



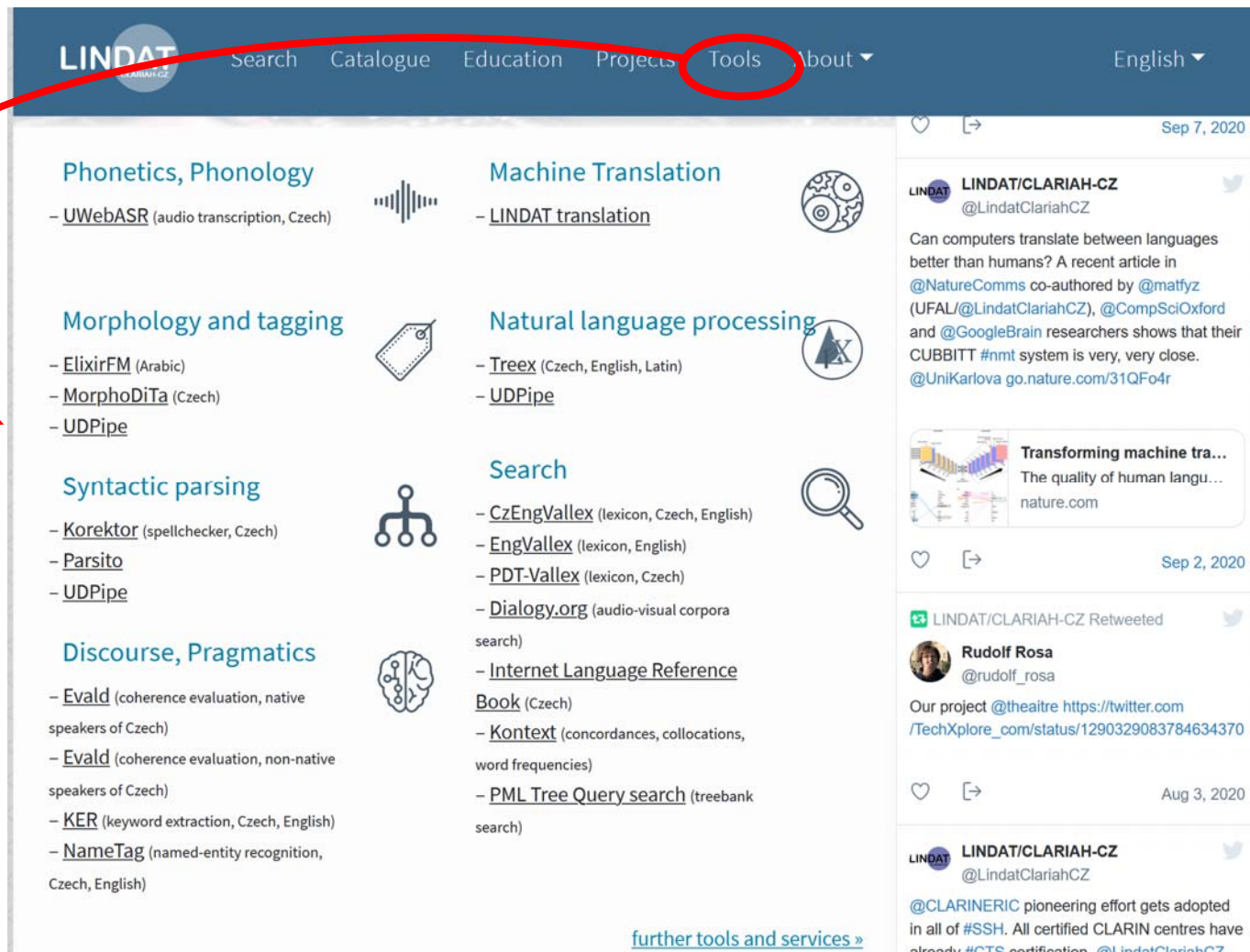
Institute of Formal and Applied
Linguistics (ÚFAL)
Projects Related to
LINDAT/CLARIAH-CZ

Jan Hajič

- LINDAT = CLARIN and DARIAH in Czechia
 - European networks for supporting research (in SSH)
 - CLARIN: Language resources and technology
 - DARIAH: Digital humanities and arts
 - LINDAT/CLARIAH-CZ: 2019-2022
 - Combines membership in EU CLARIN and DARIAH networks
 - 11 institutions in the Czech combined network
 - UK, MU, Academy of Sciences, National and Moravian Libraries, NG, NFA
 - Center for Visual History Malach: under LINDAT (& Herzl Center)
 - <http://lindat.cz> (redirected to <https://ufal.mff.cuni.cz/lindat>)
 - Provides also large part of computing infrastructure of UFAL
 - New hardware due in 2021 (OP VVV VI 2)
 - Lots of GPUs 😊

LINDAT/CLARIAH-CZ in 2020/1

- New web at <https://lindat.cz>



The screenshot shows the LINDAT website interface. The navigation bar at the top includes 'LINDAT', 'Search', 'Catalogue', 'Education', 'Projects', 'Tools', and 'About'. The 'Tools' menu item is circled in red. A red arrow points from this menu item to the 'Morphology and tagging' section of the main content area.

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

- [LINDAT translation](#)

Natural language processing

- [Treeex](#) (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)
- [PML Tree Query search](#) (treebank search)

[further tools and services »](#)

Twitter Feed:

- LINDAT/CLARIAH-CZ** @LindatClariahCZ (Sep 7, 2020): Can computers translate between languages better than humans? A recent article in @NatureComms co-authored by @matfyz (UFAL/@LindatClariahCZ), @CompSciOxford and @GoogleBrain researchers shows that their CUBBITT #nmt system is very, very close. @UniKarlova go.nature.com/31QFo4r
- LINDAT/CLARIAH-CZ Retweeted** (Sep 2, 2020): **Rudolf Rosa** @rudolf_rosa: Our project @theaitre https://twitter.com/TechXplore_com/status/1290329083784634370
- LINDAT/CLARIAH-CZ** @LindatClariahCZ (Aug 3, 2020): @CLARINERIC pioneering effort gets adopted in all of #SSH. All certified CLARIN centres have already #CTS certification. @LindatClariahCZ

Sep. 25, 2020

- “Internal” projects – offer datasets, annotation, curation, tools, services
- Repository development (CLARIN DSpace), HW and other IT support
- Universal Dependencies project
 - General support, official editions
 - Tools: UDPipe (now UDPipe 2), soon to have: NER
- Prague Dependency Treebank – Consolidated edition
 - Approx. 4MW, manual morphology
 - PDT, PCEDT, PDTSC, Faust; TR annotation same, a-layer: auto (PDT: manual), new morph. dictionary
- SynSemClass
 - Multilingual (cs, en so far) event type ontology
 - A.k.a synonym dictionary of verbs, with valency, linked to corpora and other lexical resources
- TEI:TOK
 - Platform for rich annotation and multimedia content
- Support for other projects (GAČR, TAČR, H2020)
- Running services (<https://lindat.cz/en/services>), maintaining HW and basic SW

- International cooperation (support or development)
 - ELG European Language Grid, EU H2020 Call 29a
 - „Marketplace“ for mainly commercial use of Language Technologies
 - Partners: **DFKI**, UK, Sheffield, ELRA, commercial partners
 - Pilot projects: organized by us (2M EUR), now 2nd Call (Oct 1)
 - Some UFAL tools provided – UDPipe, MT; harvesting repository
- SSHOC, Infrastructural H2020 project
 - Part of CLARIN ERIC participation
 - UFAL/LINDAT provides basic language tools for IR, MT for Social Surveys (COVID)
- Humaine-AI-Net: AI Center of Excellence, Call 48
 - Small participation, w/DFKI Saarbrucken: microproject (anyone...?)
- ELG and SSHOC 2019-2022, Humaine-AI-Net 2020-2023
- Two more submitted in 2020: ELG continuation, LANGEQ

Shared Tasks

- Supporting Shared Tasks by data, manpower, participation
- CoNLL Shared Task on Meaning Representation Parsing (MRP)
 - Run in 2019 (and 2020, Cz added)
 - Stephan Oepen & colleagues
 - Co-organization
- Data/repo support for WMT

The Future

- Evaluation 2021 (!!!)
 - We have to pass w/ either of the two highest marks, otherwise LINDAT will not get funding 2023- (same for equipment)
- Internal LINDAT members:
 - your main work for October-November/December 2020!
- Others:
 - we might ask for help
- 80 page questionnaire, with partners, attachments, indicators, budget, ...
 - Deadline: Dec. 10, 2020; interview in front of panel: May/June 2021; verdict: July 2021, full proposal for 2023-2026(9?) due in early 2022
- Extended to EHRI (Holocaust research infrastructure)
 - Natural partner for CVHM, lead: Masarykův ústav AV ČR
- Main points: SoA and new tools, equipment renewal 23-29, new resources, annotation, support to Digital Humanities, new services at partners, international cooperation (support for more projects)