AKADEMIE VĚD ČESKÉ REPUBLIKY
Národní 3, 117 20 Praha 1

List PT1

ZPRÁVA O PRŮBĚHU PRACÍ

na programovém projektu „Informační společnost“

Rok	*Identifikační kód projektu*
2005	**1ET101120413**

01 Řešitel projektu

Jméno: Mgr. Barbora Vidová Hladká, PhD.

02 Příjemce 1

Název: Univerzita Karlova v Praze

Název projektu:

Data a nástroje pro informační systémy

Doba řešení: 1.7.2004 – 31.12.2008

Splnění cílů projektu * ☒ není ohroženo ☐ je ohroženo

Datum: 13.1.2006

Podpis řešitele: .....

Podpis a razítko statutárního zástupce Příjemce 1: .....

** vhodnou odpověď zaškrtněte*

03 Přehled uplatněných výsledků za rok 2005

Aplikace[*]	2					
Publikace v recenzovaných časopisech	*publikované*	1	*ks*	*přijaté k publikaci*	3	*ks*
Ostatní publikace	*publikované*	4	*ks*	*přijaté k publikaci*	0	*ks*
Patenty	*udělené*	0	*ks*	*podané*	0	*ks*
Jiný (specifikovat)	0					

[*]/ Podrobnější informace o formě aplikací uveďte v textové části zprávy.

AKADEMIE VĚD ČR

ZPRÁVA O PRŮBĚHU PRACÍ V ROCE 2005

na programovém projektu „Informační společnost“

Podrobně jsou výsledky projektu dokumentovány na stránce http://ufal.mff.cuni.cz/rest. Anotační nástroj LAW a implementace perceptronového taggeru jsou aplikačními výstupy projektu.

Řešitelské pracoviště MFF UK

Český akademický korpus Proběhla revize diskrepancí ručních korektur, včetně ručního dořešení zbylých nesrovnalosti. Hlavním formátem Pražského závislostního korpusu se stal tzv. Prague Markup Language (PML), který podporuje koncepci odděleného uložení anotací jednotlivých rovin anotování. Abychom kontinuálně zajistili kompatibilitu ČAK s PZK, provedli jsme konverzi Korpusu do formátu PML. Dle původního plánu měla být první verze Korpusu vydána na konci roku 2005. Vzhledem k neplánované konverzi vnitřního formátu byla kontrola morfologických anotací zahájena později, a tudíž se protáhla do začátku roku 2006. Paralelně s přípravou ČAK 1.0 byla navržena konverzní procedura syntakticko-analytických anotací. Na specifikaci konverzní procedury se podíleli také kolegové z ÚJČ. Práce na ČAK odčerpaly 38% prostředků OON.

Nástroj pro lexikální anotaci Anotační nástroj LAW (současná verze 0.6.0) je testován v rámci dvou anotačních projektů: morfologické anotování části testovacích dat paralelního česko-anglického korpusu PCEDT a zjednoznačňování nejednoznačných hodnot vybraných morfologických kategorií na datech z PDT 2.0. Aktuálně jsou zapracovávány podněty od anotátorů. Na vývoj nástroje bylo spotřebováno 22% prostředků OON a na jeho testování 35%.

Modifikace morfologického analyzátoru Nový morfologický analyzátor založený na konečných automatech je téměř dokončen. Jeho úplné dokončení se oproti plánu poněkud opozdilo, ale zkušební provoz bude zahájen během několika týdnů. Byl dokončen editor pro morfologický slovník (SLED), který navíc kromě editorských funkcí dokáže "odhadovat" vzory nově zadávaných lemmat. Jeho praktické použití je vázané na zpracování dat pro morfologický analyzátor. Začali jsme s vývojem nového "guesseru", nástroje, který dokáže určit lemma a morfologickou značku neznámého slova. Prvním krokem byla automatická analýza slov z ČNK, jejímž cílem bylo sestavení seznamu nových "předpon" užívaných při vytváření novotvarů v češtině.

Nové metody tagování Jedna z nejnovějších metod korpusového modelování tzv. „průměrovaný“ perceptron Michaela Collinse (publikováno 2002) byla implementována pro potřeby tagování češtiny v rámci studentského projektu Morče – oceněno druhým místem v ACM Student Research Competition 2005. Úspěšnost algoritmu závisí především na zvolené sadě rysů popisujících kontext, na jehož základě se značky (tagy) vybírají. Proto na studentský projekt navazovala diplomová práce, která v rámci daných možností důkladně mapovala možné sady rysů, jejich úspěšnosti a vztahy mezi nimi. Přetrénování tohoto taggerů, HMM taggeru a exponenciálního taggeru nemohlo být uskutečněno vzhledem k nedokončené kontrole dat (viz výše).

Čerpaní **investičních** prostředků je specifikováno ve finančním výkazu. **Cestovní náklady** byly odčerpány na zahraniční letní školy – Machine Learning Summer School, Chicago, květen, Ribarov, Smrž, Vidová Hladká; ESSLLI, Edinburgh, červenec, Mírovský - a na účast na konferenci TLT, Barcelona, prosinec, Vidová Hladká (další ročník konference TLT v roce 2006 bude hostit ÚFAL MFF UK). V položce **služby** bylo hrazeno ADSL připojení, konverze vnitřního formátu dat a oprava počítačů. V položkách **DHM a NHM** činily největší položky nákupy tiskárny a 2 LCD panelů.

Ve změnovém listu A/4 uvádíme kompletní personální obsazení projektu pro rok 2006.

Spoluřešitelské pracoviště ÚJČ AV ČR

- vytvořen základ kompletní elektronizace dat historického **Česko-německého slovníku Fr. Št. Kotta** (1878-1906, 10 204 stran, ~ 250 000 hesel) včetně skenování, obnovy historických knižních vazeb a elektronizace textové podoby, vypracovány softwarové nástroje a za jejich pomoci dokončena první etapa textových korektur
- vytvořena elektronická podoba **Lexikálního standardu ÚJČ** (1970, 1000 stran) včetně nástrojů pro automatické i věcné revize elektronické podoby dat a jejich vnitřní konzistence
- vytvořena elektronická báze kompletních **heslářů** *PSJČ, SSJČ, SSČ, SSJ-jména, Lex. standard* a *FSČ* a vypracovány softwarové nástroje pro jejich konfrontaci a lexikografické využití
- vytvořeny **automatické nástroje** pro revizi vnitřní konzistence elektronické podoby všech textových dat PSJČ (elektronická podoba pořízena již dříve)
- dokončena **statisticky využitelná verze ČAK** (softwarová úprava prohlížeče, pracovní frekvenční slovník tvarů aj.)
- souběžně probíhaly expertní a pomocné práce pro *elektronizaci lexikálního archivu ÚJČ* (zajišťuje jiný projekt ÚJČ)
- spoluřešitelé se účastnili spolupráce na projektu a přípravě dat pro *lexikální databázi ÚJČ*

Z **rezervy věcných prostředků** byl pořízen nákup *odborné literatury* a zdrojů pro lexikální excerpci. Z investičních prostředků bylo obnoveno vybavení *třemi počítači* (dosud užívané PC z inventáře ÚJČ odepsány).
Cestovní náklady byly čerpány na zahraniční cesty (Novi Sad – *účast na konferenci* (referát Uhlířová), Budapešť – *příprava dat trojjazyčného slovníku* (Králík), Graz – *příprava konference kvantitativní lingvistiky* (Králík), Moskva – *konzultační pobyt* (Rangelova) a na účast na domácích konferencích (*Gramatika a korpus* (referát Uhlířová, Klímová, Holubová), *Slovní poklad češtiny* (referát Králík, referát Klímová, referát Holubová)).

Publikace (za celý projekt)

Hladká B, Králík J.: Český akademický korpus mezi dvěma tisíciletími. *Slovo a slovesnost*, přijato k tisku, 2006.

Hlaváčová J.: Average Reduced Frequency. 2o Coloquio de Ling Comp, Mexico City, Mexiko, 2005.

Hlaváčová J:. Korpusové chyby. Ve *Sborník konference Gramatika a korpus*, s. 22-24, ÚJČ AV, Praha, ČR, 2005.

Hlaváčová J.: Orwell's 1984 - playing with Czech and Slovak versions. Ve *Sborník konference SLOVKO*, v tisku, 2005.

Králík J., Uhlířová L.: The Czech Academic Corpus (CAC), its history and presence, *Journal of Quantitative Linguistics*, přijato k tisku, 2006.

Spousta M.: *Automatické přiřazování tvaroslovných tvarů v češtině*. Diplomová práce, MFF UK, 2005.

Urrea A. M.; Hlaváčová J.: Automatic Recognition of Czech Derivational Prefixes, In LNCS/Lecture Notes in Artificial Intelligence/Proceedings of the 6th International Conference CICLing, pp. 189-197 (eds. Alexander Gelbukh), Mexico City, Mexico, Feb. 13-19, 2005.

Votrubec, J.: *Volba vhodné sady rysů pro morfologické značkování češtiny.* Diplomová práce, MFF UK, 2005.

AKADEMIE VĚD ČR

PROGRAM PRACÍ NA ROK 2006

na programovém projektu „Informační společnost“

Řešitelské pracoviště MFF UK

Český akademický korpus

- **Vydání CD ROM ČAK 1.0**
  - duben 2006, nakladatelství Karolinum
  - náklad 150ks – brožurka s CD ROM
  - obsah CD ROM

```
data/             # ČAK 1.0
      pml/        # data ve formátu PML (w-soubory, m-soubory)
      csts/       # data ve formátu CSTS
doc/              # průvodce ČAK
tools/            # nástroje
      Bonito/     # pro vyhledávání v Korpusu
      LAW/        # pro anotování
      Morphology/ # pro morfologickou analýzu a tagování
```

- **Konverze syntakticko-analytických anotací**. Konverzní procedura bude aplikována nejdříve na vybranou část Korpusu, která bude následně předložena zkušenému anotátorovi. Anotátor data opraví a poskytne podklady pro vylepšení konverzní procedury. Podle charakteru připomínek bude zvolena strategie dalšího zpracování s ohledem na poměr ruční práce a možnosti automatizace.

Nástroj pro lexikální anotaci

- podpora PML formátu
- zapracování možností pro „libovolné“ lexikální anotování
- zpracování vstupního textu parsery „na vyžádání“
- zapracovávání aktuálních připomínek anotátorů

Modifikace morfologického analyzátoru

- doplnění morfologického analyzátoru o guesser
- testování editoru SLED během anotování nástroje LAW
- ladění dle zpětné vazby od uživatelů

Nové metody tagování

- přetrénování taggerů na ČAK 1.0
- vyhodnocení a porovnání úspěšností
- aplikace perceptronového modelu tagování na arabské texty
- kombinace perceptronového modelu s ručně navrženými pravidly

Plán. Pracoviště za spoluúčasti několika projektů vybudovalo základ 64-bitového výpočetního clusteru, který postupně přejímá veškeré výpočetně náročné procesy. Provedeme buď upgrade dvou uzlů tohoto výpočetního clusteru, nebo pořídíme jeden nový uzel clusteru.

Plán OON
Většina prostředků je plánována na práci s ČAK (vydání CD ROM, konverze syntakticko-analytických anotací). Dále budou prostředky určeny na další vývoj a testování anotačního nástroje LAW.

Spoluřešitelské pracoviště ÚJČ AV ČR

Data

- dokončit kompletní elektronizaci dat (textové korektury) historického **Česko-německého slovníku Fr. Št. Kotta**
- vypracovat koncepci elektronizace **dalších základních** (historických) **slovníků** češtiny a zahájit jejich skenování
- připravit první testovací verzi **trojjazyčného slovníku** anglicko-česko-maďarského
- využít ke kvantitativní exploataci **CD verzi ČAK**

Nástroje

- interně (na CD) uživatelsky zpřístupnit data historického **Česko-německého slovníku Fr. Št. Kotta**
- vytvořit softwarové nástroje pro automatickou revizi OCR a vnitřní konzistence textových dat **Jungmannova slovníku**
- nalézt, adaptovat a aplikovat vhodný software pro doplňování rozsáhlé databáze **lexikálního archivu ÚJČ**

Plánované zahraniční cesty

- ústav MTI Budapešť (dlouhodobá spolupráce)
- universita Tallin (konference kvantitativní lingvistiky)
- universita Trevír (dlouhodobá spolupráce)