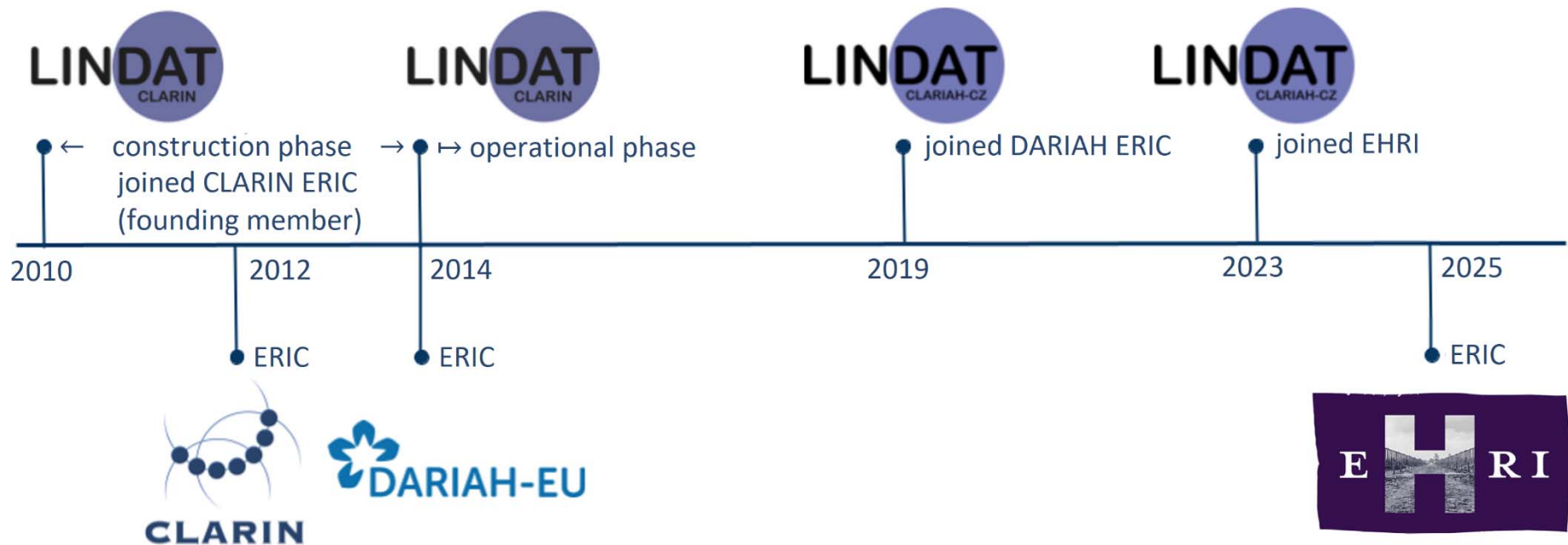
2)
A



LINDAT/CLARIAH-CZ 2010-2026



- LINDAT timeline + Czechia in ERICs



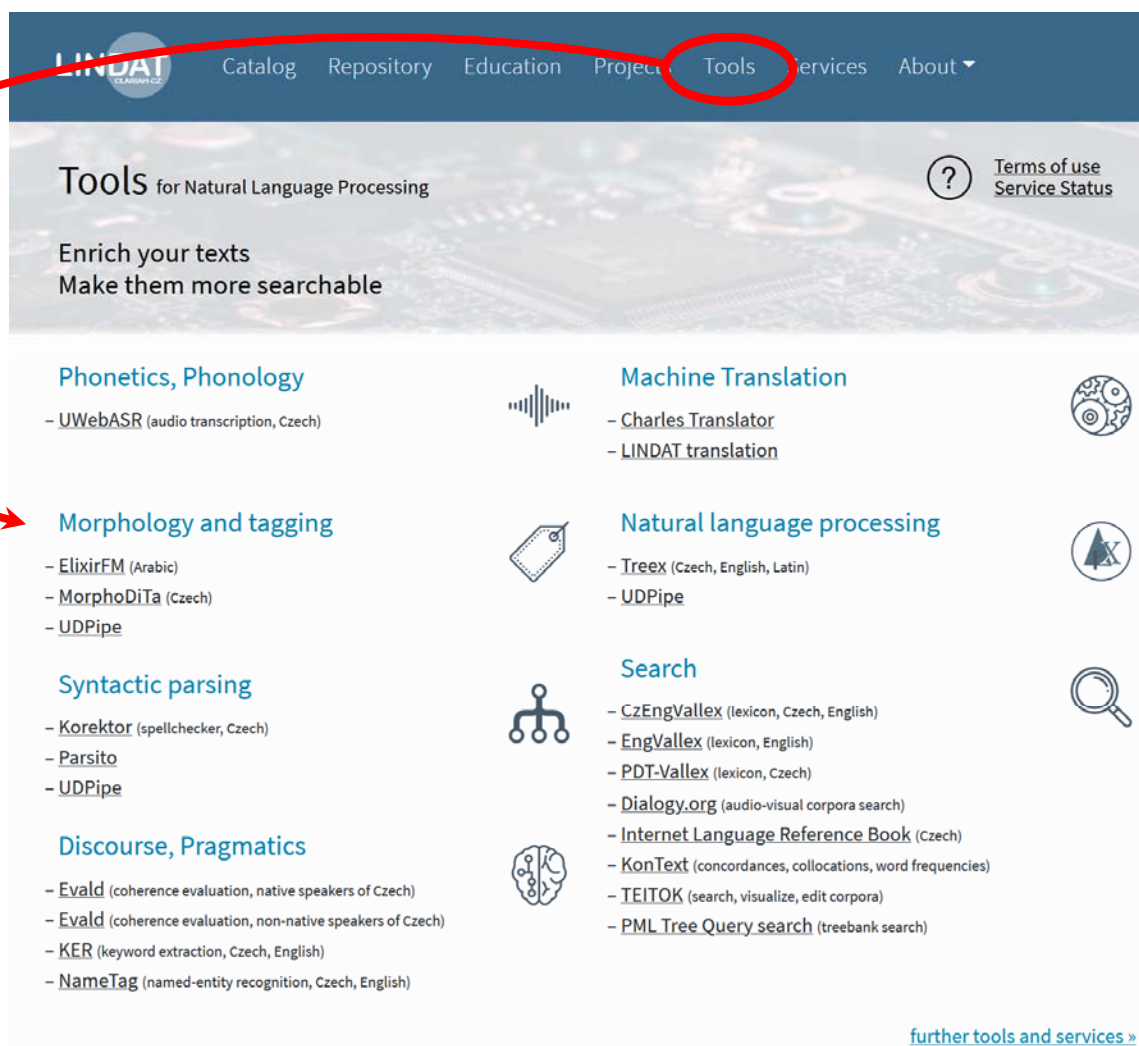
LINDAT/CLARIAH-CZ in European Research Infrastructures

- ERIC = European Research Infrastructure Consortium
 - Supervising body
 - ESFRI (European Strategy Forum on Research Infrastructures)
 - Members are countries (EU member or associated states)
 - ... or international organizations (NTU); associated countries possible (U.S.)
- CLARIN ERIC
 - Original network (since 2010)
 - Consortium members in
 - NCF, SCCTC, CLIC, Annual Conference committee, award committees
- DARIAH ERIC
 - Since 2014
 - Consortium members in
 - NCC, legal VCC
- EHRI network (soon to be ERIC, January 2025)
 - EHRI: 3 preparatory projects by EU
 - EHRI CZ: partner in all 3 preparatory projects

LINDAT/CLARIAH-CZ

- <http://lindat.cz> (redirected to <https://ufal.mff.cuni.cz/lindat>)
- Data, Tools, Services, Expertise
 - Language Resources Data Repository
 - Software tools repository
 - Web services with both API and User Interactive Interface
- Computing infrastructure
 - New hardware for deep learning computations due in 2024-2026
 - Coming as we speak – for 7 partner institutions
- Cooperation with other Czech RI from SSH
 - Czech National Corpus (CNC)
 - Czech Literary Bibliography (CLB)
 - Archaeological Information System of the Czech Republic (AIS CR)

- Web at <https://lindat.cz>




The screenshot shows the LINDAT website interface. The 'Tools' menu item in the top navigation bar is circled in red. A red arrow points from this menu item to the 'Morphology and tagging' category in the main content area. The website header includes the LINDAT logo and navigation links: Catalog, Repository, Education, Projects, Tools, Services, and About. The main content area is titled 'Tools for Natural Language Processing' and includes a sub-header 'Enrich your texts Make them more searchable'. Below this, there are several tool categories, each with a list of tools and an icon:



- Phonetics, Phonology**
 - UWebASR (audio transcription, Czech)
- Morphology and tagging**
 - ElixirFM (Arabic)
 - MorphoDiTa (Czech)
 - UDPipe
- Syntactic parsing**
 - Korektor (spellchecker, Czech)
 - Parsito
 - UDPipe
- Discourse, Pragmatics**
 - Evald (coherence evaluation, native speakers of Czech)
 - Evald (coherence evaluation, non-native speakers of Czech)
 - KER (keyword extraction, Czech, English)
 - NameTag (named-entity recognition, Czech, English)
- Machine Translation**
 - Charles Translator
 - LINDAT translation
- Natural language processing**
 - Treex (Czech, English, Latin)
 - UDPipe
- Search**
 - CzEngVallex (lexicon, Czech, English)
 - EngVallex (lexicon, English)
 - PDT-Vallex (lexicon, Czech)
 - Dialogy.org (audio-visual corpora search)
 - Internet Language Reference Book (Czech)
 - KonText (concordances, collocations, word frequencies)
 - TEITOK (search, visualize, edit corpora)
 - PML Tree Query search (treebank search)


At the bottom right of the page, there is a link: [further tools and services »](#)

- Repository development (CLARIN DSpace)
 - Upgrade to version 7.x of DSpace – major undertaking, w/external partners
 - Will be used by Charles University Library for Open Data Library
- Universal Dependencies project
 - General support, official editions (every 6 months), now at v2.13 (Nov. 2023)
 - Tools: UDPipe (now UDPipe 2 -> LinPipe soon)
- Uniform Meaning Representation project
 - With partners in the U.S. (Brandeis, Colorado, New Mexico)
 - Official release repository for UMR 1.0

- Prague Dependency Treebank – Consolidated ed. (PDT-C 1.0 -> 2.0)
 - Approx. 4MW, manual dependency syntax for all parts underway
 - PDT, PCEDT, PDTSC, Faust; TR annotation same, a-layer: full manual, corrections
- Prague Discourse Treebank – on top of PDT
 - Adding cross-sentence annotation
- SynSemClass
 - Multilingual (cs, en, de, es, [kr in the works]) event type ontology, now v5.0
 - A.k.a synonym dictionary of verbs, with valency, linked to corpora and other lexical resources
 - New search tool in 2023
- TEI:TOK
 - Platform for rich annotation and multimedia content
 - Many new datasets, continued ParlaMint project (led by CLARIN-SI)
- New ASR service (UWB Pilsen)
- Consolidation of MT interface underway

Virtual Language Observatory Search Contributors Help CLARIN 

VLO / Faceted search / Search results  

parliament 

Showing 1 to 10 of 19 results within selection for parliament Public or Academic Czech Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language ⌵

Type to filter or search for more

- Czech ✕
- English (8172)
- Bulgarian (206)
- Modern Greek (59)
- Swedish (32)
- Finnish (20)
- French (20)
- Slovenian (19)
- German (16)
- Polish (16)
- Danish (15)
- more...

Collection ⌵

<< < 1 2 > >>

Czech Parliament Meetings

(Part of LINDAT / CLARIAH-CZ Data & Tools)

⊕ The corpus consists of recordings from the Chamber of Deputies of the **Parliament** of the Czech Republic. It currently consists of 88 hours of speech data, which corresponds roughly to 0.5 million tokens. The annotation process is semi-automatic, as we are able to perform the speech recognition on the data with high accu...

Czech

[Landing page for this record](#)

18 19 CC P ⓘ Ⓔ

Large Corpus of Czech Parliament Plenary Hearings

(Part of LINDAT / CLARIAH-CZ Data & Tools)

⊕ We present a large corpus of Czech **parliament** plenary sessions. The corpus consists of approximately 444 hours of speech data and corresponding text transcriptions. The whole corpus has been segmented to short audio snippets making it suitable for both training and evaluation of automatic speech recognition (ASR) s...

Czech

[Landing page for this record](#)

1 CC P ⓘ

Český parlamentní korpus, Poslanecká sněmovna, 2015-10-22 ps2013-033-07-001-205 [ParCzech.ana]

Speed: 100% 9:59.949 / 13:58.775 Zoom: 120 pps

audio.mp3

▶ play from start of transcription

Děkuji. Nyní pan ministr financí Andrej Babiš, poté pan poslanec Volný. Prosím, pane ministře.

Místopředseda vlády ČR a ministr financí Andrej Babiš

To je klasická debata o ničem. Pět miliard je lež vašeho Kalouska a dobře to víte. Tak zase lžete. Nebo kabely. Žádné **kabely** jsem nenosil. V životě jsem tady nebyl v Poslanecké sněmovně předtím, než jsem se stal politikem. Tak zase lžete. Tak to je zbytečná debata. Rozdíl mezi těma dvěma knížkami, že jednu napsal Nadační fond proti korupci a tu druhou napsal nějaký novinář na objednávku, to je všechno. Já jsem řekl můj názor a já se nikomu omlouvat nebudu. Já si za tím stojím. A to je všechno.

Místopředseda PSP Vojtěch Filip

Nyní pan poslanec Jan Volný, potom pan poslanec Vácha. Prosím, pane poslanče.

- New ASR service (UWB Pilsen)
- Consolidation of MT interface underway

- Running services (<https://lindat.cz/en/services>)
- Each service available as
 - Software to download – for those able to run it locally, modify etc (Open Source)
 - Models to download – e.g. for machine translation, for locally run software
 - API, for running programmatically but remotely
 - User interface (running on small samples, cut & paste, etc.) – open to everybody
- Examples
 - UDPipe 2.0
 - Linguistic analysis: segmentation, morphology, lemmatization, POS, parsing, NER
 - 100+ languages
 - Machine translation
 - en-cs and en-fr at human level of accuracy
 - Includes uk-cs
 - Also on Android (Charles Translator)
 - Dictionaries
 - Valency, semantics

- Running
- Each s
- Soft
- Mod
- API
- Use
- Example
- UD
- Ma
- Dic

lindat.mff.cuni.cz/services/udpipe/

Service

The service is freely available for testing. Respect the CC BY-NC-SA licence of the models – explicit written permission of the authors is required for any commercial exploitation of the system. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

Model: UD 2.12 (description) UD 2.10 (description) UD 2.8 (description) PDT-C 1.0 (description) EvaLatin20 (description)

japanese-gsd-ud-2.12-230717

Actions: Tag and Lemmatize Parse

Advanced Options

A Input Text

大リーグの大谷翔平選手がドジャースに移籍することを自身のインスタグラムで発表した。契約は10年で7億ドル (約1015億円).

Process Input

A Output Text

Save Output File

```
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe
# udpipe_model = japanese-gsd-ud-2.12-230717
# udpipe_model_licence = CC BY-NC-SA
# newdoc
# newpar
# sent_id = 1
# text = 大リーグの大谷翔平選手がドジャースに移籍することを自身のインスタグラムで発表した。
1 大 大 NOUN 接頭辞 _ 2 compound _ SpaceAfter=No|TokenRange=0:1
2 リーグ リーグ NOUN 名詞-普通名詞-一般 _ 6 nmod _ SpaceAfter=No|TokenRange=1:4
3 の の ADP 助詞-格助詞 _ 2 case _ SpaceAfter=No|TokenRange=4:5
4 大谷 大谷 PROPN 名詞-固有名詞-人名-姓 _ 6 compound _ SpaceAfter=No|TokenRange=5:7
5 翔平 翔平 PROPN 名詞-固有名詞-人名-名 _ 6 compound _ SpaceAfter=No|TokenRange=7:9
6 選手 選手 NOUN 名詞-普通名詞-一般 _ 10 nsubj _ SpaceAfter=No|TokenRange=9:11
7 が が ADP 助詞-格助詞 _ 6 case _ SpaceAfter=No|TokenRange=11:12
8 ドジャース ドジャース PROPN 名詞-固有名詞-一般 _ 10 obl _ SpaceAfter=No|TokenRange=12:17
9 に に ADP 助詞-格助詞 _ 8 case _ SpaceAfter=No|TokenRange=17:18
10 移籍 移籍 VERB 名詞-普通名詞-サ変可能 _ 12 acl _ SpaceAfter=No|TokenRange=18:20
11 する 為る AUX 動詞-非自立可能-サ行変格 _ 10 aux _ SpaceAfter=No|TokenRange=20:22
12 こと 事 NOUN 名詞-普通名詞-一般 _ 18 obj _ SpaceAfter=No|TokenRange=22:24
13 を を ADP 助詞-格助詞 _ 12 case _ SpaceAfter=No|TokenRange=24:25
14 自身 自身 NOUN 名詞-普通名詞-副詞可能 _ 16 nmod _ SpaceAfter=No|TokenRange=25:27
15 の の ADP 助詞-格助詞 _ 14 case _ SpaceAfter=No|TokenRange=27:28
16 インスタグラム インスタグラム NOUN 名詞-普通名詞-一般 _ 18 obl _ SpaceAfter=No|TokenRange=28:35
17 で で ADP 助詞-格助詞 _ 16 case _ SpaceAfter=No|TokenRange=35:36
18 発表 発表 VERB 名詞-普通名詞-サ変可能 _ 0 root _ SpaceAfter=No|TokenRange=36:38
19 し 為る AUX 動詞-非自立可能-サ行変格 _ 18 aux _ SpaceAfter=No|TokenRange=38:39
20 た た AUX 助動詞-助動詞-タ _ 18 aux _ SpaceAfter=No|TokenRange=39:40
21 。 。 PUNCT 補助記号-句点 _ 18 punct _ SpaceAfter=No|TokenRange=40:41

# sent_id = 2
# text = 契約は10年で7億ドル (約1015億円).
```

ng, NER

- Running
- Each s
 - Soft
 - Mod
 - API,
 - Use
- Exampl
 - UD
 -
 -
 -
 -
 -
 -
 -
 -
 -
 - Ma
 -
 -
 -
 -
 - Dic
 -

lindat.mff.cuni.cz/services/udpipe/

UDPipe is a free software distributed under the [Mozilla Public License 2.0](#) and the linguistic models are free for non-commercial use and distributed under the [CC BY-NC-SA](#) license, although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using [Semantic Versioning](#).

Copyright 2017 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic.

Description of the available methods is available in the [API Documentation](#) and the models are described in the [UDPipe 2 models list](#) and [UDPipe 1 models list](#).

Service

The service is freely available for testing. Respect the [CC BY-NC-SA](#) licence of the models – **explicit written permission of the authors is required for any commercial exploitation of the system**. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

Model: UD 2.12 ([description](#)) UD 2.10 ([description](#)) UD 2.6 ([description](#)) PDT-C 1.0 ([description](#)) EvaLatin20 ([description](#))

japanese-gsd-ud-2.12-230717

Actions: Tag and Lemmatize Parse

Advanced Options

Input Text Input File

大リーグの大谷翔平選手がドジャースに移籍することを自身のインスタグラムで発表した。契約は10年で7億ドル (約1015億円)。

Process Input

Output Text Show Table Show Trees

Save Tree as SVG

Previous 1 2 Next

大リーグの大谷翔平選手がドジャースに移籍することを自身のインスタグラムで発表した。

NER

LINDAT Translation

Translate
Docs

The translation service is available for **personal and non-commercial use** (see [terms of use](#) for more details).

Source

Ukrainian

advanced

Input sentences

Нова стрічка режисерки Софії Копполи – біографічна драма "Прісцилла" – стала однією з найочікуваніших прем'єр цьогорічного Венеційського кінофестивалю. Щоб представити стрічку на острові Лідо, виконавці головних ролей – Кейлі Спені (яка, до речі, отримала у Венеції відзнаку за свою гру) та Джейкоб Ейлорді – отримали дозвіл від Гільдії кіноакторів США, яка на той час перебувала на піку історичного акторського страйку.

Translate
Choose file

Target

English

Translation

Director Sofia Coppola's new film, the biographical drama Priscilla, has become one of the most anticipated premieres of this year's Venice Film Festival. To launch the film on the island of Lido, the leading roles - Kaley Spaney (who, incidentally, won an award in Venice for her play) and Jacob Aylordy - were given permission by the American Screen Actors Guild, which was at the height of a historic acting strike at the time.

Credits: The service runs systems trained by:

Martin Popel
 CUBBITT models, en<->cs, en<->fr, en<->pl as described in
 Popel, M., Tomkova, M., Tomek, J. et al. *Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals*. Nat Commun 11, 4381 (2020). <https://doi.org/10.1038/s41467-020-18073-9>

CUBBITT models, en<->cs, en<->fr, en<->pl as described in

Popel, M., Tomkova, M., Tomek, J. et al. *Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals*. Nat Commun 11, 4381 (2020). <https://doi.org/10.1038/s41467-020-18073-9>

LINDAT/CLARIAH-CZ partner services (1)

- Institute of the Czech Language, Czech Academy of Sciences (CAS)
 - Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
 - In cooperation with Masaryk Univ., Faculty of Informatics
- Faculty of Arts, Charles University (FF UK)
 - LINDAT team participates at building the Center for Digital Humanities at FF UK
 - Corpora (diachronic, educational), Exhibitions (manuscripts)
 - Didaktikon project – educational project with National Library, other universities
- Masaryk University Brno – Faculty of Arts
 - Own digital repository at <https://digitalia.phil.muni.cz/en> (Digitalia MUNI ARTS | PHIL)
- University of West Bohemia, Faculty of Applied Sciences
 - Technology – speech recognition services
- Libraries (National Library, Moravian Library, Library of CAS)
 - Upgrade to Kramerius 7 (underlying library management system)
 - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)

• Insti

– I

– I

• Facu

– L

– C

– D

• Mas

– C

• Univ

– T

• Libra

– U

– C

[Hlavní stránka](#)

[O příručce](#)

[Nápověda](#)

[Mobilní verze](#)

[Návštěvnost](#)

[English version](#)

Související odkazy:

[Jazyková poradna](#)

[ČSN 01 6910](#)

[Zajímavé dotazy](#)

[Databáze dotazů](#)

Hledání konkrétního slova nebo tvaru slova.

procházet

dělení: pro-chá-zet¹

	jednotné číslo	množné číslo
1. osoba	procházím	procházíme
2. osoba	procházíš	procházíte
3. osoba	prochází	procházejí, prochází ²
rozkazovací způsob	procházej ³	procházejte
příčestí činné	procházel	
příčestí trpné	procházen	
přechodník přítomný, m.	procházej ⁴	procházejíce
přechodník přítomný, ž. + s.	procházejíc	
verbální substantivum	procházení	

příklady: *Procházel jednu položku za druhou, ale chybu stále nenacházel. Procházel se po parku celé odpoledne.*

Heslové slovo bylo nalezeno také v následujících slovnících: [SSČ](#), [SSJČ](#)

Další slovní charakteristiky a příklady: [ČNK](#)

Odkazy k výkladové části Internetové jazykové příručky:

¹[Dělení slov na konci řádku](#)

²[Slovesa vzoru „sázet“](#)

³[Rozkazovací způsob 4. slovesné třídy \(leč – léči\)](#)

⁴[Přechodníky](#)

acc/yr

JK

ES

| PHIL)

LINDAT/CLARIAH-CZ partner services (1)

- Institute of the Czech Language
 - Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
 - In cooperation with Masaryk Univ., Faculty of Informatics
- Faculty of Arts, Charles University (FF UK)
 - LINDAT team participates at building the Center for Digital Humanities at FF UK
 - Corpora (diachronic, educational), Exhibitions (manuscripts)
 - Didaktikon project – educational project with National Library, other universities
- Masaryk University Brno – Faculty of Arts
 - Own digital repository at <https://digitalia.phil.muni.cz/en> (Digitalia MUNI ARTS | PHIL)
- University of West Bohemia, Faculty of Applied Sciences
 - Technology – speech recognition services
- Libraries (National Library, Moravian Library, Library of CAS)
 - Upgrade to Kramerius 7 (underlying library management system)
 - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)

FACULTY OF ARTS
Charles University

Search

Centre for Digital Humanities

Home > Faculty > Departments and Institutes > Centre for Digital Humanities

The Centre for Digital Humanities brings together primarily the faculty and students at the Faculty of Arts, Charles University with a shared interest in digital humanities.

The mission of the CDH is to support and coordinate pedagogical, scientific, research and outreach activities at the faculty, especially in the field of digital humanities.

These include:

- Developing methodologies for acquiring, processing, analyzing, presenting and preservation of data in digital form for the humanities, social sciences and arts.
- Collaboration with other departments of similar focus, creating conditions for coordinating data creation and sharing, preparing grants and projects, including the publication of specific types of research outputs in the field of digital humanities.
- Outreach in general as well as seeking recognition of specific types of digital humanities outputs in the context of scientific assessment.

Members Projects Courses Tools and Links

Structure of CDH

The CDH was established and is regulated by OD 28/2022 "Statutes of the Center for Digital



- Ins
- Fa
- Ma
- Un
- Lib

á jazyková příručka”) – 40 mil. acc/yr
formatics
)
for Digital Humanities at FF UK
manuscripts)
tional Library, other universities
ni.cz/en (Digitalia MUNI ARTS | PHIL)
plied Sciences
ry, Library of CAS)
agement system)
. pages, 300,000+ books)

LINDAT/CLARIAH-CZ partner services (1)

- Institute of the Czech Language
 - Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
 - In cooperation with Masaryk Univ., Faculty of Informatics
- Faculty of Arts, Charles University (FF UK)
 - LINDAT team participates at building the Center for Digital Humanities at FF UK
 - Corpora (diachronic, educational), Exhibitions (manuscripts)
 - Didaktikon project – educational project with National Library, other universities
- Masaryk University Brno – Faculty of Arts
 - Own digital repository at <https://digitalia.phil.muni.cz/en> (Digitalia MUNI ARTS | PHIL)
- University of West Bohemia, Faculty of Applied Sciences
 - Technology – speech recognition services
- Libraries (National Library, Moravian Library, Library of CAS)
 - Upgrade to Kramerius 7 (underlying library management system)
 - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)

Digitalia MUNI ARTS – trustworthy infrastructure for FF MU digital platforms.



Digitalia MUNI Arts is a research infrastructure whose main purpose is to support digital research in the humanities and arts at the Faculty of Arts of Masaryk University (FF MU). The aim is to make available and preserve research data and research results of academics working at FF MU in the long term, to support research by making available tools and preparing a virtual research environment that enables the development and strengthening of scientific collaboration and re-using of research data.

[DIGITALIA MUNI ARTS](#)
[DIGITAL HUMANITIES & FF MU](#)
[DARIAH-EU](#)
[LINDAT/CLARIAH-CZ](#)

- Libraries (National Library, Moravian Library, Library of CAS)
 - Upgrade to Kramerius 7 (underlying library management system)
 - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)

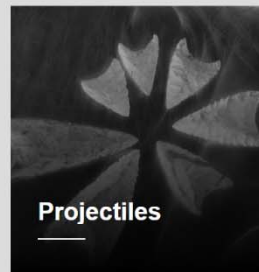
Involved platforms

We present a list of individual platforms involved in the Digitalia MUNI ARTS infrastructure. By joining the infrastructure, long-term protection and quality care of the contained data is ensured. Links to other faculty platforms can be found in a [separate catalogue](#).



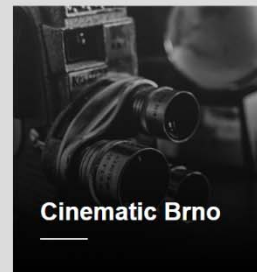
**Digital Library of
Arne Novák**

The aim of the project is to make the extensive work of Arne Novák accessible. The collection includes digitized materials from the Central Library of the FF MU.



Projectiles

The factographic database provides access to archaeological data and visualisations of Early Bronze Age arrowheads from Moravian and Slovak sites.



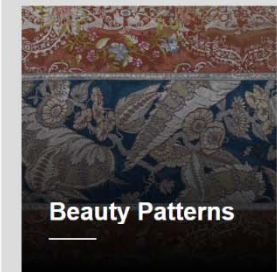
Cinematic Brno

The database contains information about individual cinemas, their programmes and films shown in Brno in 1918–1945.



**Digital Library of
the FF MU**

The digital library contains freely available full texts of journals, monographs and proceedings published at the Faculty of Arts of Masaryk University.





Beauty Patterns

The image database offers several hundred samples of liturgical vestments and other ceremonial fabrics, mostly from the 18th century.


LINDAT/CLARIAH-CZ partner services (1)

- Institute of the Czech Language
 - Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
 - In cooperation with Masaryk Univ., Faculty of Informatics
- Faculty of Arts, Charles University (FF UK)
 - LINDAT team participates at building the Center for Digital Humanities at FF UK
 - Corpora (diachronic, educational), Exhibitions (manuscripts)
 - Didaktikon project – educational project with National Library, other universities
- Masaryk University Brno – Faculty of Arts
 - Own digital repository at <https://digitalia.phil.muni.cz/en> (Digitalia MUNI ARTS | PHIL)
- University of West Bohemia, Faculty of Applied Sciences
 - Technology – speech recognition services
- Libraries (National Library, Moravian Library, Library of CAS)
 - Upgrade to Kramerius 7 (underlying library management system)
 - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)


UWebASR
UNIVERSITY OF WEST BOHEMIA -
Select speech model: Czech ▾ ●



Recognize from File



HTTP API documentation

UWebASR HTTP API

UWebASR HTTP API is a simple interface to speech recognition engine. The input data can be passed directly within the HTTP request (POST method) or as a link to a file in the form of a URL (GET method). Live audio stream recognition from a given URL is supported, as well. The output format includes plain text, machine-readable XML and JSON formats, and the WebVTT format for web captions. Recognition results (except TRS format) are streamed continuously.

For the UWebASR use the following values of variables:

- app_id - use one of the following values to use the models provided by the UWebASR service:
 - malach/en - English Wav2Vec 2.0 model
 - malach/de - German Wav2Vec 2.0 model
 - malach/cs - Czech Wav2Vec 2.0 model
 - malach/sk - Slovak Wav2Vec 2.0 model

In other words, the full HTTP API endpoint for Czech is `https://uwebasr.zcu.cz/api/v2/lindat/malach/cs`.

The UWebASR HTTP API uses the underlying SpeechCloud platform. The platform has an architecture employing a set of real-time workers. Therefore it is possible that all of the workers are allocated and no further requests could be processed at the moment. This situation is indicated by the 503 HTTP status.

Each session employing the worker is limited to at most 3600 seconds. Longer sessions are automatically terminated. Since the speech recognizer runs faster than realtime in most situations, it means that the maximum length of the audio processed in the API is actually longer than 3600 seconds.

LINDAT/CLARIAH-CZ partner services (1)

- Institute of the Czech Language
 - Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
 - In cooperation with Masaryk Univ., Faculty of Informatics
- Faculty of Arts, Charles University (FF UK)
 - LINDAT team participates at building the Center for Digital Humanities at FF UK
 - Corpora (diachronic, educational), Exhibitions (manuscripts)
 - Didaktikon project – educational project with National Library, other universities
- Masaryk University Brno – Faculty of Arts
 - Own digital repository at <https://digitalia.phil.muni.cz/en> (Digitalia MUNI ARTS | PHIL)
- University of West Bohemia, Faculty of Applied Sciences
 - Technology – speech recognition services
- Libraries (National Library, Moravian Library, Library of CAS)
 - Upgrade to Kramerius 7 (underlying library management system)
 - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)

The screenshot shows a search results page for 'Karel Čapek' on the LINDAT/CLARIAH-CZ platform. The search bar at the top contains 'Karel Čapek' and the results are ordered by relevance. The left sidebar shows active filters for 'Karel Čapek' and various availability and licence options. The main content area displays a grid of search results, including portraits, book covers, and document thumbnails, each with a red label indicating its type (e.g., 'Graphic', 'Book').

– Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)

LINDAT/CLARIAH-CZ partner services (2)

- Institute of Philosophy CAS
 - MEMORI project – early middle ages, Czech Medieval Sources Online
 - Jan Patočka Archive – Czech philosopher, dissident 1968-89
- Institute of History CAS
 - Czech historical bibliography
- National Gallery
 - Complete overhaul of digital collection hardware and software environment
- National Film Archive
 - Uses LINDAT repository for all open film documents (historical so far)
 - Direct play from repository
- EHRI group
 - Starting to integrate with LINDAT and Center for Visual History Malach
 - Documentation of documents digitized earlier (pre-LINDAT) (NA, ITI, PT)
 - EHRI CZ presentation and organization
 - Preparation of CLARIN-EHRI workshop, Prague, March 27-28, 2024



LINDAT/CLARIAH-CZ Repository Home / View Item

Mobilisation and Demobilisation 1938

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

Mobilisation and Demobilisation 1938, 1938, Národní filmový archiv, 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/20.500.12801/3901359-02>

Share: [f](#) [t](#)

Authors (:unav) Unknown author NFA

Item identifier <http://hdl.handle.net/20.500.12801/3901359-02>

Date issued 1938

Type clip, video

Language(s) Nolinguistic content

Description Rough footage capturing both the mobilisation and the subsequent demobilisation of the Czechoslovak Army in 1938. The first part was shot on 23 September after the declaration of general mobilisation, and shows civilians as they are issued military uniforms. The second part shows the demobilisation authorised at the extraordinary meeting of the Ministerial Council on 6 October. Soldiers are shown in the barracks handing in their uniforms. An officer passes out certificates of praise for military service.

Navigation: Browse, All of the Repository, My Account, Login, Statistics, General Information, Deposit, Cite, Submission Lifecycle, FAQ, About, Help Desk

- Starting to integrate with LINDAT and CVHM
- Documentation of documents digitized earlier (pre-LINDAT) (NA, ITI, PT)
- EHRI CZ presentation and organization
- Preparation of CLARIN-EHRI workshop, Prague, March 27-28, 2024

The screenshot shows the EHRI website with a navigation menu (ABOUT EHRI, NEWS & PRESS, TRAINING AND EVENTS, KRISTEL FELLOWSHIPS, ONLINE SERVICES) and a search bar. The main content area features a 'Home' link and a prominent article titled 'Call for Applications CLARIN-EHRI Workshop | "Natural Language Processing Meets Holocaust Archives"'. The article is dated Monday, 4 December, 2023, and is for a 'Hands-on Workshop CLARIN & EHRI | Call for applicants' held from March 27-28, 2024, in Prague, Czech Republic, with a deadline of 15 January 2024. A sidebar on the right lists various resources such as the EHRI Portal, Digital Tools Guides, Document Blog, Geospatial Repository, Online Course, Online Editions, Podcast, and Publication Repository.

- EHRI CZ presentation and organization
- Preparation of CLARIN-EHRI workshop, Prague, March 27-28, 2024

Internal & User Publications in 2023

- Research Infrastructure made for (research) users
- LINDAT/CLARIAH-CZ
 - Distributed, virtual, open
- RI-Internal publications
 - About building the RI
 - Hardware, repository software, organizational matters, legal, networking (ELE)
 - Creating datasets, software, services
 - Papers typically at LREC, ACL, ...
- User publications (NB: hard to find...) – using Zotero to collect
- 2023: [incomplete yet]
 - RI-Internal publications: 47
 - User / LINDAT partners: 339
 - User / rest of the world: 725

Educational support

- CLARIN K-Centre for treebanking (with CLARINO Norway)
- **Didaktikon** (UK & City of Prague, @Campus Hybernská)
 - Educational activities for primary and secondary school level
- Courses
 - Regular (FF UK, FF MU)
 - E.g., Digital Humanities (fall 2023) at FF MU,
 - Continued education (FF UK)
 - Specialized open seminars
 - **CVHM MFF UK** (Holocaust topics) for teachers
 - Support for teaching in general (tools, data, services/UI)
- Students
 - Employed as data curators, annotators (universities)
 - Technological support (**HW cluster**, software, data)

Hybernská campus | Charles University | UK Forum | Contacts

Search term Search

[DIDAKTIKON](#) •
 [EXHIBITS](#) •
 [SCHOOLS](#) •
 [RINGS](#) •
 [EXHIBITIONS](#) •
 [EVENT CALENDAR](#)

Přihlaste se k odběru našeho newsletteru

DIDAKTIKON - A NEW DIRECTION IN THE POPULARIZATION OF SCIENCE

Didaktikon is a popularization and educational center of Charles University, which was established in cooperation with the Hybernská Campus, or the capital Prague. This is a place intended for everyone who is interested in an informal approach to education.

School excursion
 At Didaktikon, we mainly focus on [excursions for primary and secondary schools](#) . In the unique [exhibitions](#) , which were created in cooperation with our faculties, you will not only learn a lot of interesting information, but you can also try everything right away. In Didaktikon, we work with the so-called heuristic method, i.e. the method of guided discovery. In this way, students will find out in practice how abstract disciplines such as philosophy and mathematics can be understood for someone. At the same time, teachers and their students can choose from a wide range of activities that Didaktikon offers. Popularization lectures and workshops are held here, which represent a wide range of fields that can be studied at Charles University.

Afternoon courses
 New this school year are [the afternoon courses](#) , which expand the scope of regular studies for those interested. The courses are intended for [primary and secondary school students who, under the guidance of Charles University teachers and students, can become scientists in various fields](#).

rnská)
level

Centrum vizuální historie Malach
...

Page · Education

Malostranské náměstí 25, Prague, Czech Republic

221 914 391

malach@ufal.mff.cuni.cz

ufal.mff.cuni.cz/malach

Open now


Not yet rated (1 review)

Centrum vizuální historie Malach
...


28 November 2023 · 🌐

Tento týden jsme zahájili seminářem pro [CET Academic Programs in Prague](#). V rámci semináře se [Jirka Kocian](#) a Tereza Juhászová věnovali tématu návratů a hledání domova po skončení druhé světové války.

K tématu doporučujeme knihu "Návraty: Poválečná rekonstrukce židovských komunit v zemích středovýchodní, jihovýchodní a východní Evropy", kterou editovaly Kateřina Králová a Hana Kubátová. Kniha vychází z rozhovorů uložených v Archivu vizuální historie USC Shoah Foundation.... [See more](#)



Photos [See All Photos](#)





UFAL AIC Search UFAL AIC

Main Page

[Main page](#) [Discussion](#) [View](#) [View source](#) [History](#)

CZ.02.2.69/0.0/0.0/17_044/0008562
Podpora rozvoje studijního prostředí na Univerzitě Karlově - VRR

 EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

 MINISTRY OF EDUCATION,
YOUTH AND SPORTS

[Contents](#) [hide](#)

- 1 [Welcome to AIC](#)
 - 1.1 [Access](#)
 - 1.2 [Jupyterlab](#)
 - 1.3 [Connecting to the Cluster \(directly\)](#)
 - 1.4 [Basic HOWTO](#)

Welcome to AIC

AIC (Artificial Intelligence Cluster) is a computational grid with sufficient computational capacity for research in the field of [deep learning](#) using both CPU and GPU. It was built on top of [SLURM](#) scheduling system. MFF students of Bc. and Mgr. degrees can use it to run their experiments and learn the proper ways of grid computing in the process.



ká)

– Technological support (HW cluster, software, data)

Specific datasets: Shared Tasks Support

- Supporting Shared Tasks
 - by data, manpower, participation
 - “Shared Task” = worldwide open competition (~kaggle etc.)
 - on a defined task
 - data and metrics provided by organizers
- CoNLL Shared Task on dependency parsing (2017, 2018)
 - Using the Universal Dependencies collection, additional data
- CoNLL Shared Task on Meaning Representation Parsing (MRP)
 - Run in 2019 and 2020, Stephan Oepen & colleagues, co-organization
- **CRAC Shared Task on Multilingual Coreference Resolution**
 - 2022, 2023 co-organized, datasets in repository (EMNLP 2023)
- WMT Shared Tasks and Subtasks
 - Data/repository support for WMT

LIN

The screenshot shows the CorefUD website with the following content:

- Navigation menu: ABOUT, EVENTS, RESEARCH (selected), PEOPLE, TEACHING, JOBS*, WIKI, PAKT, ICCL. Search bar is present.
- Breadcrumbs: COREFUD > CRAC22 > CRAC23. A "BACK TO PROJECTS" link is also visible.
- Section Header: CRAC 2023 Shared Task on Multilingual Coreference Resolution
- Section Header: Overview
- Text: Coreference resolution is the task of clustering together multiple mentions of the same entity appearing in a textual document (e.g. *Joe Biden, the U.S. President and he*). This CodaLab-powered shared task deals with multilingual coreference resolution and is associated with the [CRAC 2023 Workshop](#) (the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference) held at [EMNLP 2023](#).
- Section Header: Official Results
- Text: The following table shows four versions of the CoNLL metric macro-averaged over all datasets:
 - head-match excluding singletons (the primary metric, see [below](#)),
 - partial-match excluding singletons,
 - exact-match excluding singletons and
 - head-match with singletons.
- Text: A more detailed evaluation will be provided in the shared task overview paper.
- Table of Official Results:

system	head-match	partial-match	exact-match	with singletons
1. CorPipe	74.90	73.33	71.46	76.82
2. Anonymous	70.41	69.23	67.09	73.20
3. Ondfa	69.19	68.93	53.01	68.37
4. McGill	65.43	64.56	63.13	68.23
5. DeepBlueAI	62.29	61.32	59.95	54.51
6. DFKI-Adapt	61.86	60.83	59.18	53.94
7. Morfbase	59.53	58.49	56.89	52.07
8. BASELINE	56.96	56.28	54.75	49.32
9. DFKI-MPrompt	53.76	51.62	50.42	46.83



(P)
ation
olution

LINDAT/CLARIAH-CZ international cooperation



- International cooperation (support or development)
 - ELG European Language Grid, EU H2020 Call 29a
 - Automatic metadata harvesting continues
- SSHOC, Infrastructural H2020 project, followup
 - Part of CLARIN ERIC participation, EOSC Marketplace
- Humane-AI-Net: AI Center of Excellence, Call 48
 - Microprojects w/DFKI, INRIA, etc.
- ATRIUM 2024-2028, Horizon Europe INFRA
 - Archaeology, MT and object recognition
 - In cooperation with Czech RI: Archaeological Information System
- Everse 2024-2028, Horizon Europe INFRA
 - Long/term preservation of research software (SSH), w/CLARIN
- Support for other Horizon Europe projects
 - RESQ+ (IR), MEMORISE (Holocaust access)
 - **HPLT (LLM data collection, LLM/TM building, 2022-2025)**





LI
inter

- International cooperation
 - ELG European Language Grid
 - Automatic metadata harvesting
- SSHOC, Infrastructure for the Social Sciences and Humanities
 - Part of CLARIN ERIC project
- Humane-AI-Net: AI Centre for the Humanities
 - Microprojects w/Digital Humanities
- ATRIUM 2024-2028
 - Archaeology, MT and Digital Humanities
 - In cooperation with the European Commission
- Everse 2024-2028,
 - Long/term preservation of digital research data
- Support for other HPC centres
 - RESQ+ (IR), MEMOIR
 - HPLT (LLM data)

H P L T HIGH PERFORMANCE LANGUAGE TECHNOLOGIES

A space that combines petabytes of natural language data with large-scale model training

- Lots of monolingual and multilingual data consistently formatted and curated
- Efficient and high-quality language and translation models
- Sustainable and reusable workflows using high-performance computing

More about HPLT

<https://hplt-project.org/> @hplt_eu

Funded by:

WELCOME TO BOARDS
We will help users find what is in language data and models, how they compare to others, and how they were built through interactive boards.

Our Focus
We will retool how language data is generated, shared, and transformed into efficient large language and translation models making HPC centres ready to large scale NLP across Europe.

- 7** petabytes of web data from the internet archive
- 5** petabytes of web data from commoncrawl
- 2.5** trillion words of monolingual text
- ~300** unique corpora
- ~80** languages to cover
- 100s** of efficient language and translation models
- 36** months to complete the project
- 8** consortium partners collaborating together

Our Partners

CHARLES UNIVERSITY UNIVERSITY OF OSLO UNIVERSITY OF EDINBURGH UNIVERSITY OF TURKU

UNIVERSITY OF HELSINKI PROMPSIT L.E. CESNET SIGMA2



EUROPEAN LANGUAGE GRID

SSHOC
social sciences & humanities open cloud



LINDAT/CLARIAH-CZ EOSC CZ

- Czech European Open Science Cloud Association and Secretariat (EOSC CZ)
 - National support for central tasks and organization
- EOSC CZ Open Science I (2024-2028)
 - Project: National Repository Platform
 - Lead by Czech eInfra “CESNET”
 - LINDAT/CLARIAH-CZ significant partner (0,8 mil. EUR)
 - CLARIN-DSpace chosen as one of four repositories recommended nationwide
 - Funds repository(ies) development, upgrades, adaptation for national use
- EOSC CZ Open Science II (2025-2028)
 - Project led by Charles University (HQ)
 - LINDAT/CLARIAH-CZ will lead SSH workpackage (4 mil. EUR total), 1 of 8
 - Czech SSH RIs as partners, FSTP-like funding for others

LINDAT/CLARIAH-CZ beyond 2023

- Future continuation
 - Contract (with cuts) until 2026 (originally: 1,8x more funding, -2029)
 - LINDAT/CLARIAH-CZ on Czech Roadmap
 - Czech Republic joins EHRI ERIC in early 2025
 - Durable equipment (hardware, computing) grant under review
 - 2024-2026, 2 mil. EUR (shared by 7 partners of the RI)
- Evaluation
 - International panel + quantitative scientometrics at national level
 - Crucial for funding 2027-???? (legal changes – special status?)
 - Starts late 2024
 - IAB may play a role
- Integration with / support to EOSC CZ



Thank you!

<https://lindat.cz>

