
Centrum počítačn lingvistiky

Centrum počítačn lingvistiky

Matematicko-fyzikln fakulta
Univerzita Karlova v Praze

Vedoucí: Prof. PhDr. Eva Hajičová, DrSc.

Centrum počítačn lingvistiky bylo založeno v rámci vdeckho programu Nrodnch center v roce 2000. Ke společnmu vzkumu se spojila pracovišt Karlovy univerzity, Ústavu pro jazyk esk Akademie vd a Zpadočesk univerzity Plzeň. Centrum je součst Matematicko-fyzikln fakulty UK v Praze. Pracuje v nm v současné době 41 pracovník, z toho 70% mladch pracovník do 35 let.

Centrum se zabv teoretickm i aplikovanm vzkumem v oblasti počítačn lingvistiky, zamřuje se na kombinovn empirickch a stochastickch metod studia prozench jazyk s drazem na eštnu. Vzkum se zabv psanm i mluvenm jazykem.

Centrum m 4 oddlen:

1. počítačn a teoretick analza eštny na zklad Pražskho zvislostnho korpusu
2. matematick zklady a metody
3. analza esk mluven řeči
4. vcejazyčn vzkum a aplikace

CKL je uniktnm vzkumnm pracovištm v ČR v oblasti počítačn lingvistiky a automatickho zpracovn prozenho jazyka, a to pedevšm tm, že

- je v nm integrovn a vzjemn posilovn vzkum jak jazyka psanho, tak mluvenho
- vzkum m pevn a originln teoretick zklady v oblasti gramatiky i lexika, ale neztrc ze zretele ani aspekty aplikačn
- ve vzjemn rovnovze jsou lingvistick i informatick aspekty tohoto vzkumu.

V tomto smyslu se d říci, že v tto integraci m CKL i vznamn postaven mezi pednmi svtovmi centry vzkumu v počítačn lingvistice, o čemž svdčí i zjem o naše vsledky.

Hlavn oblasti vzkumu

1. Teoretick aspekty počítačn lingvistiky

Teoretick aspekty počítačn lingvistiky se zamřenm na vzkum psan i mluven eštny a s ohledem na možn aplikace. Používn Pražskho zvislostnho korpusu (dle PZK), kter obsahuje poloautomatickou analzu stovek tisíc eskch vt, umořňuje vzkum na všš úrovni než dve.

2. Jazykov zdroje

Zamřujeme se na využívan bohatch jazykovch zdroj pro eštnu, zejména proto, že eštna je jedin jazyk s bohatou morfologi, kter byl analyzovn v podobn mře.

3. Vvoj statistickch i strukturnch metod pro eštnu

Metodologie je založena na prohloubenm studiu, porovnvn a kvalifikovanm využítn postup strukturnch i statistickch metod strojovho uen, s ohledem na specifick typologick vlastnosti eštny jako vysoce flexivnho jazyka. V tomto ohledu vyvjme originln metodologii, neboť dosud navrhovan pstupy erpj ze zdroj v angličtin a v jinch jazycch s nzkm stupnm flexe.

4. Matematické základy a metody

Pozornost je věnována matematickým a výpočetním základům vyvíjených metod, algoritmů a postupů zpracování přirozeného jazyka.

5. Rozpoznávání řeči

Výzkum v oblasti systémů rozpoznávání mluvené řeči je založený na statistických metodách. Zvláštní důraz je kladen na studium suprasegmentálních jevů, jako je význam větné prozodie, a jazykového modřelování, opět s přihlédnutím k tomu, že čeština je jazyk s vysokou inflexí. V průběhu řešení projektu bylo Centrum pozváno jako spoluřešící (a pro evropskou oblast koordinující) pracoviště mezinárodního projektu [MALACH](#) (Multilingual Access to Large Spoken Archives), jehož cílem je vývoj systémů pro automatický přepis svědeckých výpovědí lidí, kteří přežili holocaust. Svědecké výpovědi byly pořízeny ve více než 30 různých jazycích a česká strana je spoluzodpovědná za zpracování jazyků střední a východní Evropy. Na projektu participují Visual History Foundation v Hollywoodu, Johns Hopkins University v Baltimore, University of Maryland, IBM, MFF UK v Praze a ZČU v Plzni. Anotační práce na zpracování svědeckých výpovědí jsou podporovány National Science Foundation v USA.

6. Vícejazyčný výzkum

Využití vícejazyčných zdrojů umožňuje získat znalosti o češtině srovnatelné s výsledky zpracování jiných jazyků. Pozornost se zaměřuje na studium a použití paralelních korpusů se zaměřením na široké spektrum aplikací, například získávání informací z vícejazyčných textů, automatický či strojově podporovaný překlad nebo základní výzkum multimediální interakce.

Mezinárodní spolupráce

Centrum využívá mezinárodních kontaktů zakládajících institucí a je zapojeno do intenzivní a vysoce plodné spolupráce s podobnými výzkumnými centry a univerzitami v zahraničí. Aktivně spolupracuje mj. s těmito institucemi: Department of Computer Science, Johns Hopkins University, Baltimore, USA; Istituto di Linguistica Computazionale, C.N.R. Pisa, Itálie; Natural Language Systems Institute, Fachbereich Informatik, Universität Hamburg; Institut für Übersetzen und Dolmetschen, Universität des Saarlandes, Německo; Institut national des langues et civilisations orientales (INALCO), Paris, Francie; Xerox Research Center Europe, Language Research Group, Grenoble, Francie.

Vedle toho probíhá dlouhodobá intenzivní spolupráce s dalšími předními pracovišti na University of Pennsylvania ve Philadelphii a University of Massachusetts v Amherstu v USA a na univerzitách v Uppsale a v Edinburghu.

Pracovníci Centra se pravidelně účastní hlavních mezinárodních konferencí v oboru (COLING, ACL, EACL, ECAI, IJCAI, TSD apod.).