The primary objective of our Institute is interdisciplinary research and education in Computational Linguistics and Natural Language Processing (NLP). To approach our research topics, we combine knowledge from linguistics, mathematics, computer science, statistical modeling, and software engineering, while fostering international cooperation along with direct participation of graduate students in all these areas. As leaders in several of these fields, we actively develop language resources for Czech and other languages of interest as well as host LINDAT/CLARIN, the Czech node of the Clarin ERIC European language resource infrastructure.

Theoretical aspects as well as application of NLP tools and techniques to diverse language-related tasks, including but not limited to multilingual text analytics, machine translation, dialog systems and multilingual, crosslingual and multimodal information retrieval and information extraction.

Development of state-of-the-art methods, such as statistical modeling, Deep Machine Learning, or Big Data processing.

Hosting the Center for Visual History Malach, an access point to the University of Southern California’s Shoa Foundation archive.

Development of tools and expertise in the area of Digital Humanities.
RESEARCH AREAS & EXCELLENCE

TEXT ANALYTICS
- Tools for basic and advanced language analysis designed for sophisticated text processing and analysis: tokenizers, language identification, lemmatization, POS and morphological tagging, parsing, semantic role labeling, named entity recognition, and sentiment analysis.
- Tools work with a number of languages (30+) and can be easily customized, e.g. by adding more languages, thanks to the language resources available in the LINDAT/CLARIN infrastructure (http://lindat.mff.cuni.cz/en/services).

MACHINE TRANSLATION
- Phrase-based statistical machine translation employing machine learning techniques and very large text corpora.
- Machine translation based on deep linguistic analysis of whole sentences.
- Modern Deep Neural Network models for machine translation.

INFORMATION RETRIEVAL AND INFORMATION EXTRACTION
- Cross-lingual as well as multimodal information retrieval of text, speech, video, and images.
- Expertise gained within EU as well as nationally funded large projects in the medical and oral history domains.

KEY RESEARCH EQUIPMENT
The Institute runs its own computing cluster, the Linguistic Research Cluster, with 1600+ CPUs and 9 TB of memory; a number of GPUs for the most demanding machine learning tasks are available. In addition, the Institute has a PB-level tape robot for storing large text and multimodal datasets and archives. For commercial cooperation, the Institute routinely sets up separate “grey-zone”, firewalled and specifically protected clusters to ensure security and privacy of the data processed.

PARTNERSHIPS AND COLLABORATIONS

ACADEMIC AND RESEARCH PARTNERS
- CLARIN ERIC Infrastructure – European Research Infrastructure for Language Resources and Technology.
- University of Southern California – Library and the Visual History Institute (Formerly USC Shoa Foundation), USA.

University partners from around the World, both formally in EU- and U.S.- based projects as well as informally, above all:
- Germany – Saarland University, RWTH Aachen University, University of Tübingen;
- UK – University of Edinburgh, University of Sheffield;
- Ireland – Dublin City University;
- Spain – University Pompeu Fabra Barcelona;
- Italy – University of Pisa;
- The Netherlands – University of Groningen;
- Sweden – Uppsala University;
- Norway – University of Bergen;
- USA – University of Pennsylvania, Stanford University, University of Colorado in Boulder, Johns Hopkins University, Brandeis University, University of Southern California,
- University of Maryland;
- China – Academia Sinica Taipei, Hong Kong University of Science and Technology;
- India – IIIT Hyderabad.

INDUSTRIAL AND PUBLIC RESEARCH INSTITUTIONS:
European Language Resources Association (ELRA); Google Research, USA; IBM Research, Czech Republic and USA; DFKI Berlin/Saarbrücken, Germany; Fondazione Bruno Kessler, Italy; Linguistic Data Consortium, USA; Institute of Computer Science Polish Academy of Science (IPI PAN), Poland; ILC Pisa, Italy; National Center for Scientific Research (CNRS), France; Scotland’s national Telehealth and Telecare organisation (NHS24), Scotland; Cochrane, Germany.

National partners
University of West Bohemia, Pilsen; Masaryk University, Brno; Czech Technical University in Prague; Brno University of Technology; Technical University of Liberec; Czech Academy of Sciences (Institute of the Czech Language, Institute for Contemporary History, Institute of Philosophy), Prague; Institute for the Study of totalitarian Regimes, Prague; National Film Archive, Prague.

INDUSTRY PARTNERS
- Datlowe, Czech Republic;
- Geneea, Czech Republic;
- IBM TLS Czech Republic;
- Lingea, Brno, Czech Republic;
- ACREA CR, Czech Republic;
- Tilde, Latvia

MAIN RESEARCH PROJECTS

INTERNATIONAL PROJECTS

SEE OUR WEBPAGE
ufal.mff.cuni.cz