Institute of Formal and Applied Linguistics

 # Natural Language Processing
 # Computational Linguistics, Text Analytics
 # Machine Translantion, Deep Machine Learning
 # Big Data Processing, Information Retrieval

Offer

- 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 soft ware 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/CLARIAH-CZ, the joint Czech node of the CLARIN and DARIAH ERIC European language resource and digital humanities infrastructures.

Expertise

- Theoretical aspects as well as application of NLP tools and techniques to diverse language-related tasks within Artificial Intelligence or independently, including but not limited to multi-lingual text analytics, machine translation, dialogue systems and human-computer interaction, natural language generation and multi-lingual, cross-lingual 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: tokenization, language identification, lemmatization, POS and morphological tagging, parsing, semantic role labeling, named entity recognition, and sentiment analysis.
- Tools for basic and advanced language analysis for 200+ languages
- Easy combination, integration and customization (incl. adding more languages), thanks to the language resources, tools and services available in the LINDAT/CLARIAH-CZ infrastructure ((http://lindat.cz/services))

Machine Translation

- Top-performing Deep Neural Network models for machine translation backed by decades of experience in classical architectures of MT including hybrid systems (http://mttalks.ufal. ms.mff.cuni.cz).
- Delving into speech translation and multimodal translation in high-profile research projects (e.g. http://elitr.eu)

Information Retrieval and Information

- 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

Conversational AI

- Text-based and voice-based conversational assistants
- Systems supporting task-based scenarios and chitchat
- Deep Neural Networks for language understanding, dialogue management, and natural language generation

Key Research Equipment

The Institute runs its own computing cluster, the Linguistic Research Cluster, with 2000+ CPUs and 10 TB of memory; almost 100 large 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 a separate “grey-zone”, firewalled and specifically protected clusters to ensure security and privacy of the data processed.

Partnerships and Collaborations

Academic and Research Partners

- Germany – Saarland University, Ludwig-Maxmillian University Munich, RWTH Aachen University, University of Tübingen, University of Leipzig, Karlsruhe Institute of Technology
- UK – University of Edinburgh, University of Sheffield, Heriot- Watt University, Cambridge University
- Ireland – Dublin City University
- Spain – University Pompeu Fabra Barcelona
- Italy – University of Pisa
- Belgium – University of Leuven
- The Netherlands – University of Groningen, University of Utrecht
- Sweden – Uppsala University
- Norway – University of Oslo, University of Bergen
- USA – University of Pennsylvania, Stanford University, University of Colorado Boulder, Johns Hopkins University, Brandeis University, University of Southern California, University of Maryland, New Mexico University
- China – Hong Kong University of Science and Technology
- Taiwan – Academia Sinica Taipei
- India – IIIT Hyderabad
- Czech Republic – University of West Bohemia, Pilsen; Masaryk University, Brno; Czech Technical University in Prague; Brno University of Technology; Czech Academy of Sciences (Institute of the Czech Language, Institute for Contemporary History, Institute of History, Institute of Philosophy, the Library), Prague; Institute for the Study of totalitarian Regimes, Prague; National Film Archive, Prague; National Library Prague; Moravian Library, Brno; National Gallery, Prague.

Industrial and Public Research Institutions:

European Language Resources Association (ELRA); Google Research, USA; IBM Research, Czech Republic and USA; IBM TLS, Czech Republic; DFKI Berlin/Saarbrücken, Germany; Fondazione Bruno Kessler, Italy; Linguistic Data Consortium (LDC), 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; Mozilla Denmark APS, Germany; Geneea, Czech Republic; Memsource, Czech Republic; ACTA Association, Czech Republic; Trask solutions, Czech Republic.

Main Research Projects

International Projects

- *EU Horizon 2020 projects*: Welcome (2020–22), SSHOC (2019–22, https://sshopencloud.eu), Bergamot (2019-21, http://browser. mt), ELG (2019–21, https://www.european-language-grid. eu), Elitr (2019-21, http://elitr.eu), HimL (2015–18), Parthenos (2015–18), QT21 (2015–18), ClarinPlus (2015–17), KConnect (2015–17).
- *EU COST networks participation:* Multi3Generation (2019–21), COST Action CA16204 Distant Reading for European Literary History (2018–21), TextLink (2015–17), PARSEME (2013–17).
- *EU Erasmus Mundus*: European Masters Program in Language & Communication Technologies (2007–26).
- *EU Marie Curie ITN*: CLARA – Common Language Resources and their Applications (2009–14).
- *Previous EU frameworks projects*: QTLeap (2013–16), Khresmoi (2010–14), Faust (2010–13), EuroMatrixPlus (2009–12), EuroMatrix (2006–09), Companions (2006–10), STEEL (1997–99).
- *USA:* Mellon Foundation CLARIN-LAPPS coordination (2019–21

Are you interested in this expertise?

Please contact CPPT UK
Web: www.cppt.cuni.cz/
Mail: transfer@cuni.cz
Phone: +420 224 491 255

Experts and their department

Assoc. Prof. RNDr. Markéta Lopatková, Ph.D.
Institute of Formal and Applied Linguistics
Web: ufal.mff.cuni.cz

Klíčová slova

 # Zpracování přirozeného jazyka
 # Počítačová lingvistika, textová analýza
 # Strojový překlad, hluboké strojové učení
 # Zpracování velkých datových souborů, vyhledávání informací