Charles University, Faculty of Arts
Institute of the Czech National Corpus
nam. Jana Palacha 2, 116 36 Prague 1, Czech Republic
tel.: +420 2 21 619 357, ucnk@ff.cuni.cz

Projekt

Český národní korpus a korpusy dalších jazyků,

vedený jako výzkumný záměr č. 0021620823, financovaný Ministerstvem školství, mládeže a tělovýchovy ČR a realizovaný na Univerzitě Karlově v Praze, Filozofická fakulta UK, odpovědný řešitel prof. PhDr. František Čermák, DrSc.

se skládá z několika podprojektů, mezi které patří

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 v 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 zarovnávání (alignování) a v některých případech i nižší úroveň značkování (podle možnosti a smysluplnosti), softwarový servis jednotlivým týmům, další vyvažování a celková koordinace uvedených 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 ve 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.