Charles University, Prague

Visiting fellowship contact
Michal Křen
michal.kren@ff.cuni.cz
Charles University, Faculty of Arts
Institute of the Czech
National Corpus
https://ucnk.ff.cuni.cz/en/

Founded in 1348, Charles University (CUNI) is the largest and oldest Czech university. Nowadays, it stands for almost one quarter of the scientific performance among the Czech universities. In CLS INFRA, CUNI is represented by two tightly cooperating departments: Institute of Formal and Applied Linguistics (UFAL) at the Faculty of Mathematics and Physics, and the Institute of the Czech National Corpus (ICNC) at the Faculty of Arts.

UFAL is specialized in computational linguistics research from formal theoretical foundations to application development. Its staff are involved in many international projects (e.g. Universal Dependencies) creating many language resources and NLP tools.

The mission of ICNC is continuous mapping of (Czech) language by building large general-purpose reference corpora that result from a broad-scale data collection. This is supplemented by the development of specialized web applications that enable user-friendly work with the language data available from the CNC research portal.

Existing tools, services and expertise

UFAL hosts the LINDAT/CLARIAH-CZ research infrastructure, a Czech national node of both CLARIN and DARIAH. Its main aim is to allow for an open access to digitized data resources of many SSH disciplines for broad research community.

Main project-related services offered by UFAL:

- automatic text annotation (tagging, parsing, named entity recognition for many languages);
- setup and supervision of (semi-)manual text annotations to enable/improve automatic text annotation for new languages or text domains.

ICNC hosts the CNC national research infrastructure. Main project-related services offered by ICNC:

- corpus compilation: systematic collection, processing and annotation of large quantities of language data to produce language corpora (written, spoken, parallel);
- application development: design and development of user applications aimed to promote empirical methods in linguistic research;
- user support: CLARIN K-centre helpdesk, comprehensive web documentation, corpus hosting, data packages on demand, consulting, education, training etc.

Aim of the fellowship

The CLS Fellowship at CUNI is directed towards: a) humanities researchers who wish to design, create, annotate and exploit their own corpora, b) linguists who want to learn how to analyze the corpus data, c) programmers who would like to participate on the design and development of corpus analysis tools.

The fellowship can be done either at UFAL or ICNC. At ICNC, the fellows will primarily be assisted in a) methodology and technology of corpus compilation, including corpus composition issues; b) quantitative statistical analysis of corpus data and text mining.

At UFAL, the fellows will primarily get assistance with a) linguistic annotation by the UDPipe and NameTag2 tools and their extension on lesser-resourced languages or project-specific entities; b) creation of gold standard data sets, most of all conforming to the Universal Dependencies formalism; c) harvesting text snippets from OCR-ed resources of major libraries.

Fellows with projects with well-defined realistic (albeit modest) outcomes are especially encouraged to apply. As a part of the fellowship, the fellows will be assigned a local mentor with a best match with their project. Therefore, it is recommended that the fellows contact the CUNI contact person before submitting the application.

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement: 101004984