

## **Intercorp – schůzka paralelistů 21.2.2006**

**Přítomni:** Čermák F., Schmiedtová V., Vavřín M., Křen M., Rosen A., Blažek D., Boková E., Čermák P., Dovalil V., Gladková H., Haušildová K., Jindrová J., Jirásek K., Kalajdžievová N., Káňa T., Klégr A., Kolmanová S., Lindroos H., Malá M., Maroszová J., Nádvorníková O., Peloušková H., Rajnochová N., Schreierová A., Skoumalová H., Šaldová H., Tláškal J.

**Omluveni:** Adamičová A., Cinková S., Dybalská R., Hrnčířová Z., Nábělková M., Štichauer P., Vondříčka P., Zemánek P.

**Úvod:** prof. Čermák přivítal přítomné z paralelních korpusů

Projekt je na 7 let.

Minulý rok - věnován nákupu HW a SW

- ujasnění, kdo bude pracovat na projektu za jednotlivé jazyky
- začátek práce, skenování prvních textů

Čínština se zatím neozvala – nenavázala spolupráci

Tento rok – vybudování základu paralelní složky ČNK o rozsahu 500 000 slovních tvarů

- práce s daty
- dokoupení HW – pokud je potřeba
- SW – používá se ParaConc + Finereader pro scener
- flashdisk – nejlépe, když si každý koupí sám, přinese paragon a UCNK zajistí proplacení

Zdroje textů - texty majetkem UCNK nabídnuty jednotlivým jazykům

- idea: texty v UCNK – vhodné, aby se od nich odvíjejí překlady
- překlady z cizích jazyků do češtiny
- společné zdroje textů: dokumenty EU – nedostačují pro kvalitu jazyka, pro budoucí vytěžování korpusu není ideální, EUROLEX, Gutenberg, Český konzervační fond- fond státní knihovny, není veřejně dostupný – přístup pouze pro jednu osobu, která bude skenovat pro všechny jazyky

Finance – byla vyplacena odměna řešitelů týmů na konci roku 2005

- byly vyplaceny odměny za skenování a zarovnávání – texty musí být zpracovány

Propagace- možno přeložení dokumentů o korpusu do jednotlivých jazyků

- každý jazyk si musí přeložit sám
- prezentace VZ na konferencích, publikace

Součástí naší práce je metodologie paralelních korpusů – pro tento obor zatím není prozkoumána.

Publikační výstupy z výzkumu musí obsahovat číslo VZ (MSM: 0021620823).

Čeká nás workshop v roce 2009 – bude multijazyčný – otázka zda bude mezinárodní nebo pouze český, následně vyjde sborník – každý jazyk by měl mít příspěvek

Je možno pozvat významné hosty za jednotlivé jazyky: M. Scott – Liverpool, prof. Gaston – Francie

Zahraniční cesty: účast na konferenci – podmíněno příspěvkem

- pozvání významného hosta
- cesta do zahraničí v rámci kontaktování se zahraničními nakladateli

## **Diskuze:** problémy a zkušenosti

Vavřín: nárůst o 10% odměny za skenování – texty zpracované – hotové páry  
Text naskenovaný, opravený poslat Vavřínovi ve Wordu – připraví verzi pro ParaConc, pošle zpět a tyto texty se zpracují v programu ParaConc, zapsané nejdříve v databázi před skenováním, aby nedocházelo ke zdvojení – aby se neskenoval stejný text dvakrát

Odměny studentům – 3,10 Kč/ kB textu – naskenování českého textu – přečtení textu  
odstranění chyb

- 4,20 Kč/kB textu - pro cizojazyčný text
  - rozdělení při zajišťování skenování třetí osobou 0,55 Kč
  - formátování, kontrola odstavců, vymazání obrázků 0,35 Kč
  - korektura českého textu 2,20 Kč, cizojazyčného textu 3,30 Kč
- cena za korekce je rozdílná podle typu jazyka

Přehled umístění počítačů a skenerů: učebna UCNK

seminář slavistiky - 7  
polonistika -8  
romanistiky – po dohodě  
němčina – knihovna  
slovenština – pracovna

Jazyky, které nemají vlastní výbavu se mohou domluvit s vybavenými pracovišti, kdy by mohli skenovat.

Nádvorníková: Je možno skenovat pře sobotu a neděli? Je potřeba vyjednat přístup na fakultu?  
Shmiedtová – neměl by být problém, skener v učebně UCNK je přístupný přes kartu, kterou lze vypůjčit v UCNK

Čermák – najít vhodné studenty pro skenování je problém. Je lepší investovat do spolehlivých studentů.

Boková: Zda můžeme proplatit knihy, které se koupí v zahraničí?

Schmidtová: Lze to, zajistíme proplacení formou smlouvy s fakultou. Pokud bude v budoucnu potřeba nakoupit knihy, lepší bude, když se zeptá nejdříve v UCNK na postup.

Jirásek: Jak to funguje s národní knihovnou s konzervačním fondem, máme spousta cizojazyčných překladů české beletrie?

Čermák: Vyjednáno s knihovnou, ale knihy nelze odnést mimo budovu, musí se najít osoba, která bude skenovat pro všechny jazyky na vybavení knihovny. Seznam knih poslat Vavřínovi. Jiná alternativa, přivést knihy z ciziny a zase je vrátit.

Haušildová: Jak funguje spolupráce s EUROMEDIA?

Nádvorníková: Poskytovatel textů EUROMEDIA je nástupce ODEONU, Knižního klubu, Ikaru. Seznam překladů, které byly vydány z překladového jazyka poslat na adresu diestlerova, [petra@euromedia.cz](mailto:petra@euromedia.cz), - uvést, že se jedná o InterCorp. Omezení – jsou ochotni poskytovat elektronické verze 3 roky zpětně, je možno se domluvit na zapůjčení originálu v knižní podobě, vždy při jednání s překladatelem uvést, že UCNK má smlouvu s EUROMEDIA – proces je legální, je dobré jako výměnu nabídnout překladateli přístup do korpusu.

Možno rozšířit i na jiné nakladatelství, pokud má někdo nápad.

Čermák: Bylo by dobré prozkoumat Památník národního písemnictví - zdroje textů.

### **Nádvorníková – prezentace zahraniční cesty do Francie.**

Cíl - získat přístup k textům v elektronické podobě, které je možno zahrnout do paralelního korpusu

- získat kontakt na instituce, které by mohli poskytnout v elektronické podobě např. České centrum v Paříži, ambasáda, česko-francouzské instituce a spolky, univerzity a překladatele
- problémy: těžké je proniknout do nakladatelství – nejasnost s autorskými právy na internetu není možnost získat elektr. texty v důsledku autorských práv
  - další problém, jak vysvětlit: co chceme, o co nám jde
  - nejlepší je získat oficiální záštitu – ve Francii Česká ambasáda

Na začátku si přeložili prezentaci UCNK do francouzského jazyka.

Je dobré začít prezentaci o čem je korpus, jak funguje, podstatné zdůraznit, že každý řádek je jedna věta z jiného textu. Projekt Intercopr je součástí velkého již fungujícího projektu UCNK, který je pod záštitou UK fil. fak. a ministerstva školství, má své web. stránky

Velmi záleží na způsobu prezentace projektu – být co nejnázornější, je dobrá prezentace na notebooku.

3. fáze prezentace – představit korpus (co to je, jak funguje), pak přejít k paralelním korpusu
- zdůraznit praktické aspekty pro tvorbu slovníku, výzkum jazyka,
  - nepředpokládat znalost korpusu
  - vysvětlit způsob financování projektu – zdůraznit, že se nejedná o elektronickou knihovnu, vysvětlit způsob ochrany dat – dohoda o poskytování textu – web UCNK, omezení textů
  - co chceme: chceme elektr. texty, překlady – beletrie, odborné texty, interní materiály
  - oplátkou nabídnout – vstup do korpusu

Výsledky: jednání s nakladateli, kteří se zavázali poskytovat texty,

- kontakty na překladatele možno získat přes obec překladatelů prof. Pognane – přislíbil přístup do francouzského korpusu
- franc. bohemisté přislíbili poskytovat své vlastní překlady
- nabídka pro franc. profesory – pozvání přednášet do Prahy

Závěrem: cestu dobře předem připravit – naplánovat

Oficiální prezentace posílat poštou než e-mailem. Nenechat se odradit, vybavit se dostatečným propagačním materiálem.

Na stránkách ministerstva zahraničí je seznam spolků v zahraničí

Doporučení na konferenci v OLOMOUCI mladých lingvistů 15. – 17. května 2006

Otázky:

Do jaké míry jsou zpracované internetové zdroje z EU?

Zpracovává někdo návody, manuály vícejazyčné dostupné k elektronice?

Koordinace rozšiřování smluv s dalšími nakladatelstvími?

Má někdo manuál pro studenty, kteří spolupracují na projektu pro práci s texty, jak postupovat při skenování?

Čermák: Desky pro prezentaci korpusu je možno vyzvednout v UCNK + CD.

Bude se konat Mezinárodní veletrh knihy – možnost navázat spolupráci s nakladatelstvími.

Letos je věnováno lotyšskému jazyku.

Rosen: Zpracování internetových zdrojů z EU. V plánu využít volné zdroje, např. počítačové manuály - zpracovat pro všechny jazyky.

Hlavní zdroje – hotové korpusy – texty jsou již zarovnané:

Ústava Evropské unie

Systémová hlášení v systému KDE

Manuály jazyka PHP

Korpus Acquis Communautaire – právní texty pro členy EU

Multext-East – text morfologicky označovaný

### **Prezentace bulharské cesty: Kalajdzieva**

- překlad prezentace do bulharské jazyka

- filozofická fakulta v Sofii – bohemisty a bulharisty – získány seznamy českých překladů do bulharštiny

- bulharská akademie věd – institut paralelního zpracování dat – zpracovaný korpus - syntaktická struktura bulharského jazyka – zdroj elek. textů – jsou ochotni poskytnou publicistické texty

- sekce počítačové lingvistiky – 35 mil slov – obávají se autorských práv – v jednání, potřeba sepsat smlouvu

- nakladatelství – asociace bulharských nakladatelů – reference – zda vydávají překlady

- kontakt překladatelů – elektr. texty 3 roky nazpět

- pokud někteří překladatelé poskytnou text – chtějí mít zpětnou vazbu od žijícího českého autora, že souhlasí s poskytnutím textu

- elektronická knihovna – obsahuje překladovou literaturu – možnost stáhnout texty, není ale uvedeno o které vydání se jedná

Otázka: Zpracovávat manuály, které jsou přiložené v elektronice apod.?

Blažek: Pozor na obsah, často jde o neodborné překlady – jazykové mutace .

Schmidtová: při získávání el. textů možno využít přes internet e-booky, pokud bude mít někdo zájem, pomůžeme finančně.

Švédština a Makedonština si nechali naskenovat texty v Akademii věd, pokud bude mít někdo zájem využít tuto cestu, proplácíme fakturu.

Ruština: získala pro práci kolegyni, umí velmi dobře ruštinu a češtinu – velká spolehlivost při zpracování textu, chybovost nulová

Bulharština – používá skupina bohemistů a bulharistů – každý zpracovává svůj jazyk

Vavřín: důležité je seznámit studenty, že na začátku je práce velmi pracná – pokud získá praxi – práce se zrychlí.

Dotaz : Bude se aktualizovat Finereader – na verzi 8.00 – je přesnější o 30%.

Vavřín: Zvažíme, cena upgrade je dost vysoká.

Schmidtová: Shrnutí za jakými účely se dá cestovat z prostředků VZ:

- prezentace Interkorpu – nutné mít příspěvek, citovat číslo VZ

- cesta za texty, možnost vybavit finančně
- možnost pozvat hosta ze zahraničí – pomůžeme zajisti ubytování, diety
  - náplň si musí koordinátoři zorganizovat sami
- pokud budete cestovat do zahraničí – prosíme o její upozornění, využití při vyúčtování
- pokud budete cestovat vlastním autem – nutno zajistit školení řidičů.

Otázka: Zda zahrnout osobní překlady manuálů do korpusu?

Čermák: Záleží na zvážení jednotlivých koordinátorů. Korpus má smysl, jestliže má kontext, nemá význam pokud je tam pouze seznam

**Další schůzka: 16.6.2006 ve 13 hod.**