

UFAL's projects (and longer-term financial/grant outlook)

JAN HAJIČ

ÚFAL MFF UK

Ongoing infra projects (1)

LINDAT/CLARIAH-CZ 2023-2026 (PI: J. Hajič)

- 17 partners (Universities, Academy of Sciences, National cultural institutions)
- <https://lindat.cz>
- Budget ~2,6 M EUR yearly, total all 17 partners
 - UFAL salaries: ~800k EUR/year
- Separate project for equipment (“OP JAK”)
 - Key for compute at ÚFAL as a whole
 - 2024-6: total of ~700k EUR in GPUs, network, data storage

Related EU projects with LINDAT/CLARIAH-CZ participation (PI: P. Straňák)

- Atrium (digitalisation of archaeological resources) (PI: P. Straňák)
- Everse (proper repository specs for research software) (PI: P. Straňák)
- FIDELIS (HPC and data resources) (PI: P. Straňák)
- CLS (with CNC; finished 2025, PI: S. Cinková)

Ongoing infra projects (2)

EOSC CZ NRP 2023-2028, OP JAK (MŠMT)

- European Open Source Cloud, CZ version (ÚFAL PI: J. Hajič)
 - National Repository Platform (Open Science I call)
- ÚFAL (LINDAT/CLARIAH-CZ) will supply Dspace 7.x adapted for data (as in LINDAT/CLARIAH-CZ itself)
- Hiring ongoing (1 1/2 developers for the repository software still open)

EOSC CZ Open Science II, 2025-2028, OP JAK (MŠMT)

- Just confirmed, accepted for funding (ÚFAL PI: J. Hajič)
- Actual funding starts 2026
- Most ÚFAL's positions start in 2027
 - Data upgrade: PDT-C, SynSemClass, UMR, ...
 - Enriched services: geographical viewer(s) – with CNC



Large OP JAK MŠMT research projects

OP JAK Mezisektorová výzva (industrial cooperation in research, 2025-28)

- „Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím“ („NLP“)
- PI: O. Bojar, companies: VUT Brno, Mama AI, Phrase, Phonexia, seznam.cz
 - Conversational systems, multilinguality and translation, speech technology, ethics/bias
 - 1,2 mil. EUR (7,5 mil. CZK/year)

OP JAK Excellent Science in the Social Science and Humanities (2025-2028)

- „HumanAID“, coordinated by FF/FSV UK (ÚFAL PI: B. Hladká)
 - ÚFAL role: WP leader for supplying finetuned LMs for various areas (including use of AI in education)
- About 3 FTEs

... talk to their PIs 😊!

CEDMO II (NPO, MPO), 2024-2026 (20M CZK), PI: ČVUT

- part of CEDMO series of projects (V. Moravec, B. Hladká)

TAČR(s) – applied research projects

- CZDEMOS4AI: (AI/LLM) avatars for adictology, law, history education, media (PI: R. Rosa)

MŠMT International cooperation (Inter-Action)

- UMR (PI: J. Hajič), building Czech, Latin UMR
- AIAI (R. Rosa): AI, authorship and interpretation (2025-2027)
- Improving stomach examinations with Artificial Intelligence: A deep learning approach for assisted gastroscopy (P. Pecina) (2024-2026) (w/Bavaria)

UK – UNCE: Multilinguality (2024-2029)

- Z- Žabokrtský (with V. Cvrček, FF UK, FF coordinating)

GAČR (5 new projects in 2025): PIs: V. Kolářová, J. Libovický, D. Macháček, R. Rosa, K. Bendová)

GAUK (6 new projects in 2025): PIs: V. Lanz, D. Taji, M. Nassajian, P. Schmidtová, N. Kumar, T. Polák

EU projects started in 2025

OpenEuroLLM – Digital Europe (EC, 2025-2027, PI: J. Hajič)

- Collect data, build 1 multilingual (36+/42 lang) LLM (+ variants), competitive
- 20 partners (academia, industry, research organizations, open-source community, HPCs)
- 34 mil. EUR (ÚFAL: 0,8 mil. EUR/year ~21 mil. CZK/year)
- Small reference models published at huggingface
 - look for OpenEuroLLM „organization“
- Hiring 😊!! (ask me – possibly up to 3-4 FTEs):
 - we have finally got the cofunding money from MŠMT 😊

ALT-EDIC4EU – Digital Europe (EC, 2025-2028, ÚFAL PI: J. Hajič)

- Coordinating Action, small ÚFAL participation (to link to OpenEuroLLM)
- ALT-EDIC: Alliance for Language Technologies European Data Infrastructure Consortium
 - As an ERIC (CLARIN, DARIAH, EHRI), but for Data; Czech Republic is a full member since 2024

PressMint - Interoperable corpora of historical newspapers

- Funded by CLARIN ERIC (PI: B. Hladká, 2025-2027)

Projects ended(-ing)

HPLT (2022-2025), Horizon Europe

- Acquire large data, build (encoder) LLMs / TMs (OpusMT)
- HPLT v3.0 published (not yet advertised... still checking for possible errors)
 - Texts extracted, cleaned, filtered, re-lang-Oded, deduplicated
 - with exceptions: en, ru, zh
 - <https://hplt-project.org/datasets/v3.0>
 - Or at huggingface
 - 198 languages, 50TB compressed (only clean(est) version published)
 - Czech: Docs: 107.80M, Tokens: 126.25B, Chars: 367.84B, Segments: 2.47B
 - English: Docs: 18.06B, Tokens: 16.28T, Chars: 72.34T, Segments: 435.23B
- Parallel data, tools, dashboard at OpusMT (<https://github.com/Helsinki-NLP/OPUS>)
- BERT-type models



? Projects under review

-
- ? ERC projects submitted / to be submitted (3)
 - ? GAČR (announces early December)
 - Several standard grants submitted
 - ? TAČR (500M Kč) Sigma DC5 for AI domain (MFF budget: 31 mil. Kč / 6 years)
 - ČVUT coordination, ÚFAL PI: O. Bojar/J. Hajič (2026-2031)
 - Multimodal LLMs (VLMs)
 - Many AI teams across Czechia
 - ? Horizon Europe (results: March-April 2025)
 - 3x (ÚFAL PIs: Hajič, Dušek, Popel) - HORIZON-CL4 calls ? ?
 - Multimodal LLMs, Explainable and robust AI
 - MT for application domains

General funding outlook

Not very positive

- National budget under stress (only 0.5% GDP to research – vs. EU avg. 0.75%):
 - Large deficit, must increase defense spending, other priorities than research
- AI currently a „buzzword“ – positive, but still no specific programs (except TAČR Sigma)

2026(7,8) budget (GAČR):

Specifické ukazatele rozpočtových kapitol v letech 2026 až 2028

Kapitola: 321 Grantová agentura České republiky

v Kč			
Specifické ukazatele - výdaje	2026	2027	2028
Výdaje Kanceláře Grantové agentury ČR	164 841 581	169 841 581	174 841 581
Dotace jiným subjektům	4 735 158 419	4 430 158 419	4 425 158 419
Výdaje celkem	4 900 000 000	4 600 000 000	4 600 000 000

Increase in MPO (Industry), MO (Defense)

- MPO seems to be now willing to discuss specific AI support

Private support? (cf. Bulgaria, Romania)

ÚFAL funding outlook

Large projects (cf. also ÚFAL director's presentation earlier)

- LINDAT/CLARIAH-CZ
 - Funding until 2026, then unclear (evaluation – condition for renewal – has not started yet)
- OP JAK projects ALL finish at the end of 2028
- TAČR projects
 - Only the one under review (soon) to go over 2029 (-2031), but not that large budget (5M Kč/year)
- GAČR all end 2027, the new ones (if accepted) in 2028
- Institutional funding: only small nominal increases expected

Project funding sources:

- OP JAK (or similar funding) after 2027 unclear, most probably 4x-5x smaller(!)
- VVI (i.e., for LINDAT) funding constant -> effectively 30% decrease
- GAČR, TAČR – continues, but see previous slide
- EU projects: end 2025, 2027, 2028 (early), even the new ones will end end of '28

... 2029 will bring a substantial loss of funding

- We need to keep watching for all possible (incl. alternative) sources