

Výsledky dosažené v roce 2008

Řešitelské pracoviště MFF UK

Podrobně jsou výsledky projektu dokumentovány na stránce <http://ufal.mff.cuni.cz/rest>. Stěžejním výsledkem projektu v roce 2008 (jakož i kompletního řešení projektu) bylo vydání **CD-ROM Český akademický korpus verze 2.0 (ČAK 2.0)**.

- CD-ROM bylo vydáno v nakladatelství Linguistic Data Consortium, Philadelphia, PA, USA (ISBN: 1-58563-491-3).
- Datová složka CD-ROM obsahuje morfologicky a syntakticky anotovaný ČAK o celkovém objemu 660 tis. slov. ČAK se přičlenil do rodiny tzv. pražských závislostních korpusů, které svým objemem a koncepcí anotací patří k nejvýznamnější počínům světové korpusové a počítačové lingvistiky.
- Složka nástrojů CD-ROM obsahuje nástroje pro automatické morfologické a syntaktické zpracování českých textů. Jsou prezentovány jejich poslední verze s dosud nejlepší dosaženou úspěšností.
- Do obsahu CD-ROM jsou začleněny demosnímky k nástrojům s grafickým uživatelským rozhraním, prezentovanými na CD-ROM. Prostřednictvím demosnímků získají prvotní představu a používání a využívání nástrojů.
- CD-ROM obsahuje dva bonusové materiály – elektronickou cvičebnici STYX a modul pro hlasové ovládání anotačního nástroje TrEd.
- Elektronická forma Průvodce ČAK 2.0 je k dispozici na adrese http://ufal.mff.cuni.cz/rest/CAC/cac_20.html; tištěný Průvodce ČAK 2.0 byl publikován v časopise Prague Bulletin of Mathematical Linguistics (Vidová Hladká a kol., 2008). Průvodce je sepsán česky a anglicky.
- Koncepce CD-ROM byla nastavena tak, aby bylo uživatelům, bez ohledu na míru jejich počítačové zdatnosti, zajištěno maximální pohodlí při práci s CD-ROM.

Alternativní způsob získávání anotovaných dat – v rámci pilotního projektu získávání anotovaných dat prostřednictvím webových her byl otevřen herní portál LGame (www.lgame.cz), byla implementována první hra a byl zpracován návrh grafického prostředí portálu.

Čestné uznání – ČAK 1.0 byl použit jako hlavní zdroj dat pro zápočtové příklady zadané při přednášce Úvod do strojového učení (v počítačové lingvistice) konané v ZS 2006/07 na MFF UK (přednášejí Hladká, Ribarov). Studentka MFF UK Jana Kravalová přihlásila, pod vedením Barbory Vidové Hladké, svoje řešení zápočtové úlohy do soutěže SVOČ 2008 a v soutěži obdržela čestné uznání (<http://cms.jcmf.cz/svoc/vys08.html>).

Videonahrávky – Byly pořízeny záznamy z téměř všech přednášek konaných v rámci Semináře z formální lingvistiky pořádaného ÚFAL MFF UK. Každá nahrávka je k dispozici v několika formátech: Flash Video – lze přehrávat on-line; H.264 MP4 – kvalitnější on-line video, zatím jsme jedni z mála, kdo tento formát využívá; Xvid – video ke stažení ve vyšší a nižší kvalitě; MP3 – pouze audio. Webové stránky (<http://lectures.ms.mff.cuni.cz>) obsahují veškerý natočený materiál.

Publikace

1. Bielický Viktor, Smrž Otakar: Building the Valency Lexicon of Arabic Verbs, In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, ISBN 2-9517408-4-0, 2008.
2. Hlaváčová Jaroslava, Hrušecký Michal: "Affisix" Tool for Prefix Recognition, In *Lecture Notes in Computer Science, Vol. 5246, Proceedings of the 11th International Conference*, TSD 2008, Springer-Verlag, Berlin Heidelberg, ISBN 978-3-540-87390-7, ISSN 0302-9743, pp. 85-92, 2008.
3. Hlaváčová Jaroslava, Lopatková Markéta: Variants and Homographs: Eternal Problem of Dictionary Makers, In *Lecture Notes in Computer Science, Vol. 5246, Proceedings of the 11th International Conference*, TSD 2008, Springer-Verlag, Berlin Heidelberg, ISBN 978-3-540-87390-7, ISSN 0302-9743, pp. 93-100, 2008.
4. Hlaváčová Jaroslava, Kolovratník David: Morfologie češtiny znovu a lépe, In *Informačné Technológie – Aplikácie a Teória. Zborník príspevkov, ITAT 2008*, Seňa, Slovakia, ISBN 978-80-969184-8-5, pp. 43-47, 2008.
5. Křen Michal, Hlaváčová Jaroslava: Corpus as a Means for Study of Lexical Usage Changes, In *Proceedings of the 13th EURALEX International Congress*, ISBN 978-84-96742-67-3, pp. 437-447, 2008.
6. Mírovský Jiří: Netgraph Query Language for the Prague Dependency Treebank 2.0, In *Prague Bulletin of Mathematical Linguistics*, Vol. 90, Charles University, ISSN 0032-6585, 2008.
7. Mírovský Jiří: Towards a Simple and Full-Featured Treebank Query Language, In *ICGL 2008 Proceedings of the First International Conference on Global Interoperability for Language Resources*, City University of Hong Kong, Hong Kong, pp. 171-178, 2008.
8. Mírovský Jiří: Netgraph 1.95 - a tool for searching in PDT 2.0, ÚFAL MFF UK, 2008.
9. Mírovský Jiří: Netgraph - Making Searching in Treebanks Easy, In *IJCNLP 2008 Proceedings of the Third International Joint Conference on Natural Language Processing*, International Institute of Information Technology, Hyderabad, India, pp. 945-950, 2008.
10. Jiří Mírovský: Does Netgraph Fit Prague Dependency Treebank?, In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008.
11. Mírovský Jiří: PDT 2.0 Requirements on a Query Language, In *ACL 2008 Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*, ISBN 978-1-932432-06-0, pp. 37-45, 2008.
12. Mírovský Jiří, Panevová Jarmila: Learning to Search in the Prague Dependency Treebank, In *Grammar & Corpora / Gramatika a korpus 2007*, Copyright (C) Academia, ÚJČ AV ČR, Prague, Czech Republic, ISBN 978-80-200-1634-8, pp. 105-111, 2008.
13. Smrž Otakar, Bielický Viktor: ElixirFM, Software prototype, [<http://sourceforge.net/projects/elixir-fm/>], SourceForge.net, 2008.
14. Smrž Otakar, Bielický Viktor, Kouřilová Iveta, Kráčmar Jakub, Hajič Jan, Zemánek Petr: Prague Arabic Dependency Treebank: A Word on the Million Words, in *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, Marrakech, Morocco, pp. 16-23, 2008.
15. Vidová Hladká Barbora, Hajič Jan, Hana Jiří, Hlaváčová Jaroslava, Mírovský Jiří, Raab Jan: The Czech Academic Corpus 2.0 Guide, In *Prague Bulletin of Mathematical Linguistics*, No. 89, Charles University, ISSN 0032-6585, pp. 41-96, 2008.

16. Vidová Hladká Barbora, Hajič Jan, Hana Jiří, Hlaváčová Jaroslava, Mírovský Jiří, Raab Jan: *Czech Academic Corpus 2.0*, LDC - Linguistic Data Consortium, ÚFAL MFF UK, Prague, Czech Republic, ISBN 1-58563-491-3, 2008 .
17. Vidová Hladká Barbora, Kebrt Michal. Videonahrávky přednášek. 2008. [<http://lectures.ms.mff.cuni.cz/video/categoryshow>]

Přednášky

1. Smrž Otakar: Programming the Arabic Treebank, Dublin Computational Linguistic Research Seminars, Dublin, Ireland, Apr 2008 .
2. Vidová Hladká Barbora: Vydáváme Český akademický korpus 2.0. Co dál? [<http://lectures.ms.mff.cuni.cz/video/recordshow/index/27/82>]. Seminář formální a aplikované matematiky, MFF UK, May 2008.

Spolurešitelské pracoviště ÚJČ AV ČR

Cílem závěrečného roku projektu 2008 bylo dovršit dlouhodobý projekt, tj. v pořadí důležitosti hlavních okruhů:

a/ dokončit a uzavřít práce na *datových zdrojích* (zejm. na *elektronických verzích historických slovníků* a na *speciálních databázích*) – splněno, viz níže

b/ prezentovat prostřednictvím webu dokončené práce na *dostupných webových stránkách* (včetně vzniku nových webových *nástrojů*) – splněno, viz níže

c/ dokončit pilotní podoby dílčích projektů *Naše řeč, české morfémy* a *slovní sítě* – splněno, viz níže

d/ převést dlouhodobé práce (zejména *lemmatizaci lexikálního archivu*) do rámce navazujících projektů – splněno, viz níže

ad a/ Ve spolupráci s Knihovnou AV ČR dokončena digitalizace historického Jungmannova *Slovníku česko-německého* – šesti svazků s autorovými úpravami, poznámkami a doplňky (cca 4755 digitálních obrazů) – uzavřeno v souladu s harmonogramem v květnu 2008.

Ve spolupráci s Památkem národního písemnictví a firmou IMD digitalizací dokumentovány a zálohovány zbylé unikátní části rukopisů k Jungmannovu slovníku (400 stran) včetně autografu čistopisu a Čelakovského kartotéky (20 000 lístků) – uzavřeno v souladu s harmonogramem v červenci 2008.

Ve spolupráci s firmou IMD vytvořen korpus současných mluvených textů z rozhlasu a televize – uzavřeno v souladu s harmonogramem v červenci 2008.

Ve spolupráci s firmou Imaging Systems dokončeny digitalizace vybraných archivních textů a publikací (zkratky, značky, akronymy, lingvistika, matematika, religionistika, regionalia, kroniky) – v souladu s harmonogramem dokončeno v prosinci 2008.

ad b/ Digitalizovaný devítisvazkový *Příruční slovník jazyka českého* zpřístupněn ve spolupráci s ředitelstvím ÚJČ plným zprovozněním vzájemného propojení elektronizované verze PSJČ a lexikálního archivu ÚJČ (v systému nástrojů www BARA) na adrese <http://bara.ujc.cas.cz/psjc/> .

Digitalizované autorské exempláře Jungmannova *Slovníku česko-německého* s rukopisnými doplňky přiřazeny k projektu webového zpřístupnění celého slovníku na adrese <http://www.slownjk.cz/slownjkc/> .

Digitalizovaný historický desetisvazkový Kottův *Slovník česko-německý* zpřístupněn s novými vyhledávacími nástroji na adrese <http://kott.ujc.cas.cz/> .

ad c/ Provedena a zpřístupněna plná elektronizace prvních 14 ročníků časopisu *Naše řeč* (do roku 1930) na adrese <http://naserec.ujc.cas.cz> .

Soustředěn, zpracován a zpřístupněn kompletní inventář českých morfémů na adrese <http://www.morfemy.cz/repertoar/> .

ad d/ Těžiště dlouhodobé části projektu – sub b/ rozvedené vzájemné propojení elektronizované verze *Příručního slovníku jazyka českého* a lexikálního archivu – bylo s předstihem plně převedeno mezi ústavní úkoly ÚJČ tak, aby bylo plynule zajištěno jak zpřístupnění, tak další postupné doplňování.

Všechna data dostupná v elektronické podobě byla zpracovávána tak, aby mohla být využita také pro budoucí morfologicko-derivační slovník češtiny. Bylo vytvořeno interní uživatelské webové prostředí pro vkládání slovních kořenů a odvozených slov, ve kterém je možné definovat slovtvorné vztahy a zobrazovat je ve stromové struktuře.

Publikace

1. Králík, J.: Statistické metody v korpusové lingvistice (PhD-B 7) (syllabus), *Korpusová lingvistika – přehledy*, ÚČNK 2008, s. 69-71
2. L. Uhlirova: On word length: The influence of a boundary condition on the modelling. *Glottology*, vol. 1, No. 1, 2008, 55-64. (Trnava) ISSN 1337 – 7892.

Zahraníční cesty

1. J. Králík – *Budapešť*: seminář ke koncepci využití dat a nástrojů při tvorbě moderních výkladových slovníků a při elektronizaci archivních dat.
2. Rangelova – *Sofia*: seminář k datovým základnám pro komutační lexikologii
3. L. Veselý – *Budapešť*: studium k získávání dat pro formální kategorizaci gramatických popisů typologicky různých jazyků
4. M. Laštovičková – *Londýn*: studium k vytváření datové základny a k analýze jazyka specifických obsahových okruhů
5. M. Laštovičková – *Poznaň*: referát na konferenci k multidimenzionálním lingvistickým studiím
6. J. Králík – *Graz*: seminář k dějinám a záměrům kvantitativní lingvistiky a příprava konference kvantitativní lingvistiky QUALICO
7. Společná publikace (Budapešť) – v tisku

Zhodnocení řešení projektu

Všechny původně vytčené cíle i jejich logické rozšíření, doplňky a zpřesnění **byly uskutečněny** dle naplánovaného harmonogramu.

Řešitelské pracoviště UK MFF: Nosným tématem projektu bylo zpracování Českého akademického korpusu do podoby kompatibilní s rodinou tzv. pražských závislostních korpusů. Zpracování v sobě zahrnuje jak ruční práci (z pohledu anotací), tak i vývoj automatických procedur. V obou dvou složkách projektu **se podařilo dosáhnout naplánovaných cílů a uceleně je prezentovat** ve formě CD-ROM, a sice CD-ROM Český akademický korpus 2.0 vydané americkým vydavatelstvím LDC. CD-ROM je koncipováno tak, že zajišťuje uživatelům maximální pohodlí při práci s korpusem a s nástroji. Spolupráce se spoluřešitelem byla vysoce přínosná v tom směru, že se podařilo navázat na vynikající výsledky, kterých dosáhlo spoluřešitelské pracoviště před více než dvaceti lety.

Bylo dosaženo i několika vědeckých výsledků nad rámec plánů. Zkušenosti s ruční anotací odborníky se promítly i do myšlenky hledání alternativního způsobu získávání anotovaných dat. Tento představují internetové hry. První pilotní hra byla implementována.

Projekt zastřešoval i pořizování videonahrávek vědeckých akcí, jejich kompletní zpracování a publikování na internetu. Tímto krokem jsme výrazně rozšířili možnosti prezentace vědeckých výsledků.

Spoluřešitelské pracoviště ÚJČ AV ČR: Původní cíl projektu byl postaven na kooperaci při využití Českého akademického korpusu a směřoval k jeho zásadnímu informačnímu povýšení a zveřejnění. Tento cíl byl **beze zbytku splněn** a postupně rozšířen o úzce související témata elektronizace a prezentace dalších rozsáhlých lingvistických dat. Také tyto následně upřesňované cíle byly **úspěšně naplněny** a zároveň otevřely i další dlouhodobější výhledy, jejichž dovršení se mohlo stát předmětem nových projektů.

Nejvýznamnější výsledky za celou dobu řešení projektu

Plná **digitalizace** a sofistikovaná **prezentace** *Českého akademického korpusu*, *Příručního slovníku jazyka českého* (s propojením na rozsáhlý *lexikální archiv*), *Kottova slovníku* a *Jungmannova slovníku*. Vytvoření rozsáhlé *diverzifikované základny dat* a *nástrojů* pro další lingvistická studia a pro bezprostřední využití v oblasti počítačového zpracování přirozeného jazyka a v oblasti lexikografie.