František Čermák
Ústav Českého národního korpusu, Karlova Univerzita v Praze

InterCorp: jeho povaha a možnosti.

1. Korpus: Jednojazyčný a vícejazyčný.

Ten protiklad se může zdát normální, ale není. Záleží totiž na hledisku, které zvolíme. A necháme-li stranou aspekty historické a technické, je jasné, že o protiklad vlastně nejde. Ten první krok je totiž nutnou podmínkou pro vznik toho druhého. Bez více napřed jednojazyčných korpusů vybavených k sobě jinojazyčnou protiváhou, jsou paralelní korpusy pochopitelně nemyslitelné. Nicméně ještě před tím, než se pustíme do vlastního výkladu o InterCorpu, na místě snad bude pár obecných úvah, které je třeba mít na zřeteli.

Potřeba odstupu při úvaze o vlastním jazyku a projektu, který se ho týká, je zdravá, ba nutná. S určitým odstupem lze nejen získat určitou objektivnost, ale snad i dospět k jistým zobecněním, která jsou možná jen díky srovnávání s více jazyky. Lidská komunita zpravidla nikdy nežije v úplné izolaci a vždy jí obklopují jiné, které obvykle mluví jazyky jinými, jak blízkými jako je vztah komunity české a slovenské, či nikoliv, a jen v etymologicky dávné minulosti se mohlo stát, jako v případě polštiny, slovenštiny či češtiny, že se další soused metonymicky nazval v tomto vztahu *Němec, Niemec* či *Nemec* a jednoduše se mu tak upřel jazyk jakýkoliv, protože ho naši předkové měli za němého, jelikož vydával jim nesrozumitelné zvuky. Dneska víme už trochu víc: nesrozumitelné zvuky vydávají všichni mluvčí všech jiných jazyků, pokud se je nenaučíme, což vyžaduje určitou námahu. Avšak od respektování jazykové odlišnosti je jen krok ke (1) **srovnávání jazyků**, tj. k hledání podobností a odlišností, zvláště máme-li k takovému srovnání potřebu, ať už profesionální jako lingvisté, či čistě praktickou, dorozumívací. Je zřejmé, že první takové srovnání bývá vždy napřed bilingvní, jak nám napovídají vůbec první slovníky, které vznikaly mezi mluvčími jazyků v pravidelném styku, zvláště u jazyků sousedících (začátky je třeba hledat v Mezopotámii). Srovnávání jazyků však přineslo a přináší takové množství poznatků a vhledů, které mj. zdravě krotí přebujelý nacionalismus, že dalo vznik řadě oblastí zahrnujících jak na jedné straně praktickou lexikografii, tak na straně druhé typologii, univerzálie a obecnou lingvistiku vůbec. V praxi však toto poznávání mělo své jasné meze: bylo jen málo lidí, kteří kvalifikovaně znali více jazyků, jakkoliv znát pár sousedních jazyků se odjakživa chápalo za výhodu a dnes dokonce za slušnost, bylo ale především málo dat, z nichž se dalo vycházet, stejně tak jako jen málo opravdu multilingvních mluvčích, kteří by byli multilingvního výzkumu schopni. To ale poněkud předbíháme, protože obrovská většina srovnávání byla dvoujazyčná, a z něho bylo jak obtížné tak nebezpečné dělat širší závěry o čemkoliv.

Až příchod (2) **korpusů** dramaticky situaci změnil a dnešní velké korpusy jako je Český národní korpus (přes 3 miliardy slov, ČNK), ještě větší německý v Mannheimu, ale i slovenský a další prokázaly za relativně velmi krátkou dobu svou nezastupitelnou užitečnost pro veřejnost i jazykovědu. Dobrým příkladem je nezbytnost korpusu například pro probíhající tvorbu současného velkého slovenského slovníku. Je třeba nekompromisně trvat na tom, že každou informaci je možné a nutné hledat napřed v textech, a to díky kontextům, ve kterých se slovo vyskytuje. Jen z nich lze pak dospívat k potřebným a prokazatelným generalizacím. Korpusy se takto nabízenou informací maximálně podobají realnému světu komunikace kolem nás a jsou tak ve skutečnosti jeho nejlepším záznamem i aproximací, kterou máme. Ve srovnání s minulostí stará předkorpusová lingvistika nikdy neměla dost dat, resp. dost dat a jejich kontextů. Obecně lingvisticky pak je zřejmé, že v korpusovém pohledu kontexty, které jsou vytvářeny z kombinací slov, nás reorientují ze starého metodologického postoje prvek a jeho pozice (item-and-slot), resp. člen a jeho třída, tj. přístupu paradigmatického, k tolik potřebnému přístupu syntagmatickému, založenému na kombinacích a jejích typech.

V případě masívního nárůstu korpusů a možností, které dává, však neplatí, rozhlédneme-li se poněkud, že s jídlem roste chuť, jen se tak znovuobjevuje už i zmíněná potřeba podpořit starou tradici srovnávání jazyků pro studium nejrůznějších souvislostí z něj vyplývajících. Takové studium bez rostoucího (3) **paralelního korpusu**, v našem případě specificky slovensko-českého a česko-slovenského korpusu, by vůbec nemohlo ani začít. Kombinuje jak možnost srovnávání tak výhody korpusu (1 a 2), není však jen jejich pouhou sumou. Je jasné, že zvýšený důraz, lingvistický i politický, na takovéto aktivity se dobře hodí do rámce toho, čemu se dnes říká globalizace, která by však neměla končit u pouhých proklamací, v lingvistice může mít velmi reálnou a dokonce potřebnou podobu.

Paralelní korpusy dnes existují pro mnohé páry jazyků, jakkoliv technologie jejich výstavby nikterak jednotná není, tím méně je takové i jejich skutečné využití. Od dob, kdy paralelní skutečně znamenalo bilingvní, se však pokročilo dále, zvláště díky čistě technologickým možnostem. Od těch dob, kdy jediným skutečně paralelním korpusem byla bible, odhlédneme-li od jejích různých překladů, se dnes pokročilo (přes mezifázi anglicko-

francouzského parlamentního *Hansardu*) ke zdánlivě velkolepým možnostem nabízeným evropskými multilingválními soubory jako je *Europarl* (debaty evropského parlamentu, 11 jazyků, průměrně 50 mil. slov na jazyk, http://www.statmt.org/europarl/) nebo *The JRC-Acquis Multilingual Parallel Corpus* (22 jazyků, 636 milionů slov, zákonodárství EU, tzv. acquis communautaire, http://langtech.jrc.it/JRC-Acquis.html). Jejich obsahová stránka je však značně limitovaná a tím i odvozené možnosti praktického využití lingvistického.

Je tedy zřejmé, že nejde jen o prosté počty slov v dostupných překladech, ale především o *obsah textů*, resp. o to, jak velkou a jak obecnou část jazyka paralelní korpusy zachycují, a tedy o to, jaké jsou cíle tvorby paralelního korpusu a, řečeno ještě jinak, jaké jsou *potřeby ale i možnosti* takového počínání. Odhlédneme-li od úzkého právnického využití, které nám oba velké evropské korpusy nabízejí, je zřejmě, že je třeba vážit skutečné potřeby, které zdaleka nejsou jen právnické. Dnešní situaci paralelních korpusů charakterizují především dvě věci. Na jedné straně existuje relativně hodně dvoujazyčných, v zásadě oportunních korpusů pro řadu jazyků, které jsou navzájem většinou inkompatibilní, na druhé straně vznikla už paralelně a vícekrát aspoň elementární technologie k jejich prohledávání. Komputační lingvisté se tomuto úkolu věnovali a věnují však jen do té míry, do jaké lze dělat věci automaticky (především pro strojový překlad), a to včetně metod alignmentu (zarovnávání) textů; stále však tu řada potřeb zůstává nepokrytá (viz Čermák-Rosén, v tisku). V návaznosti na to tedy zůstává především na straně lingvistů úkol paralelní data shromažďovat a začít je využívat především v těch aspektech, které dosud realizovatelné nebyly. *Cíl srovnávat více jazyků* je v dnešní multilingvální Evropě velmi smysluplný, nejde však jen o příznivé politické ovzduší. Takové srovnávání může a musí v širším smyslu především naplnit a ospravedlnit staré diktum, že *jazyk je prostředek přenosu významu od myšlenky k formě*. A to se dá snadno doplnit o diktum další, totiž že *jazykové srovnávání je i mostem umožňujícím přenos významu i mezi jazyky* navzájem.

Paralelní korpusy jsou pochopitelně možné jen tam a tehdy, jsou-li dostupná podkladová, tj. překladová data mezi jazyky. Dostupnost dat je však často problém, který se nedá nijak řešit, pokud neexistují. Tento problém, který je pro jednojazyčný korpus jedním z mnoha, se v tomto případě stává zcela zásadní a primární. Pro ty jazyky, které nemají možnost využívat existence společného fondu překladů z krásné literatury (tj. překladů z nich nebo do nich), ba ani svou roli v mezinárodním kontextu (například jako jeden z oficiálních jazyků EU), se tato překážka stává zásadní natolik, že nebezpečně omezuje další růst paralelního korpusu, protože příslušná data prostě neexistují.

Cílem našeho příspěvku není věnovat se technickým aspektům výstavby paralelních korpusů, jakkoliv jsou hojné a netriviální, protože jsou vždy novátorské a realizované pro daný jazyk poprvé, a to jak v oblasti segmentace vět, tokenizace, zarovnávání textů (alignment), tak i lingvistické anotace. Každý jazyk tu vyžaduje specifické nástroje vyvinuté jen pro něj. V dalším se zaměříme proto především na obecnější aspekty netechnické.

Můžeme si pak snadno v dané situaci představit, že vznikající disciplína *srovnávací korpusová lingvistika* tu může najít dostatečnou a zásadní vzpruhu, pokud se multilingvní korpusy dostatečně rozrostou, budou usilovat o přijatelnou míru reprezentativnosti a zaměří svůj výzkum na aspekty skutečně multilingvální. Protože každé srovnání potřebuje své *tertium comparationis*, je zřejmé, že si lingvisté pak musejí být jisti tím, že srovnávání v takovém širším měřítku musí nalézt i společný širší metodologický rámec, který by měl být typologický. Ale realizace je tu dosud daleko.

2. Jazykové kontakty a překlad. Česká jazyková situace.

Skutečnost, že jak bilingvní tak multilingvní korpusy jsou v zásadě podmíněny **jazykovými kontakty** a zakládají se tedy na dostupných překladech mezi jazyky a že jejich počet roste jen postupně, má své důvody i důsledky. Z kulturního a historického hlediska představuje úhrn dostupných překladů z jednoho jazyka do druhého sumu nejrůznějších nitek i proudů zájmu, ať už podmíněného dobově (jako v případě módních románů) či reálných a užitečných, které daná komunita měla a má po dané časové období, vázaných na komunitu jinou a její texty. To je hned nápadné, srovnáme-li takovou sumu přeloženého mezi dvěma malými jazyky, kam se dostalo do centra pozornosti často leccos užitečného či zajímavého, čeho si překladatelé všimli. Pokusíme-li se zobecnit, zdá se, že platí, že velikost průniku dostupných přeložených textů u více jazyků je nepřímo úměrná množství těchto jazyků; jinými slovy *počet textů sdílených mnoha jazyky se úměrně počtu jazyků snižuje*.

Takto lze nahlížet i kulturní, politické a další vlivy mezi komunitami, pokud studujeme počet, typ a rozšíření překladů mezi nimi v jejich úhrnu, a to nejen pro jeden jazyk a etnikum za ním, ale i pro komunitu větší a multilingvální, jakou je třeba Evropa. Jakkoliv existuje mnoho různých typů překladu *z* (a *do*) velkého jazyka (=*zdrojový jazyk*), jsou recipienty překladů ve většině případů, řečeno zjednodušeně, malé jazyky, tj. ty, do kterých se texty překládají (=*cílový jazyk*). To se promítá nepřímo i do skladby paralelních korpusů.

Za dané geopolitické situace se většina pozornosti, až na několik výjimek, upírá k paralelním korpusům, které se orientují na páry složené ze dvou velkých jazyků (jako je angličtina a francouzština v Hansard Corpus) či na takové páry, ve kterých je aspoň jeden z jazyků velký, jako je angličtina. Díky rozšířenosti angličtiny a některých dalších jazyků je ale také jasné, že páry dvou malých jazyků v tomto pohledu dost stádají. Přitom v kontrastu k všední praxi a jejím potřebám reálné lingvistické požadavky ukazují jinam, k potřebě srovnávání ve velkém měřítku a kvalifikovanějšímu studiu (všech druhů) jazyků i všech druhů textů. Proto je nutné, aby se rozumně shromažďovala srovnatelná data z co největšího počtu jazyků.

To platí i o **českém jazyce**. Je to slovanský jazyk, kterým mluví 10 milionů lidí, tedy jeden z těch malých jazyků. Jako typický flektivní jazyk má rysy, které se jen stěží najdou v angličtině, francouzštině, němčině či čínštině, jako je bohatá flexe o sedmi pádech, slovesný vid, volný slovosled, bohatá verbální prefixace, bohatá derivace substantiv, desubstantivní adjektiva (typ *vlakový*), množství partikulí aj., jakkoliv většinu těchto rysů má společných, ne však identických s ostatními slovanskými jazyky. Historicky jím mluví lidé ve středu Evropy, kde čeština byla vždycky jazykem na rozcestí v důsledku vlivu jiných, mezi něž patří především němčina či polština a slovenština a na druhé straně po několik desetiletí dočasně i nesousední ruština.

Čeština měla tradičně dva druhy těsných jazykových kontaktů se svými sousedy, slovanskými na jedné straně (slovenštinou a polštinou) a německým (rakouskou a německou němčinou), a oba představují nutnost ve výzkumu věnovat pozornost velmi odlišným problémům. Z nich představuje zvláště velmi dlouhý kontakt s němčinou nejen nutnost, ale i možnost, kterou lze využít zajímavěji, půjdeme-li hlouběji, za pouhé výpůjčky, totiž do sémantiky, kalků a vlivů na gramatický systém.

Všechny tyto faktory měly a mají svůj vliv, který se promítá do češtiny, v níž tento zajímavý vlivový souběh nabízí předmět jak specifického výzkumu tak výzkumu obecného, především v typologickém rámci. Ten, obohacený o pohled zvenku, by měl být zajímavý už šíře a nejen pro české mluvčí a lingvisty, jakkoliv sem patří dnes i nejnovější vliv globální angličtiny. Odtud tedy představa a pozadí mnohojazyčného korpusu, majícího češtinu uprostřed mezi koncentricky navázanými dalšími jazyky, a tedy představa **InterCorpu**. Dodejme hned, že takto výlučné postavení češtiny nikterak neupozaďuje žádný další z participujících jazyků, které tu lze zkoumat také, dokonce i s vyloučením češtiny.

3. Projekt InterCorp a jeho povaha.

Na rozdíl od jiných projektů (viz Čermák, v tisku, a Čermák-Rosén, v tisku) je otevřený a usiluje stále o pokračující růst, kdekoliv je to možné, tj. všude tam, kde jsou k mání dostupné texty a finance k tomu potřebné. Jeho hlavní filozofie je stejná jako u velkého jednojazyčného korpusu ČNK: *v zásadě čím víc dat tím lépe.* A protože česká data dostupná už jsou a byla k mání o něco dříve, je v zásadě třeba získávat jen nečeské překladové texty, ať už se najdou hotové (v elektronické podobě), nebo se naskenují a dodá se jim potřebná úprava; dnes se však podle potřeby získávají i některé texty, kde nová, resp. nově naskenovaná může být i čeština a zveřejněné jsou až v podobě v InterCorpu.

Seznam a počet jazyků vstupujících do InterCorpu je stále pragmaticky otevřený, tj. jediným ohledem je dostupnost textů; proto je i řada textů, které, protože jsou vždy teprve v procesu zpracování a čekají na své plné začlenění do systému a zveřejnění. Je jasné, že každý jazykový pár je odlišný (kromě společné češtiny) jak co do rozsahu tak obsahu. Ukázalo se totiž, že původní představa, že existují překladové texty společné většině, ne-li všem relevantním jazykům, se dosud nepotvrdila; můžou však existovat texty, které dosud nebylo možné získat a zařadit.

Takový je tedy **obecný cíl** výstavby korpusu InterCorp i vlastní implementace projektu, který ho rámcuje. **Politika jeho výstavby** je jednoduchá a snad i skromná:

(**1**) Shromažďují se jen *současné texty*, vymezené tak, že nemají sahat před rok 1945 (i když můžou zahrnovat i texty starší, jsou-li vydané znovu po tomto datu). Tato časová hranice je stanovena vědomě: kromě klasické literatury začíná skutečná četba textů a tedy i současného jazykového úzu zhruba právě tady. Je to jinými slovy rozhodnutí nejen čistě praktické, ale i způsob, jak specifikovat distinkci synchronie-diachronie. Takováto hranice však pro některé jiné budovatele korpusů se nezdá zřejmě důležitá. Je to otevřený a obtížně řešitelný problém, který záleží na konkrétní situaci v jednotlivých jazycích, který je dobře vidět i tehdy, když zdrojový text je starší než překlad, který mohl vzniknout až po roce 1945, resp. jde o nový poválečný překlad. V takovém případě je zřejmě nejlepší prosté pragmatické řešení, které respektuje kvalitu a povahu textu.

(**2**) Jakkoliv by bylo ideální řešení dosáhnout u každého jazykového páru jistý druh *rovnováhy* co do počtů výchozích textů na každé straně, zůstává to z pragmatických důvodů zatím jen zbožným přáním a není ani jasné, má-li se o něj důsledně usilovat za každou cenu, tj. např. o stejný počet překladů *z* a *do* češtiny. I proto se to dosud nestalo kritériem určujícím výstavbu InterCorpu. Přesto může mít stejný počet původních textů na obou stranách páru pro jisté cíle své zřejmé přednosti.

(**3**) Kvůli očividnému nedostatku *textů sdílených více jazyky* bylo rozhodnuto, že do InterCorpu se začlení i některé texty, jejichž původním jazykem není ani jeden z jazyků daného páru, což je případ zvláště malých a nesousedních jazyků. Takové „třetí" texty pak pocházejí zpravidla ze široce překládaných jazyků. Takto má povahu takového *třetího jazyka* např. v případě dostupného česko-srbského subkorpusu (na podzim 2011) 6 z 15 titulů v česko-srbské části, kdy jde zvl. o texty z angličtiny, ale i italštiny, polštiny, portugalštiny a ruštiny. Obecnou zásadou je mít přehled o častěji opakovaných překladech do různých jazyků, vybírat z něj a zajistit tak co nejširší vazbu na co nejvíce jazyků. Takto se dává přednost titulům překládaným do více jazyků. Toto rozhodnutí, tj. připustit v některých případech existenci neoriginálního jazyka na obou stranách jazykového páru, je třeba pak brát v úvahu při těch rozborech korespondence, kde na tom může záležet; obecně ho však odmítnout nelze, je to nouzový prostředek tam, kde jiná, přímá cesta není. Techniku a kritéria vyhodnocování relevantnosti tohoto druhu *nepřímé ekvivalence* při začlenění třetího jazyka je třeba teprve hledat, především v kontrastu k ekvivalenci přímé.

(**4**) *InterCorp* se snaží být lingvisticky co nejobecnější, aby mohl sloužit mnoha *různým cílům*: lingvistickým, nelingvistickým, akademickým, praktickým, výuce překladu aj. Proto je mj. důležité jím zachytit co nejvíce různých druhů a typů jazyka a lexikonu. Je však třeba připomenout, že vybudovat vyvážený paralelní korpus (dvoujazyčný i vícejazyčný) je mnohem těžší než korpus jednojazyčný. Důvody jsou aspoň čtyři.

(**a**) Některé textové typy, ale většina mluvených textů se překládají jen málokdy; sem patří většina případů novinářského jazyka, který je naopak tak důležitý v jednojazyčných korpusech. To je také důvod pro pragmatické řešení soustředit se na to, co dostupné je: proto se *InterCorp* skládá výlučně z textů *psaných*, jakkoliv, aspoň zatím teoreticky, nejsou mluvené texty vyloučené.

(**b**) Pokud jde o *prózu neliterární* a její převažující žánr, *žurnalistiku*, je k dispozici multilingvální zdroj (*Project Syndicate*, viz už výše, je to mezinárodní asociace novin vydávající komentáře a analýzy od předních autorů ovlivňujících veřejné mínění) a objevuje se i další slibný kandidát v tomto směru, *Presseeurope* (http://www.presseurop.eu), což je portál monitorující přední evropské deníky, který se v současnosti překládá do 10 jazyků včetně češtiny.

(**c**) Realizují se postupně snahy začlenit do InterCorpu více typů textu, zvláště z oblasti *specifického jazyka* parlamentních diskusí EU (*Europarl* http://www.europarl.europa.eu/), dokumenty zákonů (*EUR-Lex* http://eur-lex.europa.eu, *JRC-ACQUIS Multilingual Parallel Corpus* http://wt.jrc.it/lt/Acquis/), o kterých už byla řeč také výše, či dále různé otevřené zdroje technické literatury a softwarových manuálů (jako je *OPUS, Open Source Parallel Corpus* http://urd.let.rug.nl/tiedeman/OPUS/) atd., jakkoliv jde o jazyk jen úzce zaměřený.

Volba takovýchto textů je jen pragmatická a záleží na jejich (A) existenci, (B) dostupnosti a (C) *legálních předpisech*, které jejich dostupnost regulují. To jsou však obecné otázky týkající se všech druhů textů. Rozhodně ale platí, že uživatel korpusu si vždy může svobodně vybrat určité texty a studovat či užívat je podle svých potřeb a zájmu a jen podle nutnosti může být jeho přístup omezen, například dalším heslem ap. Kvůli pragmatické povaze projektu InterCorp je obtížné plánovat nějakou definitivní podobu korpusu, který se do značné míry stále mění, resp. roste.

(**d**) V InterCorpu však kvantitativně převládá *nespecifický jazyk,* především beletristický, který se vnímá jako priorita zaměřená primárně na maximálně pokrytí základního lexikonu, pokud je dostupný, tj. toho, který je důležitější a obecnější než zmíněné specifické typy jazyka, protože je v komunikaci univerzální.

Takovéto teoretické a praktické ohledy lze tedy hledat za ideou velkého multilingválního korpusu s češtinou v centru. *InterCorp* (http://korpus.cz/intercorp) je součástí širšího projektu *Český národní korpus* (ČNK, http://korpus.cz). Vlastní centrální představa v jádru InterCorpu je tedy lingvisticky triviální, jakkoliv se neozývá často; mít vlastní jazyk bohatě pokrytý jednojazyčným korpusem, tj. zevnitř, nemusí stačit: *jazyk se musí také studovat zvenčí, prizmatem druhých jazyků.*

Tento projekt je nepochybně jedinečný co do svého rozsahu, volby textů (jakkoliv dosud převládá beletrie), ale i podstatného vkladu manuální práce mnoha lidí (při narůstající kvalitě zarovnání, identifikace hranic věty a

zmenšování počtu chyb). Účastníci projektu, pozvaní ke spolupráci v r. 2005 do týmu vedeného ústavem *Českého národního korpusu FFUK*, pocházejí z většiny jazykových kateder, ústavů a jejich oddělení filozofické fakulty Univerzity Karlovy v Praze a několika dalších akademických institucí, mezi které patří i spolupracovníci ze Slovenska, Polska aj., patří k nim však v neposlední řadě také početní studentští pomocníci. Současný stav zpracovávaných jazyků je 25 (plus čeština), online zveřejněných jazyků je 23 (a tedy už víc než má acquis communautaire). V současnosti se ke zpracování i studiu užívá paralelní konkordanční program PARK (=*par*alelní *k*orpus, http://korpus.cz/Park), vyvinutý a stále zlepšovaný na FFUK. K InterCorpu se může zdarma přihlásit každý registrovaný uživatel ČNK (http://korpus.cz/english/dohody.php).

Níže uvedená **tabulka** nabízí v přehledu čísla pro jednotlivé jazyky platné pro současnou verzi korpusu (čísla pro češtinu jsou vysoká proto, že se česká data v jednotlivých jazykových párech opakují). „Titulem" se zde míní především román jako dominantní typ textu, který je v InterCorpu zastoupený nejvíc. Některé jazyky však mají výhodu větší vyváženosti textů, která je zřejmější u většího počtu textů.

Písmeno *S* (± Syndicate) v jednom sloupci upozorňuje na počet titulů pocházejících z politických komentářů korpusu *Project Syndicate* (viz výše). V současnosti jeho dostupná vydání obsahují data česká, anglická, francouzská, německá, ruská a španělská z let 2000-2008, k nimž brzo přibudou další z nových vydání, zvláště arabská a čínská; čísla z tohoto korpusu jsou započítána do celkových čísel a jejich rozsah se pohybuje mezi 1,5-2 milióny slov pro daný jazyk. Celkový přehled stavu InterCorpu tedy v současnosti představuje data z března 2010 (čísla jsou udána v tisícovkách):

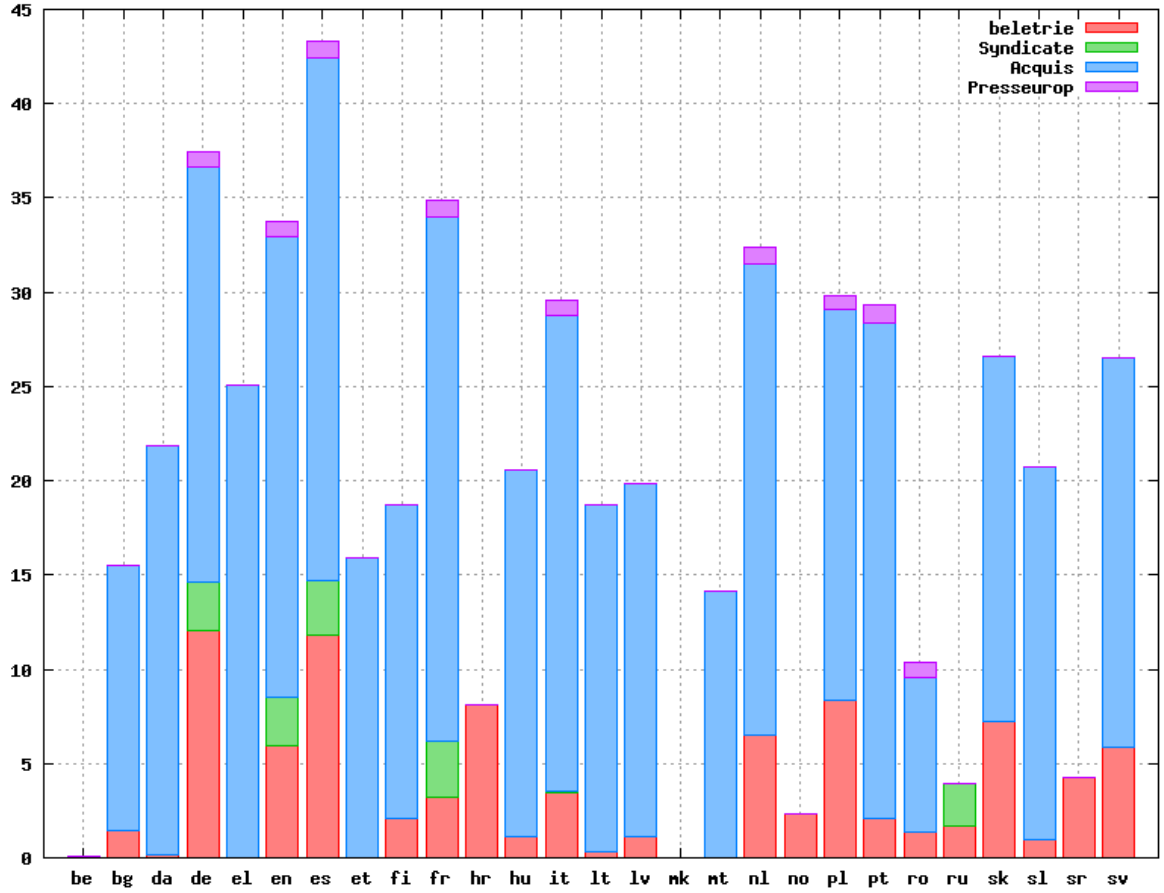
Jazyk (L2)	**Slovní tokeny v češtině** x 1000	**Tokeny v jiném jazyce (L2)** x 1000	**Počet titulů**
angličtina	4,041	4,705	S + 34
bulharština	1,057	1,049	14
dánština	80	102	4
finština	497	423	11
francouzština	2,415	3,120	S + 21
chorvatština	4,363	4,599	69
italština	2,254	2,591	26
maďarština	1,030	985	15
němčina	6,466	7,480	S + 70
nizozemština	2,448	2,046	45
litevština	318	272	7
lotyština	1,121	1,067	23
polština	2,450	2,422	40
portugalština	1,261	1,436	18
rumunština	461	564	4
ruština	2,873	2,902	S + 23
slovenština	352	351	7
slovinština	813	901	15
srbština	1,129	1,209	19
španělština	7,210	8,427	S + 82
švédština	1,439	1,643	25
Celkem	**44,077**	**49,293**	**572**

Každý z jazykových párů je zjevně odlišný, liší se rozsahem i obsahem, což je, jak se ukazuje, přirozené a odpovídá to skutečné distribuci překladů v různých jazycích. Podle stavu dat z loňského roku je nejčastěji zastoupený titul M. Kundery *Nesnesitelná lehkost bytí* (v 9 jazycích včetně češtiny) a dalších pár (aspoň 7) je ve zpracování. Kunderův román *Žert* je zastoupený v 18 případech a po nich následuje J.K. Rowlingová se svým *Harrym Potterem a kamenem mudrců* (*Harry Potter and the Philosopher's Stone*, 14x) a J.R.R. Tolkienův *Pán prstenů* (*Lord of the Rings*, 12x). I tady jsou některé verze dosud ve stadiu zpracování a zpřístupní se časem.

K aspoň stručné demonstraci poměrně značného pokroku od jara 2010 (viz tabulku výše) si uveďme aspoň zběžně novější (plná) čísla reprezentující stav pro 22 jazyků mimo češtinu (stojící v pozadí) z května 2012 (opět v tisícovkách slov/tokenů, včetně neliterárních textů); uvádějí se napřed počty slov z vlastních knižních textů a v závorce, uvozené pomocí *vč A*-, všechny počty slov všech textů, tj. včetně textů z Acquis communautaire, jejichž kvalita i hodnota je ovšem omezená, resp. nižší.

angličtina 9 229 (vč A-35 031), bulharština 1 466 (vč A-16 356,) dánština 190 (vč A-23 362), finština 2 140 (vč A-20 054), francouzština 6 939 (vč A-34 682), chorvatština 8 207 (vč A-8 207), italština 4 374 (vč A-30 297), litevština 358 (vč A-20 445), lotyština 1 116 (vč A-21 526), maďarština 1 249 (vč A-21 926), němčina 16 355 (vč A-39 998), nizozemština 7 426 (vč A-33 857), norština 2 335 (vč A-2 335), polština 9 264 (vč A-31 554), portugalština 2 959 (vč A-30 220), rumunština 2 162 (vč A-10 852), ruština 4 162 (vč A-4 162), slovenština 7 408 (vč A-28 270), slovinština 1 015 (vč A-22 161), srbština 4 351(vč A-4 351), španělština 15 832 (vč A-45 198), švédština 5 997 (vč A-27 971).

Graficky zachycuje vnitřní poměry čtyř obecných typů textů následující přehled pro 25 jazyků (včetně některých, dosud periferních „nových“):

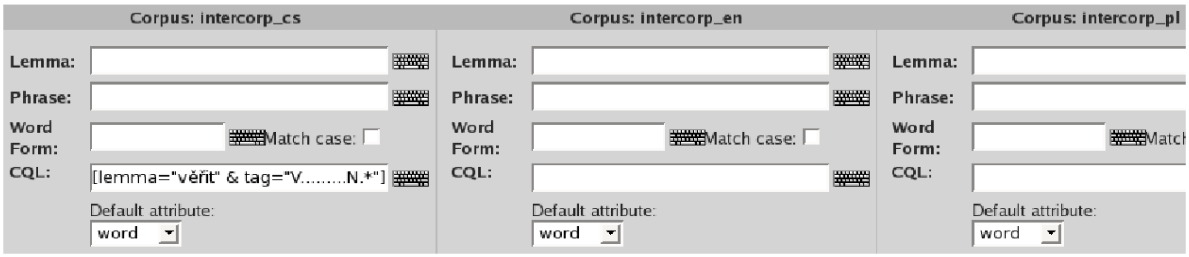


4. Užití InterCorpu, studium jeho výsledků a důsledky.

InterCorp se rozvíjí postupně dál, a to týká nejen nárůstu a povahy jeho textů, ale i obslužného softwaru a dodávané lingvistické informace; bibliografické údaje jsou uloženy ve zvláštní databázi (http://korpus.cz/intercorp/?req=page:info). Software PARK (autorem je M. Štourač), fungující na pozadí korpusového manažeru Manatee (P. Rychlý) umožňuje prohledávání více jazyků zároveň a tolik textů, kolik je třeba: volí si je badatel podle nabídnutého seznamu dostupných paralelních textů, a to buď jen v jednom směru, např. slovensko-českém, směru opačném či bez rozdílu a směru. Hledat je možné, jako v každém větším korpusu a jeho manažeru, tvary i lemmata (lemmatizovaná a morfologicky značkovaná je však zatím jen část textů), popř. i kombinace či výsledky zadané výrazem v CQL (pokud je k dispozici značkování), kdy se ve výsledku zadaná forma barevně označí jen ve zdrojovém jazyce a v cílovém jazyce či jazycích (může jich být víc) se objeví jen automaticky vybraný odpovídající kontext, v němž uživatel musí sám rozhodnout v nabídnutém kontextu, co považuje za ekvivalent; výsledky je možné prohlížet v konkordanci horizontálně nebo vertikálně a listovat v nich po stránkách. Podle potřeby lze nabídnutý automatický výsledek rozšířit na potřebný větší kontext. Výsledky lze ukládat do formátu spreadsheetu (Excel), kde je lze podle potřeby i řadit.

Dotazovací tabulka, která se objeví po zadání jazyků a textů, s nimiž chceme pracovat, ukazuje příklad (ve starší verzi dat InterCorpu) na hledání českého tvaru verba *věřit* (viz zadané *word*) v korpusu anglickém a polském (zadání se pro ilustraci zde uvádí i jazyce CQL). Ekvivalenty se hledají, opět pro ilustraci, jen v Kunderově

románu *Nesnesitelná lehkost bytí* a *Žert* a jeho ekvivalentech (překladech) do angličtiny, polštiny a ruštiny. Počet cílových jazyků lze ovšem libovolně zvětšovat. Srov.



Z výsledků se zde uvádějí pouze první tři nálezy. Počet tokenů (tvarů) zaznamenává software a uvádí nahoře za jménem korpusu (např. intercorp_pl (166365 tokens) uvádí, že v polském odpovídajícím subkorpusu se našlo 166365 odpovídajících, resp. kandidátních tvarů). Tabulka uvádí konkordanci v horizontálním zobrazení s výsledky ve sloupcích vedle sebe (lze zvolit i vertikální zobrazení ukazující každý jazyk na zvláštním řádku pod sebou).

intercorp_cs *(160320 tokens)*	intercorp_en *(192071 tokens)*	intercorp_pl *(166365 tokens)*	intercorp_ru *(162671 tokens)*
Show options Kwic			
Nevěřit (ustavičně a systematicky , bez chvile zaváhání) si vyžaduje obrovského úsilí a také tréninku , to jest častých policejních výslechů .	show context Maintaining non-belief (constantly , systematically , without the slightest vacillation) requires a tremendous effort and the proper training - in other words , frequent police interrogations .	show context Nieufność (ciągła i systematyczna , bez chwili wahania) wymaga ogromnego wysiłku i treningu , to znaczy częstych przesłuchań policyjnych .	show context Неверие (постоянное и систематическое , без тени колебания) требует колоссального усилия и тренировки , иными словами , частых полицейских допросов .
Nebylo možno nevěřit jeho upřímnému hlasu .	show context There was no doubting that forthright voice of his .	show context Nie można było nie wierzyć szczerości jego głosu .	show context В искренности его голоса сомневаться было нельзя .
Řekla jim , že to ví , ale že nevěřila , že by soudruh Jahn ...	show context She said yes , she knew , but she would never have believed that Comrade Jahn ...	show context Powiedziała , że wie , ale że nie wierzyła , żeby towarzysz Jahn ... Spytali , czy dobrze mnie zna .	show context Она сказала им , что знает , но не могла бы поверить , что товарищ Ян ...

Je nicméně důležité připomenout, že InterCorp lze nastavit libovolně jako *jakoukoliv konstelaci jazykových párů* či trojic, čtveřic, pětic jazyků aj., v nichž ani nemusí být explicitně uvedená čeština a výsledky v ní. Jednotlivé jazyky lze dokonce "odpoutat" od ostatních a studovat je přímo v rámci větších možností, které nabízí manažer Bonito.

Uveďme si pár dalších příkladů, tentokráte už přímo lingvisticky zaměřených, které se pokusí ukázat jak možnosti hledání tak problémy, které nalezené výsledky můžou přinášet. Podívejme se na **lexikální ekvivalenci**, a to napřed v podobě velmi jednoduchého výchozího českého lemmatu *stůl*, tedy slova v podstatě monosémního a na jeho ekvivalenty v angličtině a italštině, a to na M. Kunderově textu *Nesmrtelnost* a J.K. Rowlingové *Harry Potter and the Philosopher's Stone* (H.P. a kámen mudrců), kdy jeden výchozí text je v češtině a druhý v angličtině. Ilustruje se tu případ typické **jedno-víceznačné ekvivalence.**

Ke všem 100 výskytům českého lemmatu *stůl* dostáváme v angličtině 89 ekvivalentů (tj. 89%) v podobě *table*, 5 jako *desk*, 1 jako *desktop* a 5 je případů, kdy se nenabízí *žádný ekvivalent* (tj. 5%). Italské výsledky jsou pestřejší a nabízejí 83 ekvivalentů v podobě základního *tavola* (83%), 3 jako *banco*, dvakrát *cattedra*, ale taky jednou *banchetto, scrivania* a *scrittoio*, zatímco *žádný ekvivalent* se nedal najít v 9 případech (9%). Na první pohled se výsledky můžou zdát poměrně jednoduché a přímočaré, ale nápadně vysoké procento žádných ekvivalentů, které je v italštině vyšší než v angličtině, nutí k zamyšlení. K nalezení odpovědi se podívejme na dva příklady, které můžou naznačovat dvě rozdílné možnosti. Prvním je česko-anglický případ založený na předpokladu, resp. implikaci (operace se provádějí na *stole*, a proto se *stůl* vůbec nemusí zmiňovat); druhý česko/anglicko-italský případ je vědomé vynechání anglického *table*, jakkoliv bezprostřední kontext k tomu nenabízí žádnou záminku.

Uveďme si konkrétní příklady s kontexty.

CZ *při nějaké nevinné operaci zemřela na operačním **stole** mladá pacientka kvůli nedbale provedenému uspání*
ENG *a young woman who in the course of a completely minor operation died because of carelessly administered anaesthetic*

CZ *Hagrid se k němu naklonil přes **stůl**.*
ENG *Hagrid leaned across the **table**.*
IT *Hagrid si chinò verso di lui.*

Nicméně důležitější je zřejmě se zaměřit na **rozmanitost ekvivalentů** k prostému českému *stolu*, které se tu nabízejí, totiž na 3 pozitivní ekvivalenty v angličtině a 6 v italštině. Zaměříme-li se v důsledku jemnějšího studia jejich kontextů na rozdíly v jejich zřejmě **komplementární distribuci**, máme před sebou velmi úrodnou půdu, z níž lze mj. čerpat možnosti vylepšování ne vždy vyhovujících slovníků, máme-li se omezit jen na zcela praktické využití. Taková možnost i potřeba je zjevná už z toho faktu, že většina italských ekvivalentů s velmi nízkou frekvencí výskytu se často nedá ve slovníku najít.

Výše už bylo naznačeno, že zkoumání paralelního korpusu **blízkých jazyků** má zvláštní cenu, pohlédneme-li na něj poněkud jemněji, jakoby z hloubky. Takovým případem jsou i čeština a slovenština, kde jen na rovině lexikální

(1) vedle *identifikace jasných rozdílů*, tj. většinou formálně jasných (banální *borůvky-čučoriedky*, polysémní *les-hora*, či jemněji a jen fonologicky *zájem-záujem* aj.), hrozí i
(2) *bagatelizace jemných* a často opomíjených *rozdílů*, jakási nebezpečná supergeneralizace jen zdánlivě jasných vzájemných korespondencí.

Takovým případem je do velmi vysoké míry korespondence polyfunkční formy *ale-ale* v obou jazycích (většinou konjunkce, často ale i partikule). Nahlédneme-li do česko-slovenského korpusu (majícího v současnosti už pozoruhodný počet 132 textů) a podíváme se jen letmo, na pár příkladech několika textů, na slovenské korespondence českého *ale*, uvidíme, že obrovská většina slovenských ekvivalentů českého *ale* tradiční intuitivní představu jednoduchého vztahu podporuje, a tedy také *ale*. Nicméně pár příkladů nás musí znepokojit. Jednak jde o rozdíl ve variabilitě slovenských ekvivalentů *lenže, alebo, hoci, no* aj., které nejsou zase tak okrajové, a jednak o specifické odchylky v partikulární funkci této formy, srov. *Ale co vás to napadá!* a slovenský ekvivalent *Čo vám to zišlo na um!*, kde se zdá, že další možnosti ve slovenštině blokuje zároveň to, že *Čo vám to zišlo na um!* je v tomto smyslu neměnný frazém vyjadřující integrálně modalitu výchozího českého *ale*.

Takové příklady jsou jen malou ilustrací, která však chce naznačit, že je třeba je důkladně zkoumat. Lze se pak nadát, že dosavadní slovníky mezi oběma jazyky, které nejsou nikterak optimální, nahradí slovníky nové, kde se už k datům v InterCorpu bude přihlížet a dospěje se k lepším ekvivalentům s jasnou informací o jejich distribuci.

Aspoň jednu výše připomínanou možnost, kterou **multilingválnost** přináší, si přiblížíme na příkladu, do kterého se výrazně promítá už i zmíněná **typologie jazyků**. Na příkladech z devíti jazyků (na datech Kunderova *Žertu*) si lze ukázat, jak se v těchto jazycích tvoří substantivní pojmenovánání několika typů. Jde o jazyky někdy typologicky vyhraněnější (*flektivní* čeština, polština a ruština a *aglutinační* finština, a sledovaným rysem výrazně i němčina a norština) i méně vyhraněné (smíšené, zvl. francouzština a španělština), zvlášť stojí *izolační* angličtina. I když všechny jazyky jsou typologicky smíšené (obv. s převažujícím jedním rysem), pro jednoduchost se tu kromě flexe, aglutinace a izolativnosti další rysy neuvádějí (srov. VAR), i když např. španělština je z jiného hlediska flektivní taky aj. Srov.

		Prosté	**Derivace**	**Kompozice**	**Kolokace**	**Kolokace**
FL	**češ**	*hodiny*	*hodinky*	*0/běžící pás*	*nákladní auto*	*toaletní stolek*
	pol	*zegar*	*zegarek*	*0/ruchoma taśma*	*0/ciężarówka*	*mały stół, toaletka*
	ruš	*časy*	*0/časy*	*0/konvejer*	*0/ gruzovik*	*tualetnyj stolik*
AGL	**fin**	*kello*	*0/kello*	*liukuhihna*	*0/kuormaauto*	*pieni pöytä*
	něm	*Uhr*	*(Armband)Uhr*	*Fließband*	*0/Lastauto*	*0/Toilettetisch*
	nor	*ur*	*0/ur*	*transportbåndet*	*0/lastebil*	*0/toalettbord*
IZOL	**angl**	*clock*	*0/watch*	*0/production line*	*0/truck*	*small table*

VAR	**špan**	*reloj*	*0/reloj*	*0/silla del peluquero*	*0/camion*	*mesa pequeña*
	fr	*montre*	*0/pendule*	*0/chaîne*	*(voiture de tourisme)*	*petite table*

První dva sloupce uvádějí napřed případ prostý, resp. nederivovaný lexém (i když v češtině je to díky plurálu poněkud méně jasné) a pak lexém derivovaný. Kromě flektivní češtiny a polštiny se všude jinde neliší *hodiny* a *hodinky*. Třetí až pátý sloupec ukazují konkurenci kolokací a kompozit, zase ale jen v neflektivních jazycích (tam je všude jen kolokace); ruština tu není typická. Nicméně je jasně vidět souvislost pojmenování pomocí kolokace a izolačního typu v angličtině, i když i zde je jednoslovná výjimka (*truck*); výjimky lze však nalézt i u francouzštiny a španělštiny a jejich povahu a status může zpřesnit jen obsáhlejší analýza.

I zde musíme odhlédnout od překladatelských idiosynkratičností, kde např. ve fr. je chybné *voiture de tourisme*, nebo tam, kde překladatel neuvádí existující možnost (opět ve fr. *table de chevet*, popř. *coiffeuse*) a spokojí se s prostým deskriptivním *petite table* aj.

V širším smyslu takováto ilustrace ukazuje také, *jaký typ výsledků* jednoho druhu se pro jednotlivé jazyky dá z výzkumu čekat.

5. Výzkum a jeho možnosti.

InterCorp může být a už i je užitečným zdrojem poznání, o čemž svědčí i první publikované výsledky; některé nalezly své místo v publikovaných sbornících na konferencích 2009 a 2011 (Čermák-Klégr-Corness, Čermák-Kocek, Čermák).

I když tu jde obecně o mnoho možností, výzkum multilingválního korpusu lze zjednodušeně chápat jako dvojího druhu (**A**) aplikovaný a (**B**) teoretický.

Aplikovaný výzkum (**A**) bude záviset na skutečné poptávce a mohl by být tradičně propojený hlavně s překladovými studiemi a lexikografií (Teubert 2001, 2007). Specificky zajímavou možností se tu jeví studium problémů interpretace téhož textu ve více různých překladech, pokud budou k dispozici. Každý překlad je třeba chápat jako idiosynkratický mj. v tom, že zachycuje vždy jen část významu výchozího textu, čímž dospíváme ke staronové otázce, co se vlastně v překladu obvykle či vždycky ztrácí aj.

Jakkoliv se multilingvální lexikografie momentálně velké popularitě netěší (na rozdíl od terminologie, srov. *Eurodicautom,* resp. *IATE*), situace se může změnit. Uveďme jen, že by např. bylo užitečné mít slovník blízce příbuzných jazyků jako je čeština, slovenština a polština, skandinávských či jižních románských jazyků, které by se daly užívat pro kontrolu falešných přátel, resp. mezijazykových homonym aj.

Praktický význam lze rozhodně hledat i v oblasti strojového překladu, automatického text-mining, automatické disambiguace aj.

Teoretický výzkum (**B**) může v prohloubeném srovnávání jazyků nabídnout vhledy i do oblastí nových či jen málo zkoumaných. Nicméně jde především o **srovnávací korpusovou lingvistiku**, kterou je třeba začít pěstovat, kde multilingvální korpus může přijít k užitku tím, že nabídne lepší data obecné lingvistice, typologii, pragmatice a minimálně i studiu diskurzu.

V takovémto rámci se může objevit řada obecných témat a otázek. Takto si žádá např. stará a dosud jen obecně formulovaná otázka *jazykové příbuznosti* přesnějšího poznání a nutnosti pracovat jak s menšími tak většími skupinami jazyků. Naproti tomu stojí zdánlivě nekonečná *různost* zvl. nepříbuzných jazyků, až dosud pokrývaná v podstatě jen typologií a univerzáliemi, která by si zasloužila jakékoliv smysluplné zpřesnění, zvláště např. v podloženém navržení *typologie rozdílů*, pokud je to vůbec možné.

Silnou stránkou tradičního většinového studia monolingvních korpusů je vždy nesporná opora v autentických textech a skutečných kontextech. Naproti tomu však studium bilingvních a multilingvních korpusů je odlišné v tom, že překlady prostě nejsou originální, autentické texty (a tedy vlastně ani ne přeložené kontexty, o čemž se málo mluví). Nabízí se tedy potřeba *metodologie hodnocení překladových protějšků*.

Je zřejmé, že s pohybem vzhůru, od prostých lexikálních jednotek skrze kolokace ke větám a jejich kombinacím se hodnota každého takového postupu musí nutně stávat problematičtější a čím dál interpretačně otevřenější, ba dokonce někdy spornější. Nicméně uvědomíme-li si, že základem a *východiskem všude musí zůstat význam*, zdá

se, že velmi zajímavé výsledky se mají hledat mj. spíše v oněch vyšších rovinách než v těch nižších, jako je slovo. Máme-li k dispozici paralelní korpus či korpusy nabízející množství kontextů a bohatou varietu ekvivalentů hledaného prvku na škále, která se dá statisticky vyhodnocovat, máme k dispozici prostředky k dosažení mnohem lepších výsledků, než jaké dosud přinášel starý a manuální přístup, založený na nesystematických a často i podivných příkladech s problematickým zobecněním.

6. Závěr.

Je zřejmé, že možnosti, které paralelní korpusy pro komparativní studium jazyků nabízejí, se vlastně teprve otevírají, a na hodnocení toho, co teprve má přijít a přijde, jak lze doufat, je příliš brzo. Zdá se, že by mohlo být užitečné tu znovu připomenout, že daná oblast paralelních korpusů, jejich výzkumu a desiderata z toho vyplývající už jednou, aspoň zčásti shrnuta byla. Připomeňme si proto hlavní závěry z panelové diskuse, nazvané *Final Panel Discussion of the 2009 InterCorp Conference in Prague* (Čermák, Klégr, Corness 2009). Dospěli k nim po živé diskusi účastníci konference InterCorpu, v níž se vedle hlavních a tíživých otázek a problémů začíná rysovat i pár odpovědí

1 Úloha třetího jazyka v bilingvních korpusech: míra a metodologie
Názory a ohlasy: Třetí jazyk je nepostradatelný tehdy, není-li počet překladových textů v jazykové dvojici příliš velký. Míra jeho zastoupení se nezdá tak důležitá. Je však třeba rozlišovat co do jeho relevantnosti mezi jazykem originálního textu a překladem; konečné výsledky by se měly ověřovat proti velkému vyváženému jednojazyčnému korpusu.

2 Vyvažování dvou jazyků v paralelním korpusu
Názory a ohlasy: Je žádoucí, jakkoliv rovnováhu v počtu dostupných textů můžou narušovat pragmatické faktory.

3 Společné textové jádro pro více jazyků
Názory a ohlasy: Jako možnost je jistě žádoucí, i když dopředu ve skutečnosti nevíme, kolik uživatelů by tento rys umožňující srovnávání více než dvou jazyků mohlo užívat; lze ale snadno najít faktory, které mluví ve prospěch této možnosti.

4 Legální problémy vztahující se ke copyrightu a vlastnictví textů
Názory a ohlasy: Nemějme obavy. Žádný korpusový lingvista se ještě nedostal před soud za porušení copyrightu tím, že dal do korpusu nějaký text. Sběr paralelních korpusů by se neměl zastavit kvůli legálním formalitám; vždyť přece vždycky existuje prostředek umožňující jen omezený přístup a bude-li třeba, i přístup zaheslovaný; ani praxe dělení textů do vzorků či přeřazování částí textu v jejich sledu se nezdá užitečná.

5 Kritický počet slov či rozsah paralelního korpusu pro praktické účely, specificky v lexikografii.
Názory a ohlasy: Rozsah záleží na cíli, kdy např. v (bilingvní) lexikografii usilující o 20 tisíc lemmat je třeba miliónů slov korpusu.

6 Různé
Návrhy: paralelní korpus by měl pokud možno zahrnovat více textových typů (tj. vedle beletrie a odborných textů). Paralelní korpus je užitečný v jazykové výuce.

Bibliografie

Barlow, M. (1992). „Using Concordance Software in Language Teaching and Research“. In Shinjo, W. et al. *Proceedings of the Second International Conference on Foreign Language Education and Technology*. Kasugai, Japan: LLAJ & IALL.

Barlow, M. (2000). „Parallel texts in linguistic analysis“. In M. Barlow and S. Kemmer (eds.) Usage-based models of language. In Botley, S. P., T. McEnery, A. Wilson (eds.), *Multilingual Corpora in Teaching and Research*, Amsterdam, Rodopi, 106-115.

Barlow, M., (2002). „ParaConc: Concordance software for multilingual parallel corpora.“ // Language Resources for Translation Work and Research, *LREC 2002,* 20–24.

Botley, S., A. McEnery & A. Wilson, eds. (2000). *Multilingual Corpora: Teaching and Research*. Amsterdam: Rodopi.

Čermák, F., ed. (2011) *Korpusová lingvistika Praha 2011 I InterCorp NLN Praha*

Čermák, F., Klégr, A., & Corness, P. (Eds.). 2010. *InterCorp: Exploring a Multilingual Corpus*. Praha:

Nakladatelství Lidové noviny.
Čermák, F. Kocek, J. (Eds.). 2010. *Mnohojazyčný korpus InterCorp: Možnosti studia*. Praha: Nakladatelství Lidové noviny.
Čermák, F., Rosén A., *The Case of InterCorp, a multilingual parallel corpus* (in print in International Journal of Corpus Linguistics)
Gage, W. W. (1961). *Contrastive Studies in Linguistics: A Bibliographical Checklist*. Washington D.C.: Center for Applied Linguistics.
Hammer, J. H., F. A. Rice (1965). *A bibliography of contrastive linguistics*. Washington, DC: Center for Applied Linguistics.
Johansson, S. (2007). *Seeing through Multilingual Corpora; On the use of corpora in contrastive studies*. Studies in Corpus Linguistics. Benjamins.
Melamed, Dan I. (2001). *Empirical Methods for Exploiting Parallel Texts.* MIT Press
Teubert, W. (2001). „Corpus Linguistics and Lexicography". *International Journal of Corpus Linguistics*, 6, Special Issue, 125-153.
Teubert, Wolfgang, ed., (2007). *Text Corpora and Multilingual Lexicography.* University of Birmingham Benjamins Current Topics 8 2007.
Vavřín, M., A. Rosen (2008). „Intercorp: A Multilingual parallel Corpus". In *Trudy Meždunarodnoj konferencii "Korpusnaja lingvistika" 2008*, Sankt-peterburgskij gosudarstvennyj univerzitet, Sankt-Peterburg, 156-162.