CHARLES
UNIVERSITY

NLP v České republice

Jan Hajič (ÚFAL MFF UK & LINDAT/CLARIAH-CZ)

EUROPEAN
LANGUAGE
GRID

ELG NCC Cz-Sk Workshop, 18.10.2021

NLP (jazykové technologie) v ČR

- Centra, ústavy a katedry
  - Univerzita Karlova, Praha (MFF, FF)
    - Zpracování textu, nástroje, slovníky, jazykové zdroje (korpusy)
    - Dialogové systémy, vícejazyčná analýza, strojový překlad
  - Západočeská univerzita, Fakulta aplikovaných věd, Katedra kybernetiky
    - Audio vyhledávání, speech recognition, syntéza řeči
  - Vysoké učení technické Brno (mluvená řeč, dialog, multimédia, AI, strojové učení)
  - Masarykova Univerzita Brno, Fakulta informatiky
    - Lexikální zdroje; eLexis projekt
  - Technická Univerzita Liberec (mluvený jazyk)
  - ČVUT – CIIRC (Centrum robotiky)
    - Umělá inteligence, dialogové systémy
- Infrastructure: LINDAT/CLARIAH-CZ, ČNK
- Spolupráce s firmami (translation, basic analysis tools, resources)
  - Geneea, Phonexia, SpeechTek, Lexical Computing, Memsource, ..

ÚFAL MFF UK & LINDAT/CLARIAH-CZ

Ústav formální a aplikované lingvistiky MFF UK

- Jazykové technologie i teoretický výzkum v počítačové (formální) lingvistice
- Založen 1991, tradice od 30. let 20. století
- Nyní 100+ výzkumníků, z toho cca 25 PhD studentů
  - Vlastní Mgr. program pro jazykové technologie
  - PhD program „Computational Linguistics“, teoretická a aplikovaná specializace
- Bohatá mezinárodní spolupráce (EU, USA)
- Projekty H2020 a předchozích rámcových programů EU
- Národní projekty (Centra Excellence, GAČR, NAKI, …)
- Hostující ústav pro výzkumnou infrastrukturu LINDAT/CLARIAH-CZ

ÚFAL MFF UK & LINDAT/CLARIAH-CZ

- LINDAT/CLARIAH-CZ
  - Výzkumná infrastruktura pro jazykový výzkum
    - UK a partnerské instituce: MU, ZČU a ÚJČ AV ČR, NK, NGP, NFA, Knihovna AV ČR, HÚ AV ČR, FLÚ AV ČR
    - Součást evropské sítě CLARIN ERIC a DARIAH ERIC (ESFRI)
  - Vybudován repozitář pro Open Data (vč. způsobu ukládání, vyhledávání, prezentace a sdílení dat): *Clarin DSpace*
    - Certifikován v rámci sítě CLARIN v EU, používán v několika dalších zemích v EU
  - Vyvinuty a volně zpřístupněny nástroje pro analýzu jazyka (text, mluvená řeč), včetně strojového překladu
  - Ukázka: https://lindat.mff.cuni.cz/services/translation/
- Klíčová je mezinárodní spolupráce (ELG, ELE, ELRC)
  - Výzkumné projekty s partnery v EU i mimo, podíl na vývoji technologií překladu pro MT@EC, nyní mluvený překlad (aut. tlumočení)
- Aplikace
  - smluvní výzkum, partnerství i se státní správou (MV, MŠMT, EK)
  - Kreativní průmysl – projekt TheAItre (Švandovo divadlo)

Výuka a vzdělávání

- Mgr. a Ph.D. programy v počítačové lingvistice, zpracování řeči
  - Na všech velkých univerzitách v ČR
  - Technické i humanitně orientované
- Organizace seminářů, letních škol a workshopů
  - LINDAT/CLARIAH-CZ, ÚFAL MFF UK, VUT Brno (např. ACL 2007, Interspeech 2021), MU+ZČU pořádají „Text, Speech, Dialogue“ každý rok
  - Technologické semináře
  - Semináře pro Digital Humanities
  - Právní aspekty (autorské právo, jak pracovat s daty (vč. ochrany dat), otevřená data)
- „Centrum znalostí“
  - LINDAT/CLARIAH-CZ společně s norským centrem CLARINO
    - Jak anotovat korpusy
  - Technická pomoc s repozitáři, návrh uchovávání otevřených dat