INTERCORP : základy, úvodní schůzka řešitelů 22.1.2005 (rozšířený zápis) FČ leden 05

0 Úvod (František Čermák, FČ)

- InterCorp je součástí sedmiletého VZ Český národní korpus a korpusy dalších jazyků (2005-2011), schválený byl s jistou finanční redukcí (č. MSM 0021620823)

- Seznámení účastníků úvodní mezi sebou a s pracovníky ČNK, dr. A. Rosen (zvl. software a příprava textů), dr. V. Schmiedtová (finance), Mgr Jan Kocek (skenování), FČ a další (za ÚČNK přibude ještě nový pracovník, pozn. 20.4.: od 1. května bude zaměstnán v ÚČNK - pošleme všem kontakt na něj)

- Zásada: každý jazyk má svého garanta (viz seznam přítomných, člen FF), který koordinuje práci jím vybíraných a hodnocených spolupracovníků, zvl. studentů

- Práce studentů bude honorována, budou odměny i pro garanty, vše podle odvedené práce

- Informace: rozdané materiály a společná adresa s informacemi: viz dál

1 Cíl: a-Vybudovat paralelní synchronní korpusy kolem 25 jazyků s češtinou ve středu, především už s hotovým základem v ČNK, společné textové jádro by mělo být co největší (třeba zkusit)

b-Srovnávací lingvistický (kontrastivní) výzkum mezi páry, popř. i mezi nimi napříč

příspěvek lingvistickým oborům FF a jejich výzkumu i učení

POZN: podle možností (právních, popř. technických) budou korpusy v různé míře zveřejněny, ideálně na webu

1.1 Metoda: -Korpusové srovnání jevu ve výchozím jazyce s druhým a jeho kvantitativní vyhodnocení (umožní odlišení typického od periferního), tertium comparationis bude vždy startovní jazyk, lze zkoumat ale i z druhé strany

- s ohledem na možnosti softwaru lze ale přistoupit k paralelnímu studiu a srovnání až 4 jazyků (budou-li texty stejné)

- velkou výhodou je bezprecedentní možnost studia **kontextu**, tedy typická a nová korpusová možnost

- **směr komparace** je jistě obojí, v praxi podle praktických možností (textových), třeba nalézt metodu vyhodnocení výsledku, půjde-li o 3. jazyk (nemusí být tak spolehlivý)

POZN: - Jde o skutečný výzkum ve všech fázích a je to tedy terra incognita, řešení a recepty nejsou, budou se teprve hledat, takto velký projekt v Evropě zatím není, budeme tedy mj. „prošlapávat cestu“

- Základní výstupy musejí být teoretické, je tu ale bohaté pole možností dalších (diplomky, pedagogické, lexikografické...)

- Bude aspoň zčásti nabídnuta publikační báze (viz dál), vždy však třeba uvádět č.VZ výše

-Relevantní a spolehlivé výsledky lze získat až z většího a vyváženého objemu dat, resp. textů, takže čím více tím lépe, aneb pravda je ve velkých číslech a různosti, kterou je třeba zprůměrovat

-**Společné** textové jádro (zatím to co je v ČNK) lze rozšiřovat (třeba návrhy) a lze zařadit i texty **překladové,** které může mít více jazyků (zvl. např. texty z angl., především nebeletristické), v části **nespolečné** (nejádrové) se textově bude nutně vycházet z toho, co bylo přeloženo specificky

-Texty **2 žánrů**: -beletrie z 2. pol. 20. stol. (**poválečné**)

-**společenskovědní** a přírodovědní (zvl. ty budou asi muset někdy čerpat z 3. jazyka), popř. i jiné **odborné**

1.2 Aspekty textové a výzkumné

A-**Dva druhy činnosti** při tvorbě korpusů:

- **Skenování** (včetně oprav) vedoucí k výslednému formátu *.txt* (tj. prostý text)

- **Alignování** (resp. zarovnávání): ruční a pracné zajištění srovnatelné podoby téhož textu v obou jazycích na úroveň **identity odstavců** (včetně jejich sjednocování n. dělení podle potřeby a vyčištění nadbytečného)

POZN: -Text je třeba získat v el. podobě (lze-li, je to jednodušší) n. naskenovat na skeneru (budou nakoupeny podle požadavků týmů včetně pg, zřejmě obvykle FineReader, orientální jazyky ale jinak)

-Chce-li někdo z jiných důvodů si zachovat vedle původní text či překlad, nic mu nebrání, pro analýzu par. korpusovou ale nebude použitelný

-Pracovat se bude jednoduše s čistým textem (*.txt*, který lze dosáhnout exportem z čehokoliv), bude-li však zvláštní důvod, lze texty obohatit o značkování XML, popř. i lingvistické taggování (event. i lemmatizace), čímž však vyvstávají nové problémy nekompatibility (viz pak dr. Rosen, tagger a lemmatizátor cizího jazyka si musí ale každý opatřit zvlášť)

Pozn. 20.4.: V současné době se pracuje na doporučení, jakým způsobem a jakými nástroji by měly být texty před zpracováním označkovány (identifikace textu, hranice odstavců, vět).

- Seznam textů k mání v ČNK bude na webu, viz *ucnk.ff.cuni.cz/beletrie.html*, viz ale texty obecné na webu (Evropská ústava, Deklarace lid. práv aj.)

Je třeba upozornit na to, že budou proplaceny hotové, tj alignované dvojice textů, ne jednotlivá díla!!!

B-**Výzkum:** -(zatím/hlavně) na ParaConcu (omezená a drahá multicence), dostane každý jazykový tým

Pozn. 20.4.: V nejbližší době budou pořízeny licence a bude možné si ParaConc instalovat.

-Pro každý hledaný jev lze vytvořit konkordanci (i uložitelnou zvlášť) jako základ studia jevu v kontextu včetně

-Možností statistického vyhodnocení, distribuce jevu (u víc textů), tvorby textových slovníků ap.

-Vyvíjí se však webový nástroj.

-Výzkum (a výstupy) se očekává až zhruba od 2.-3. roku VZ dál, až bude minimální množství textů

C-**Plánované výstupy** (musejí se každoročně vykazovat), zvl. studie a články

Při zanášení do bibliografických hlášení je nutné uvádět číslo našeho Výzkumného záměru (viz nahoře).

- Rozsah korpusových dat- viz dodaný materiál po rocích (lze jistě rozšířit) do r. 2010 je soubor korpusů o rozsahu min. 2,5 milionu slov/tvarů

-2009: (mezinárodní) tématický workshop a 2010 publikace sborníku z něj

-2011: konference o vytěžování paralelních korpusů (sborník třeba uvážit)

Publikační možnosti další: - kdo si co zajistí ve svém oboru a jeho časopisech, konferenčních sbornících aj.

- na webovských stránkách ÚČNK

POZN: Předpokládá se účast všech zapojených jazyk. týmů, pozvání zahraničních hostů třeba vážit podle financí

1.3 Aspekty organizační

-Odpovědnost

-ÚČNK: dr. Rosen a nový pracovník (zřejmě od dubna), adresa *Alexandr.Rosen@ff.cuni.cz*

Pozn. 20.4.: Nový pracovník nastoupí v květnu.

-jazyky: jednotliví garanti

-Třeba si vést výkaz (elektronický, např. WORD) o vykonané, popř. svěřené práci

-Zaškolovat podle potřeby bude na skenování (ale po skupinách) Mgr. Jan Kocek (*Jan.Kocek@ff.cuni.cz*), v práci se softwarem dr. A. Rosen, třeba se s nimi domluvit.

-Texty zatím přebírá dr. Rosen, ten také proti podpisu distribuce softwaru.

-Prezentace výzkumu, zvl. v cizím jazyce: lze zveřejnit, ale odpovídá (po dohodě s ÚČNK) vždy garant

1.4 Aspekty archivační

-Zatím každý tým zajistí **vlastní** evidenci a uchování korpusů

-vznikne **databáze centrální** v ÚČNK (nový pracovník) a **centrální uložiště** korpusů (paralelní s týmovým)

tam třeba vše předávat s průvodní dokumentací, tabulkou apod. (text třeba doprovodit patřičnostmi jako jméno autora, překladatele, místo a rok vydání, rok překladu, popř. žánr apod.). Potřeba vše zálohovat.

1.5 Aspekty právní (zvl. nečeských textů)

-Vzniklé korpusy jsou majetkem tohoto VZ, nelze je dávat komukoliv (neplést si se zpřístupněním), bude třeba zřejmě na to mít smlouvu

-Dostupnost, resp. zveřejnitelnost textů závisí na legislativě toho kterého státu a je třeba zjistit (odpovědnost garantů)

Obecně platí, že naskenovat si pro sebe může každý cokoliv, ale nelze s tím obecně už nakládat; ta omezení je třeba zjistit a podle toho se ne/zveřejnění korpusů musí řídit (obvyklý copyright je dnes 80 let v Evropě), a to na škále zcela volný (pak bude na webu) - různě vázaný (přístup udělovaný na povolení jen někomu) - zcela vázaný (nezveřejnitelný). V posledním případě bude třeba sondovat aspoň možnost využívat v onom malém týmů tvůrců kolem daného garanta (navenek pak korpus nebude tedy existovat).

-ÚČNK své texty (všechny) získává od dodavatelů n. agentur na základě právním, podepsané smlouvy, viz formulář na web. stránkách ÚČNK (*http://ucnk.ff.cuni.cz*, tam *Dohody a registrace*), lze příp. napodobit

1.6 Aspekty finanční

-Odhady: na VZ zřejmě jsou na 1 rok peníze v rozsahu 940 000 Kč/rok, tj. na cca 500 průměrných knih kompletně zpracovaných a předaných (viz 1.2A), lze dát cca 1500 Kč za 1 takový text (ale je věc dohody a financí), to obnáší

tedy finance na skenování (dosavadní zkušenost ÚČNK jen za skenování je ale je 1500-1700 Kč za průměrný román cca 300 stran) i pracné alignování: je třeba postupovat u ceny pragmaticky též podle toho, za kolik peněz ten který student bude ochotný věc dělat. Vyšroubuje-li se každá jednotlivá složka nahoru, bude počet knih pak bohužel menší a výsledek jak menší tak výzkum z něj vzdálenejší. Pracovat se pak bude do vyčerpání peněz, je třeba to vyzkoušet.

-Rozumný výsledek za 1 rok je 10-15 textů (cca románů) v obou jazycích, a to především pro malý jazyk, větší týmy a jazyky více

-Odměna koordinátorovi

-Možnost získání skeneru, popř. PC, popř. OCR programu, popř. i jiné drobné vybavení (dr. Schmiedtová)

-Jsou k dispozici finance na cca 10 střídmých zahran. cest (konference s relevantní tématikou, ev. kvůli textům)

-Existuje navíc i možnost si požádat o avl. grant na cestu (viz ing Rosén)

-Budou finance aspoň na 1 konferenci a sborník

2. **Demonstrace programu ParaConc** (Alexandr Rosen, AR)

2.1. **Základní funkce programu ParaConc** (viz i stručný rozdávaný popis v angličtině):

ParaConc je určen primárně pro jednoho uživatele (2-5 jazyků) a v druhé fázi projektu se počítá s přechodem na korpusový manažer, který bude fungovat podobně jako systém využívaný pro hledání v Českém národním korpusu, s možností přidávání a údržby textů. Jeho aspekty:

-Jednoduché hledání (ukázka hledání v česko-anglickém zarovnaném souboru textů z časopisu Reader's Digest a románu 1984 George Orwella)

-Možnosti zobrazování výsledků hledání, využití funkce Hot Words pro zvýraznění možných ekvivalentů hledaného výrazu ve druhém jazyce

-Využití statistických funkcí programu pro vyhledání kolokací hledaného výrazu

-Základní postup při zpracování nových paralelních textů: uložení v textovém formátu, zarovnávací funkce programu, včetně rozdělování a spojování odstavců a členění delších textů na kratší úseky, ukázka zarovnávání úryvku z románu Thomase Hardyho Neblahý Juda

-V nejbližší době bude k dispozici jednoduchý návod pro práci s ParaConkem, včetně typického postupu pro zpracování nových textů

(Pozn. 20.4.: viz http://utkl.ff.cuni.cz/~rosen/INTERCORP/paraconc.pdf, nebo přes „Dokumenty“ na stránkách projektu)

2.2 **Z diskuse a k ní**

-Licenční podmínky - odpověď: program bude pořízen z prostředků projektu pro každé zúčastněné pracoviště

demo verze vystavená na internetových stránkách (viz adresa níže) neobsahuje všechny funkce, je zastaralá

POZN*:* v současnosti už obsahuje, omezení: zobrazí se max. 150 vyhledaných výskytů a výsledky nelze ukládat

-Jak zařídit, aby se při dělení a spojování odstavců během zarovnávání zachovalo původní členění textů – odpověď: bude připraven návrh

-Jaké kódování ParaConc vyžaduje – odpověď: lze pracovat v jednobytovém kódu MS Windows, pro každý jazyk zvlášť, nebo kódu Unicode (UTF-8), je třeba mít nainstalováno příslušné národní prostředí v operačním systému

-Kde lze získat příručku k programu – odpověď: http://www.athel.com/para.html, tam je k dispozici i demo verze

Ke zkopírování je 1 exemplář v ÚČNK k dispozici

2.3 **Jiné dotazy a poznámky**

-Měla by existovat stránka projektu nebo jiné materiály, na které by se dalo odkazovat (podmínka podpory ze strany jiných institucí) – odpověď: zvážíme to, pak by byla v rámci ČNK, mj. aby se na ni dalo odkazovat

-Nejsou-li texty zarovnány po větách, je využití výsledků hledání velmi obtížné (zkušenost s některými literárními díly, např. od Franze Kafky) – odpověď: zarovnávání po větách je obtížný úkol a v rámci projektu se bude považovat za „nepovinné“ (FČ), lze využít automatické metody a při zarovnávání se spokojit s jistým procentem chyb (AR)

-Kontakty mezi účastníky projektu: stačí poslat mail na adresu intercorp@ff.cuni.cz nebo paralel-korpus@ff.cuni.cz a mail se rozešle všem účastníkům projektu