Projekt ***Intercorp***

Cílem projektu *Intercorp* je vytvoření paralelních, tj. o překlady opřených korpusů mezi češtinou a všemi velkými jazyky evropskými i většinou tzv. malých, a to na základě dat Českého národního korpusu a ve spojení s ním. Zahrne zejména angličtinu, němčinu, francouzštinu, ruštinu, slovenštinu, španělštinu a italštinu. Z dalších jazyků se takto má pokrýt i arabština, bulharština, dánština, finština, litevština, maďarština, makedonština, norština, polština, portugalština, slovinština, srbština, švédština, popř. jazyky další, jako např. japonština a čínština (zajišťovat je jako garant bude vždy příslušný expert z jednotlivých jazykových pracovišť FF UK). Budou sloužit v mnoha směrech, jak pro vlastní výzkum, především kontrastivní, tak výstavbu a zdokonalování slovníků, ale i pro praktickou výuku na fakultách ve velké řadě jazykových oborů. Jednotlivé hlavní a vzájemně propojené úkoly budou zahrnovat především tyto složky:

(a) postupná tvorba jednotlivých paralelních korpusů zahrnující vytvoření celkové i jazykově specifické strategie, řešení otázek copyrightu, formy katalogizace, evidence a propojení s ČNK, školení a počítačovou podporu partnerů, koordinaci týmu a podtýmů ap.,

(b) postupné získávání textů (v rozsahu aspoň cca 400 000 – 1 000 000 slov, podle dostupnosti, kulturní významnosti jazyka a objemu dostupných překladových dat), jejich katalogizace, získání a testování softwaru, určeného v první fázi aspoň pro bilaterální zpracování apod.,

(c) zpracování textů, zvláště náročné alignování a v některých případech i nižší úroveň značkování (podle možnosti a smysluplnosti), softwarový servis jednotlivým týmům, další vyvažování a celková koordinace uvedené dvacítky týmů pro jednotlivé jazyky, extrakce ekvivalentů,

(d) návazný výzkum (až v pozdější etapě, po vzniku dostatečně rozsáhlých korpusů), a to podle aktuálnosti témat a potřeb zastoupených oborů včetně publikování jeho výsledků, mj. na workshopu a sbornících,

(e) zveřejnění paralelních korpusů (podle možností a stupně uvolnění ze strany poskytovatelů), především pro studijní účely a aspoň pro interní přístup, případně na webu, bude-li do té doby k dispozici vhodný software, který je třeba vyvinout.

Strategie a metody sběru a zpracování

Bude v zásadě vycházet z přístupů dříve osvědčených, které se však budou nutně podle potřeby modifikovat a rozšiřovat, mj. heuristicky na základě zpřesňování kritérií, dílčího testování výsledků aj. Sběr dat (elektronických), tradičně náročný, vychází z kontaktů se stovkami poskytovatelů dat, podchycení takového vztahu smluvně a vlastního přebírání a zpracování dat technického (náročné konverze z množství formátů, čištění, unifikace aj.) a lingvistického (analýza typu každého textu, jeho zařazení a označení v tzv. hlavičce a katalogizace); nověji se zkoušejí i metody další, zvl. Internet.

Texty především diachronní a cizojazyčné je třeba pracně skenovat, opravovat a pak dále zpracovávat. Jistou spolupráci tu lze očekávat i od ÚJČ AV ČR.

U paralelních korpusů bude třeba vypracovat jednotnou koordinaci i metodologii optimálního sběru dat (mj. ve srovnání se zkušenostmi zahraničními) i metodologii korpusového konfrontačního výzkumu, kde budou dominovat mj. aspekty statistické.

Testování softwaru bude probíhat zvláště ve spolupráci s externě spolupracujícími pracovišti a pracovníky. Vlastní potřeby z analýzy dat povedou ke zpětné vazbě na partnery, zvláště v podobě požadavků na další vývoj a modifikaci softwaru.

Návazný výzkum se zaměří na vytipování aktuálních témat vhodných podle oborů i širší potřeby, navržení optimální metodologie pro každý výstup (jiná bude u studií gramatických, jiná u lexikálně orientovaných ap.), obecně však vždy s důrazem na nové možnosti zachytit především syntagmatické aspekty jevu v plné šíři, což bývalo až dosud, v předkorpusové době, nemožné.

Forma zveřejňování bude ve více podobách, u korpusů v rámci ČNK v intervalech zvláště na internetu (jinak vzhledem k rozsahu není možné). Cílem je výsledky podle možností vždy zpřístupnit širšímu okruhu odborných uživatelů, který zvláště v případě projektu Intercorp může být poměrně velký (jazykové katedry univerzit ap.). U studií půjde jak o publikace sborníkové, články, popř. CD, někdy i jako výsledek předchozího workshopu či konference.

Datové výstupy:

2005–2007:
primárně tvorba paralelních korpusů všech jazyků různého rozsahu

2008–2010:
další a prohloubené získávání dat a další výstupy

2011:
podle možností zpřístupnění na internetu ve vytvořeném/získaném softwaru

Harmonogram:

2005:
vypracování koncepce a vývoj programového vybavení pro zpracování a elementární využívání paralelních korpusů

2006:
vybudování základu paralelní složky ČNK o rozsahu 500 000 slovních tvarů (tj. přibližně 200 000 párů česko-jinojazyčných překladových ekvivalentů)

2007:
rozšíření paralelní složky ČNK na minimálně 1 000 000 slovních tvarů

2008:
rozšíření paralelní složky ČNK na minimálně 1 500 000 slovních tvarů

2009:
rozšíření paralelní složky ČNK na minimálně 2 000 000 slovních tvarů

2010:
rozšíření paralelní složky ČNK na minimálně 2 500 000 slovních tvarů

2011:
internet