Tisková zpráva k závěrečné zprávě projektu Centra komputační lingvistiky

Cílem Centra komputační lingvistiky (CKL) byl výzkum a vývoj v oblasti moderní počítačové lingvistiky na zcela nové úrovni založené na jedinečné vícerovinné analýze velmi rozsáhlého korpusu. Činnost Centra, díky kterému se podařilo vytvořit u nás jediný integrovaný tým pro výzkum psané i mluvené řeči, měla a má velký význam pro aplikace v mnoha oborech služeb a průmyslu, které pracují s komunikací člověka s počítačem.

Stěžejním projektem CKL bylo vybudování Pražského závislostního korpusu, což je soubor českých textů s bohatou informací o morfologii, větné stavbě a významové struktuře vět (první verze korpusu, "Prague Dependency Treebank, Version 1.0" byla vydána na CD-ROM v roce 2001, druhá verze, "Prague Dependency Treebank, Version 2.0" bude vydána v roce 2005). Takový soubor textů slouží jednak dalšímu teoretickému zkoumání češtiny, zejména jde však o velké množství lingvisticky zpracovaných dat, která jsou nezbytná pro automatické zpracování přirozeného jazyka pro jakýkoliv aplikovaný úkol – strojový překlad, vyhledávání informací (tzv. data mining), automatické "porozumění" textu i jeho generování.

Druhým základním směrem Centra byl statisticky založený výzkum v oblasti rozeznávání mluvené řeči. Výsledky tohoto směr výzkumu byly dán k dispozici odborné veřejnosti jako "Czech Broadcast News Corpus" a " Czech Broadcast News Transcripts" na dvou CD-ROM v roce 2004. Zásadním přínosem bylo zapojení Centra do mimořádně rozsáhlého mezinárodního projektu MALACH (Multilingual Access to Large Spoken Archives), jehož cílem je vývoj systémů pro automatický předpis svědeckých výpovědí lidí, kteří přežili holocaust. Svědecké výpovědi byly pořízeny ve více než 30 různých jazycích a česká strana je prostřednictvím Centra spoluzodpovědná za zpracování jazyků střední a východní Evropy.

Dalším cílem výzkumu Centra bylo vytváření a využívání vícejazyčných zdrojů. Pozornost byla věnována zejména studiu a uplatnění paralelních korpusů se zaměřením na strojový a strojem podporovaný překlad – v roce 2004 byla vydána unikátní sada počítačových databází a nástrojů pro automatický překlad "Prague Czech-English Dependency Treebank, PCEDT 1.0". Takto pojatá výzkumná činnost vedla k získání dalších znalostí o češtině srovnatelných s výsledky výzkumu jiných jazyků.

Nepostradatelnou součástí činnosti Centra komputační lingvistiky jako centra základního výzkumu byl výzkum teoretických aspektů komputační lingvistiky se zaměřením především na češtinu v podobě psané i mluvené a s ohledem na možné aplikace. Metodologie výzkumu v rámci Centra byla založena na prohloubeném studiu, porovnávání a kvalifikovaném využití postupů strukturních i statistických včetně metod strojového učení, s ohledem na specifické typologické vlastnosti češtiny jako vysoce flexívního jazyka.

Jak ukázala veřejná vědecká rozprava o výsledcích Centra konaná ve dnech 29.-30. listopadu 2004, za účasti 7 předních zahraničních vědců z oboru komputační lingvistiky, tyto výsledky mají přední místo v evropském i světovém výzkumu a jsou ve světě přijímány s vynikajícím ohlasem.

Činnost Centra bohatě naplnila očekávané možnosti v navazování a udržování těsných kontaktů s českým a mezinárodním průmyslem využívajícím počítače, o čemž svědčí i zájem partnerů a uživatelů z oblasti aplikační sféry o vhodně zpracované a užitečné zdroje pro široce založený vývoj a aplikace.

V Praze, dne 20.1.2005

prof. PhDr. Eva Hajičová, DrSc.
(řešitelka projektu)