AKADEMIE VĚD ČESKÉ REPUBLIKY
Národní 3, 117 20 Praha 1

List PT1

ZPRÁVA O PRŮBĚHU PRACÍ

na programovém projektu „Informační společnost“

Rok	*Registrační číslo*
2004	**T101120413**

01 Řešitel projektu

Jméno: Mgr. Barbora Vidová Hladká, PhD

02 Příjemce

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: 3.1. 2005

Podpis řešitele 1: .....

Podpis a razítko statutárního zástupce Příjemce 1: .

.....

* *vhodnou odpověď zaškrtněte*

AKADEMIE VĚD ČR

ZPRÁVA O PRŮBĚHU PRACÍ V ROCE 2004

na programovém projektu „Informační společnost“

(Doporučený rozsah zprávy je 1 strana textu.
*Při její formulaci se soustřeďte na **zdůvodnění použitých** účelových finančních **prostředků**.)*

Projekt je rozdělen do čtyř tématických celků - cíle naplánované pro jednotlivé celky pro první rok řešení byly splněny. Podrobně jsou výsledky dokumentovány na domácí stránce projektu http://ufal.mff.cuni.cz/rest; zde uvádíme základní přehled pro řešitelské a spoluřešitelské pracoviště.

Řešitelské pracoviště MFF UK

Český akademický korpus (původně Korpus věcného stylu) Konverze originálního vnitřního formátu korpusu do kódování SGML proběhla plynule. Tímto konverzním krokem a formální úpravou korpusu a anotačního nástroje TrEd bylo možné zajistit uživatelsky příjemnou práci s korpusem – na úpravu TrEd byla odčerpána část prostředků OON. Původní strategie značkování ČAK byla konzultována s kolegy z ÚJČ. Ukázalo se, že z textu byla vypuštěna interpunkce. Zároveň ciferné zápisy čísel byly odstraněny. Z pohledu procedur, které budou ČAK používat jako zdroj dat, se jedná o zásadní problém, který se nedá vyřešit automatickou procedurou, naopak ruční korektura dat je nutná; na ruční kontrolu ČAK byla odčerpána většina prostředků OON.

Nástroj pro lexikální anotaci V návrhu nástroje byly vyhodnoceny klady a zápory původního anotačního nástroje a rozpracovány aktuální požadavky.

Modifikace morfologického analyzátoru Podrobná analýza současného stavu, návrh struktury slovníkového editoru a návrh vzorů byly vypracovány dle plánu.

Nové metody tagování Hledání nových možností tagování bylo zahájeno revizí prostředků, které jsou k dispozici. Na dopracování HMM taggeru byla určena část prostředků OON.

Investiční prostředky byly odčerpány na nákup diskového pole a notebooku pro hlavního řešitele projektu. Z neinvestičních prostředků byla nakoupena odborná literatura a software (antivirové programy, licence). Vzhledem k tomu, že projekt byl zahájen až v druhé polovině roku, byly cestovní prostředky využity na předplacení cest na odborné akce roku 2005.

Spoluřešitelské pracoviště Ústav pro jazyk český AV ČR

Pracoviště spoluřešitele, Ústav pro jazyk český AV ČR bylo z účelových finančních prostředků dovybaveno pracovní stanicí PC a z týchž prostředků byl zajištěn její provoz i chod všech potřebných prací. K úkolům projektu byla zakoupena příslušná odborná literatura.

Byla provedena finální revize statistické části dat Českého akademického korpusu (ČAK, pův. název Korpus věcného stylu KVS), implementována propracována softwarová varianta uživatelského prostředí pro statistiky a doplněny prohlížeče revidovaných dat k dalšímu přímému využití pro účely kvantitativní lingvistiky. Automatické uživatelské zpřístupnění komentovaných kódů s plnou gramatickou a syntaktickou informací bylo rozšířeno o navigační propojení textových celků s názvy zdrojů. Na pracovním CD budou v dohledné době zpřístupněny všechny dosud disparátní datové části ČAK (texty, tagované texty, seznamy textů, soubory substantiv, adjektiv, sloves, elektronické verze jejich retrográdních seznamů atd.). Vznikne tak statisticky přímo využitelná paralela k jinak koncipovanému CD ČAK verze 0.25, zejména pro účely kvantitativní lingvistiky.

S perspektivou vybudování specifického informačního systému bylo zahájeno rozsáhlé skenování archivního celku lexikálních dat (tzv. I. vrstvy lexikálního archivu budované od r. 1911). Získaná data jsou určena k budoucímu zpřístupnění po elektronických sítích. Skenování se provádí formou dodavatelských zakázek po výběrové řízení.

Ve spolupráci se zahraničními pracovišti byly zahájeny práce na precizování koncepce trojjazyčného elektronického slovníku česko-anglicko-maďarského.

Z prostředků projektu byly jednak obnoveny, jednak navázány přímé konzultační kontakty s pracovišti, kde se řeší příbuzné problémy, zejm. v Budapešti a Paříži.