INTERCORP: MOŽNOSTI POROVNÁVANIA A ŠTÚDIA VIACERÝCH JAZYKOV

(Mnohojazyčný korpus InterCorp: Možnosti studia. Ed. F. Čermák – J. Kocek. Praha: Nakladatelství Lidové noviny 2010. 292 s. ISBN 80-7422-058-6

InterCorp: Exploring a Multilingual Corpus. Ed. F. Čermák – P. Corness – A. Klégr. Praha: Nakladatelství Lidové noviny 2010. 254 s. ISBN 978-80-7422-042-5)

Beáta Kmeťová – Adriana Žáková
Slovenský národný korpus Jazykovedného ústavu Ľudovíta Štúra SAV Panská 26
811 01 Bratislava, e-maily: beatak@korpus.sk, adriana163@korpus.sk

Cieľom projektu InterCorp (http://www.korpus.cz/intercorp/), ktorého riešiteľmi boli predovšetkým pracovníci Ústavu Českého národného korpusu, je budovať paralelné korpusy pre jazyky, ktoré sú predmetom štúdia na Filozofickej fakulte Karlovej Univerzity v Prahe. Prvé výstupy a výsledky tohto jedinečného projektu boli prezentované na medzinárodnej konferencii InterCorp Praha 2009, ktorá sa konala 17. – 19. septembra 2009 na pôde FF UK v Prahe. Na konferenciu bolo prihlásených 45 prednášateľov, v prvý deň prebiehali rokovania v dvoch sekciách. Výstupom z podujatia sú dva recenzované zborníky InterCorp zložené z českého (obsahuje 25 príspevkov v českom a slovenskom jazyku) a anglického zväzku (15 príspevkov v anglickom jazyku). V oboch zborníkoch sa prezentujú presvedčivé a materiálovo podložené náhľady do výskumov v oblasti korpusovej lingvistiky s využitím paralelných korpusov rôznych jazykov.

Zborník Mnohojazyčný korpus Intercorp: Možnosti studia prináša príspevky v českom a slovenskom jazyku vecne usporiadané do piatich prehľadných celkov: (1) gramatický výskum, (2) výskum lexiky, frazeológie a slovotvorby, (3) kategórie a javy, (4) aplikácie a technické aspekty paralelných korpusov, (5) preklad a jeho aspekty. Sústreďuje kontrastívny, porovnávací výskum češtiny s celkovo 13 jazykmi (slovanskými, románskymi, germánskymi, ale aj s litovčinou či fínčinou).

Porovnávaniu gramatických javov medzi dvomi alebo viacerými jazykmi pomocou dát z paralelných korpusov sa venujú autori príspevkov v prvom bloku. Problém prekladania postupne sa rozvíjajúcich prívlastkov z češtiny do francúzštiny analyzuje F. Esvan. Keďže francúzština sa na rozdiel od češtiny väčšiemu počtu adjektív (už viac ako dvom) bez ohľadu na ich umiestnenie pred alebo po substantíve skôr vyhýba, autor na základe textov z InterCorpu sleduje, aké riešenia volia v prípadoch kumulatívnych sledov atribútov profesionálni prekladatelia. Tí častokrát automaticky prekladajú jedno z českých adjektív substantívom alebo adjektíva rôzne rozmiestňujú okolo substantíva, niekedy sa snažia znížiť počet adjektív zhrnutím výz-

namov rôznych prvkov do jedného adjektíva alebo, ako krajné riešenie, pridávajú nové prvky. Striktné držanie sa originálu a vytváranie neobvyklých spojení s prítomnosťou troch adjektív podľa autora príspevku neprospieva kvalite prekladu, zdôrazňuje však, že tento jeho názor by bolo potrebné konfrontovať so skutočným úzom týchto konštrukcií v pôvodných textoch, nielen v prekladoch. P. Pečený prezentuje prehľad tendencií, ktoré sa prejavujú v distribúcii porovnávacích spájacích výrazov v češtine (*jak* a *než*) v konfrontácii s použitím nemeckých ekvivalentov *wie* a *als*. Súčasťou práce je aj pokus o zmapovanie faktorov, akými sú sémantika porovnávacej štruktúry, štýlový charakter komunikátu a formálne hľadisko na syntaktickej rovine, ktoré môžu mať vplyv na výber konkrétneho spájacieho prostriedku. L. Uhlířová vo svojom príspevku skúma pomocou korpusových dát výhradne smerom od bulharčiny k češtine, ako často sa bulharské posesívne zámeno prekladá do češtiny opäť posesívom, ako často sa vyjadruje inými výrazovými prostriedkami a ako často dochádza len k implikovaniu významu kontextom. Z jej výskumu vyplýva, že na obmedzených paralelných dátach (žánrovo, rozsahom, autorsky aj prekladateľsky – román Blagy Dimitrovovej Lavína a jeho preklad do češtiny od Aleny Maxové) sa nedá dospieť k všeobecným typologickým záverom. Ukázala sa však väčšia explicitnosť pri vyjadrovaní posesívnych vzťahov, väčšia nominálnosť a častejšie využitie pasívnej vetnej perspektívy v bulharčine, v češtine zasa výraznejšie obsadzovanie podmetovej pozície agentom a iné. H. Confortiová porovnáva výskyt slov *každý, všechen* a *celý* v češtine, angličtine a španielčine. Skúmaným materiálom z ČNK sa stal český text (Zdeněk Jirotka: Saturnin), ktorý vyhovel podmienke prekladu do oboch cudzích jazykov. Asymetrii valenčných vlastností českých a anglických slovies pohybu porovnávaním textov z dvoch formálne rozdielnych zdrojov (z Prague Czech-English Dependency Treebank – PCEDT a korpusu InterCorp) sa vo svojom príspevku venuje J. Šindlerová. Autorka zdôrazňuje, že poznanie charakteru asymetrií valenčných rámcov je dôležité pri vytváraní valenčných slovníkov i pri implementácii elektronických valenčných slovníkov do systému strojového prekladu. V pokračujúcom výskume si zároveň kladie za cieľ overiť univerzálnosť špecifickej valenčnej teórie vytvorenej pri práci s českými dátami a možnosť jej použitia na dáta iných jazykov.

Príspevky v ďalšom bloku sú venované výskumu lexiky, frazeológie a slovotvorby. D. Blažek si všíma úskalia jazykovej príbuznosti češtiny a slovinčiny, konkrétne medzijazykovú homonymiu, a porovnáva jednotlivé významy prefigovaných slovies so sémanticky totožnými slovotvornými základmi. P. Čermák a P. Štichauer skúmajú typológiu českých ekvivalentov talianskej konštrukcie *fare* + infinitív a španielskej konštrukcie *hacer* + infinitív. Ich korpusové analýzy ukázali, že na tento dominantný španielsky a taliansky prostriedok vyjadrovania kauzatívnosti sa v češtine viaže veľmi pestrý obraz rôznych prostriedkov, čiže v českej časti korpusu

nemožno hovoriť o dominantnosti jedného spôsobu vyjadrovania kauzatívnosti. Ďalšie dva príspevky sa venujú téme deminutív. H. Gladkova porovnáva pomocou dát česko-bulharského paralelného korpusu kategóriu deminutívnosti v češtine a bulharčine a T. Káňa, vychádzajúc zo svojich predchádzajúcich prác s dátami česko-nemeckého paralelného korpusu, hľadá ekvivalenty českých substantívnych deminutív v nemčine a angličtine a sleduje, aký je medzi nimi formálny rozdiel. Zjavným sa ukazuje fakt vyplývajúci z typológie porovnávaných jazykov, že čím je jazyk analytickejší, tým má menší sklon tvoriť syntetické deminutíva a skôr siaha po iných prostriedkoch. Úlohou ďalších výskumov na oveľa reprezentatívnejších dátach zostáva zistiť, či je táto „strata" v texte nejako kompenzovaná alebo či sa dá považovať za prirodzenú daň rozdielnej typológii. Cieľom práce K. Jiráska bolo overiť možnosti využitia paralelného korpusu na komparatívne štúdium prirovnaní ako špecifickej časti frazeológie a idiomatiky v českom a chorvátskom jazyku. Prichádza k záveru, že z výsledkov analýzy skúmanej vzorky dát sa vzhľadom na v tom čase malý rozsah paralelného korpusu InterCorp nedajú robiť príliš všeobecné závery o rozdieloch medzi oboma jazykmi. Nateraz vidí InterCorp ako nástroj vhodnejší skôr na kontrastívne štúdium a na translatologické analýzy a prax, než na čisto lingvistickú a lexikografickú prácu. F. Martínek vo svojom článku ukazuje možnosti využitia paralelného česko-nemeckého korpusu pri výskume prekladových ekvivalentov určitého typu slovesno-menných spojení – českých tzv. analytických verbonominálnych spojení (AVNS) v nemčine. Ide o spojenie významovo „vyprázdneného" slovesa a abstraktného substantíva, prípadne predložkovej frázy obsahujúcej abstraktné substantívum, ktoré autor v súlade s F. Čermákom považuje za kvázifrazémy (*udělit pokyn, mít povědomí, brát v úvahu*). Nemeckou obdobou sú tzv. Funktionsverbgefüge v češtine. V ďalších analýzach považuje F. Martínek za vhodné vychádzať oddelene z textov preložených z češtiny do nemčiny a z textov preložených z nemčiny do češtiny, čo by jednak spresnilo výsledky a jednak by sa získali dáta na skúmanie vplyvu štruktúr východiskového jazyka na jazyk cieľový. K. Rysová sa vo svojom príspevku zameriava na problematiku tzv. bezpríznakového slovosledu v českej a nemeckej výpovedi. Na dátach Pražského závislostného korpusu skúma či jednotlivé druhy českých voľných slovesných doplnení majú tendenciu vyskytovať sa vo výpovedi v určitom vzájomnom poradí, pričom výsledky svojho výskumu porovnáva s výsledkami germanistu Waltera Flämiga, ktorý realizoval podobný výskum v rokoch 1981 a 1991 na nemčine. Materiál ČNK, resp. InterCorp bol použitý na porovnanie slovosledu českých a nemeckých výpovedí v konkrétnych príkladoch. Z porovnávacieho výskumu autorky vyplýva, že v češtine aj v nemčine je slovosledné správanie sa kontextovo nezapojených voľných slovesných doplnení veľmi podobné a závisí od toho, či je toto doplnenie obligatórnou alebo fakultatívnou súčasťou valenčného rámca svojho riadiaceho slovesa.

Analýzou vybraných konštitutívnych častíc, ich významov a ich ekvivalenciou v textoch česko-slovenského a slovensko-českého paralelného korpusu sa zaoberala vo svojom príspevku M. Šimková. V jej typológii ide o podskupinu modálnych častíc, ktoré vyjadrujú významy *želanie, žiadosť, súhlas, prisvedčenie, zápor* a majú funkcie otázky alebo odpovede. Na podrobnejšiu analýzu si vybrala lexémy *akurát, bodaj (by, -že), kiež (by), nech* s rôznou významovou štruktúrou a synonymiou a zároveň skúmala mieru konštitutívnosti českých ekvivalentov *akorát, zrovna, bodejť, kdyby, aby, kéž (by), nechť, ať* a i.

Ďalší blok príspevkov otvára F. Čermák porovnaním jedného jazykového javu vo viacerých jazykoch. Zamýšľa sa nad povahou všeobecného subjektu a nad jeho lexikálnym vyjadrením vo vzájomnom porovnaní štyroch jazykov (nemecké *man*, anglické *one*, francúzske *on,* české *člověk*) a snaží sa zistiť, či sú tieto formálne prostriedky ekvivalentné, a ak áno, tak do akej miery. Naráža však väčšinou na rozdiely alebo na nerovnakú povahu východiskovej kategórie. H. Peloušková sa vo svojom príspevku, ktorý je časťou obsiahlej štúdie, venuje nemeckým konštrukciám s *es* a ich českým ekvivalentom. Cieľom príspevku Ł. Bańczykovej je gramatický a sémantický opis poľských vetných konštrukcií s neosobnými tvarmi slovesa na *-no/--to*, prezentácia výsledkov porovnania s ich českými prekladovými ekvivalentmi, ktoré ponúka paralelný korpus, a napokon analýza úrovne tejto ekvivalencie.

Príspevky v ďalšom bloku spájajú aplikácie, technické aspekty tvorby a využívanie paralelných korpusov. Príspevok P. Kopřivu predstavuje návrh na automatizovanú identifikáciu lexikálnych štruktúr v paralelných korpusoch, tzv. chunkov (časti viet, ktoré tvoria syntaktický funkčný celok). Ich zarovnanie by umožnilo lepšie poznanie štruktúrnych podobností a rozdielov medzi jazykmi. Špecifickým problémom segmentácie, zarovnávania a budovania česko-arabského paralelného korpusu sa venuje J. Milička. P. Vondřička prináša vo svojom príspevku informácie o TCA2, jednom z nástrojov na zarovnávanie paralelných textov. O. Nádvorníková, A. Polická, J. Šotolová a P. Vurm predstavili projekt medziuniverzitnej spolupráce romanistov na vysokoškolských kurzoch francúzskej filológie s použitím francúzsko-českej časti InterCorpu. Príspevok J. Skoumalovej obsahuje experimenty s paralelnými korpusmi, vykonanými s cieľom pomôcť pri zostavovaní chýbajúceho všeobecného česko-litovského dvojjazyčného slovníka. Slovník získaný extrakciou z veľmi obmedzených korpusových dát poslúži ako surový materiál na zostavenie skutočného slovníka lexikografmi.

Posledný blok príspevkov je zameraný na preklady a jeho aspekty. Ide hlavne o preklady medzi príbuznými jazykmi a na výskum boli použité dáta z príslušných paralelných korpusov. A. Adamovičová vo svojom príspevku analyzuje tri typy prekladateľských zlyhaní doložených v srbských prekladoch (K. Čapek, B. Hrabal, O. Pamuk cez angličtinu) a vplyv deformácie originálu na korpusové dáta. M. He-

bal-Jezierska vo svojom príspevku ukazuje, ako sa české derivačné univerbizáty, ktorých je v nespisovnej češtine dvakrát viac ako v poľštine, prekladajú do poľského jazyka. M. Nábělková podáva podrobnejší náhľad (konfrontáciu) na česko-slovenskú slovotvornú adjektívnu diferenciu *-ný/-ní* na báze paralelného korpusu. Kritikou prekladu frazém na báze paralelného česko-ruského korpusu sa vo svojom príspevku zaoberá L. Stěpanova.

Zborník príspevkov v anglickom jazyku INTERCORP: EXPLORING A MULTILINGUAL CORPUS obsahuje pätnásť príspevkov rozdelených do štyroch celkov: možnosti využitia paralelných korpusov, výskum gramatiky na korpusových dátach, výskum lexiky a prekladov a iné témy súvisiace s výskumom na korpusových dátach.

F. Čermák vo svojom príspevku predstavil viacjazyčné paralelné korpusy a vyzdvihol ich prínos pri porovnávacom výskume jazykov. Synchrónny korpus InterCorp, ktorý obsahuje ručne zarovnané, prevažne beletristické texty v 21 jazykoch okrem češtiny (v roku 2013 už korpus obsahoval texty v 31 jazykoch), napomáha pri budovaní nástrojov a technológií počítačovej lingvistiky, no predovšetkým možno paralelné korpusy využiť v oblasti všeobecnej lingvistiky, pragmalingvistiky, sociolingvistiky a i. M. Barlow vo svojom príspevku prezentuje využitie korpusov pri skúmaní kauzálnych viet. Korpusové dáta dokazujú rozdiely v štruktúre a význame kauzálnych viet. Z jeho výskumu vyplýva, že každý jedinec uprednostňuje istý typ kauzálnych viet. Otázkam prekladu sa vo svojich štúdiách venujú W. Teubert (prekladové ekvivalenty Konfuciových Analektov) a S. Johansson (možnosti využitia paralelných korpusov anglického, nórskeho a nemeckého jazyka a zároveň spomína obmedzenia pri výbere textov z dôvodu absencii textov preložených do oboch jazykov).

Paralelné korpusy ponúkajú aj možnosť štúdia morfologickej a syntaktickej roviny jazyka. S. Cinková sa napr. venuje komparatívnemu výskumu švédsko-českých konzekutívnych viet, R. von Waldenfels skúma možnosti paralelných korpusov odhaliť vzdialené sémantické vlastnosti dvoch príbuzných poľských a českých modálnych slovies *musieť* a *môcť* s adaptovaním kvantitatívneho a kvalitatívneho prístupu. O. Nádvorníková vo svojej štúdii analyzuje používanie ekvivalentu českého prechodníka vo francúzštine, pričom konštatuje, že ho v súčasnosti pri prekladoch nahradilo prítomné príčastie. Paralelné korpusy sú veľmi cenným a bohatým zdrojom pri skúmaní prekladových ekvivalentov i neplnovýznamových slovných druhov, napr. predložiek. M. Malá, P. Šaldová a A. Klégr sa venovali anglickým ekvivalentom predložiek *v/vo*, R. Novotná anglickým ekvivalentom českej predložky *na*.

Autorky L. Chlumská a D. Kováříková vo svojej štúdii zistili, že sa v prekladoch odráža vplyv tradičných vyučovacích materiálov, pričom sa nekladie dostatočný dôraz na aktuálne a moderné používanie jazyka. Nabádajú prekladateľov, aby pri práci využívali korpusové dáta, ktoré demonštrujú reálne používanie jazyka.

P. Corness sa vo svojom výskume zaoberá prekladovými ekvivalentmi frekventovaného slova *povedať*. Autor vo svojom príspevku ukazuje, že pri preklade beletrie dochádza k sémantickým posunom prekladaných slovies, a naznačuje, že faktory vplývajúce na tieto posuny by mohli byť predmetom ďalšieho skúmania. Výsledky svojej štúdie o kľúčových slovách v svetoznámych fantasy knižkách o Harrym Potterovi predniesli A. Čermáková a L. Fárová. Autorky odporúčajú používať techniku generovania veľmi frekventovaných kľúčových slov, ktoré môžu prekladateľom vopred zabezpečiť jednotné prekladové ekvivalenty.

Otázkou vplyvu kontextu na slovnodruhové zaradenie slov sa zaoberá V. Cvrček. Vo svojom výskume autor predpokladá, že korelujúce kontexty slov majú ten istý, príp. podobný význam alebo funkciu, tzn. že je možné identifikovať skupiny slov, ktoré vytvoria triedu slov s rovnakým významom alebo funkciou. A. Rosen sa zamýšľa nad problémom viacerých tagsetov – súborov značiek a pravidiel značkovania existujúcich pre český jazyk. Vo svojej štúdii autor ponúka jednotnú medzijazykovú hierarchiu kategórií (sémantické, syntaktické a morfologické vlastnosti), ktorá by vylepšila morfosyntaktickú anotáciu v paralelných korpusoch zarovnaných na slová. V záverečnom príspevku M. Vachková analyzuje sémanticko-lexikálne javy v paralelnom česko-nemeckom korpuse. Analýzou kolokačných profilov dvoch nemeckých adjektív *schlecht* a *böse* ukazuje na variabilitu prekladu. Autorka dokazuje, ako sa dá pomocou dvojjazyčného korpusu skvalitniť informačná ponuka vo vznikajúcom Veľkom nemecko-českom akademickom slovníku.

Multilingválny paralelný korpus súčasnej češtiny InterCorp zaznamenal od roku 2009 výrazný nárast v kvantite slov, ako aj v kvalite nástrojov spracovania prirodzeného jazyka. V poradí šiesta verzia InterCorp-u z apríla 2013 obsahuje 728 miliónov slov a je dostupná v 31 cudzích jazykoch. Kontinuálne rozširovanie korpusu podnecujú výskumy v danej oblasti a výsledky z korpusovo zameraných konferencií, napr. aj ďalšej konferencie InterCorp v r. 2011. O rozmachu korpusovej lingvistiky svedčí aj rozširovanie a budovanie paralelných korpusov v iných krajinách, napr. v Bulharsku, Poľsku, na Slovensku. Jednotlivé príspevky z roku 2009, z prvej fázy výstavby InterCorp-u, predstavujú rozmanité využitie korpusových databáz na vedecko-výskumné ciele a poskytujú mnohé námety na budúce smerovanie vývoja výskumu v oblasti spracovania prirodzeného jazyka a korpusovej lingvistiky.