ZDROJ JAZYKOVÝCH DAT PRO VÝZKUM A VÝUKU NA FAJ

THE SOURCE OF DATA FOR LANGUAGE TEACHING AND RESEARCH ON THE FAJ

TOMÁŠ KÁŇA

Abstract
The following article introduces the multilingual corpus InterCorp and its basic features. The impact of using corpus data for FAJ is shown on exploring three German phrasemes in contrast with Slovak and English.

Keywords: corpus based research, contrastive research, phraseme, InterCorp

Abstrakt
Následující stať představuje elektronický korpus InterCorp a jeho základní parametry. Na příkladech tří frazémů z ekonomické oblasti je naznačena možnost využití dat korpusu ve výzkumu na FAJ.

Kľúčové slová: korpusový výzkum, kontrastivní výzkum, frazém, InterCorp

Úvod

Již od roku 2005 vzniká na Ústavu Českého národního korpusu Filozofické fakultě Univerzity Karlovy multijazykový paralelní korpus InterCorp. Do dnešní doby se rozrostl na 33 jazyků, další se plánují.

Tento ve světě ojedinělý nástroj vhodný především k lingvistickému výzkumu může – a měl by – velkou měrou vstupovat do moderní výuky cizích jazyků. Důvodů je několik a všechny vlastně vyplývají už z definice korpusu.

Jazykový korpus

„Jazykový korpus je strukturovaný, unifikovaný a často označkovaný velmi rozsáhlý soubor jazykových dat (obv. v podobě textů), (…) jejichž výběr chce být vzhledem k vytýčenému cíli reprezentativní“ (1, 1997, s. 135). V této definici (ne náhodou pochází z pera nestora české korpusové lingvistiky, prof. Čermáka) se skrývají všechny důležité aspekty moderní výuky: Korpus je rozsáhlý soubor jazykových dat. Tím je dána objektivita výstupu. Z korpusu statistickou metodou získáváme jevy typické, převažující, které by měly stát v centru zájmu výuky jazyků. Jazyková data jsou uložena v podobě ucelených textů. Současný *„postkomunikační“* (2, 2006, s. 83 – 160) a na jazykové konání zaměřený přístup k výuce pracuje jen s celými texty. Korpusy sice běžnému uživateli většinou neumožňují získat celý text, autentičnost získaných pasáží je však naprosto zřetelná. Korpusy je možné navolit nebo vybrat tak, aby texty v nich obsažené byly pro daný úkol nebo jev reprezentativní.

Podle zaměření práce si uživatel volí typ korpusu. Existuje jich několik. Jedno ze základních dělení liší korpusy jednojazyčné a vícejazyčné. (Další typy viz Káňa 2014, s. 24 – 30.)

Jednojazyčný korpus je databanka (většinou psaných) textů jen jednoho jazyka. V paralelních korpusech jsou zařazeny originály a jejich oficiální překlady. Vedle paralelních korpusů existují ještě tzv. srovnávací korpusy. O těch zde ale řeč nebude. Ke srovnávacím korpusům viz např. Benko, 2014.

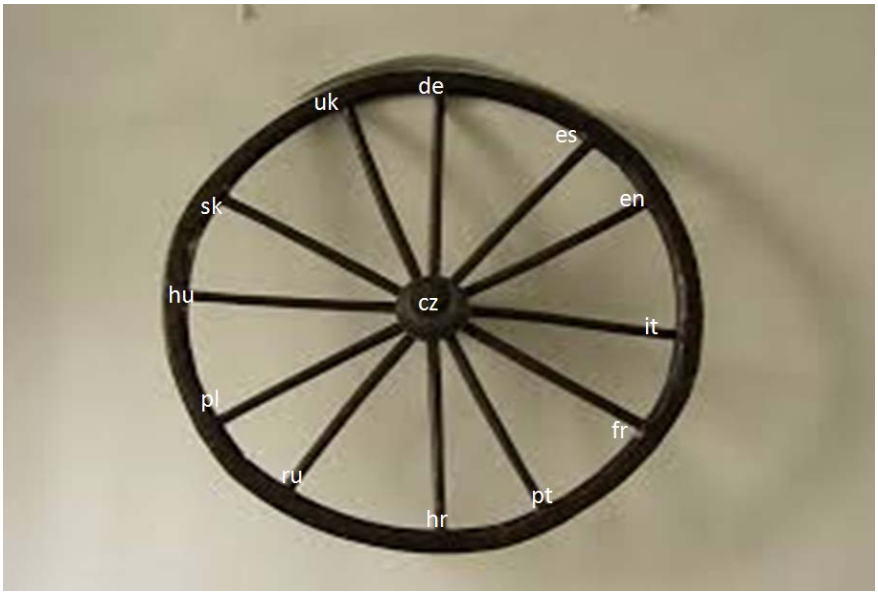
K čemu slouží jazykové korpusy

Oblastí, v nichž se mohou jazykové korpusy uplatnit, je mnoho. Ve vysokoškolském vzdělávání to je především (kontrastivní) lingvistický výzkum, výuka jazyka/jazyků v kontrastu, autonomní učení, překladatelství.

V obecné rovině mohou korpusy sloužit k prověření nabytých znalostí o jazyce a také k demokratičnosti vědění o jazyce. Princip typického a užívaného totiž odsouvá reglementární příručky směrem ke smetišti dějin a měl by instituce, které tyto příručky vydávají, nutit k pravidelné aktualizaci, resp. rekodifikaci jazyka. Plaidoyer za užívání korpusu při kodifikaci se však vždycky bude jevit jako liché, pokud nebude existovat v jazykové komunitě, ale i v odborných kruzích základní konsensus o fungování jazyka. Ten, bohužel, není, proto se jeví soulad pravidel (slovníků, gramatik) a úzu zatím v nedohlednu. Platí to pro slovenštinu, češtinu, ale i němčinu, kde sice poslední gramatika nakladatelství Duden (2005) obsahuje doklady z korpusu, ovšem pouze ty, které odpovídají pravidlům, jež se stále dokola opisují už několik desetiletí (srov. např. postavení plnovýznamového, pomocného a modálního slovesa na konci vedlejší německé věty; více k tomuto tématu Káňa, 2014, s. 155 – 161).

InterCorp

Projekt InterCorp je multilingvální korpus s 33 (převážně evropskými) jazyky. Tento elektronický nástroj lze přirovnat ke kolu od žebřiňáku, v jehož středu je čeština, jak znázorňuje obrázek 1.



Obrázek 1 Brněnské kolo jako alegorie InterCorpu (Foto a grafika autor)

Znamená to, že všechny paralelní texty mají českou verzi. K té existuje vždy minimálně jedna verze v jiném jazyce.

Texty korpusu InterCorp se dělí na tzv. jádro (ručně zarovnané texty) a paralelní texty různých projektů (Syndicate), portálů (Presseurope), jiných paralelních korpusů (open subtitles). „Jádro“ korpusu představují především beletristické texty posledních padesáti let. Ve verzi 6 byly velikosti jednotlivých paralel následující:

Tabulka 1: Velikosti paralel InterCorp verze 6

VELIKOST KORPUSU V TISÍCÍCH SLOV							
Zkratka	**Jazyk**	**Jádro**	**Syndicate**	**Presseurop**	**Acquis**	**Europarl**	**Celkem**
ar	arabština	29	0	0	0	0	29
be	běloruština	1 308	0	0	0	0	1 308
bg	bulharština	3 979	0	0	13 816	9 083	26 879
ca	katalánština	1 758	0	0	0	0	1 758
da	dánština	190	0	0	21 680	13 916	35 785
de	němčina	17 256	3 050	1 715	21 724	13 089	56 835
el	řečtina	210	0	0	25 070	15 404	40 683
en	angličtina	10 019	3 083	1 863	24 208	15 580	54 753
es	španělština	14 552	3 479	1 948	27 001	15 885	62 865
et	estonština	0	0	0	15 963	10 900	26 862
fi	finština	2 131	0	0	16 667	10 241	29 040
fr	francouzština	3 816	3 535	2 054	27 352	17 178	53 936
hi	hindština	155	0	0	0	0	155
hr	chorvatština	12 625	0	0	0	0	12 625
hu	maďarština	2 511	0	0	19 168	12 307	33 985
it	italština	4 081	247	1 893	24 850	15 489	46 560
lt	litevština	358	0	0	18 433	11 020	29 811
lv	lotyština	1 337	0	0	18 745	11 689	31 770
mk	makedonština	2 664	0	0	0	0	2 664
mt	maltština	0	0	0	14 133	0	14 133
nl	nizozemština	9 426	0	2 082	24 746	15 563	51 817
no	norština	2 301	0	0	0	0	2 301
pl	polština	12 710	0	1 660	20 464	12 805	47 640
pt	portugalština	2 318	0	2 103	28 599	16 481	49 502
ro	rumunština	2 433	0	1 917	8 200	9 446	21 995
ru	ruština	4 937	2 651	0	0	0	7 588
sk	slovenština	8 152	0	0	19 222	12 734	40 108
sl	slovinština	1 855	0	0	19 646	12 241	33 741
sr	srbština	6 972	0	0	0	0	6 972
sv	švédština	7 205	0	0	20 615	13 874	41 694
uk	ukrajinština	1 493	0	0	0	0	1 493
celkem		**138 779**	**16 044**	**17 237**	**430 300**	**264 926**	**867 287**
cs	čeština	61 962	2 741	1 639	20 285	12 920	99 547

Zdroj: ÚČNK FF UK Praha

Při výběru různých kombinací jazyků dopovídá velikost dat průniku jednotlivých paralel. Např. při výběru české, slovenské a anglické paralely můžeme hledat jen v textech, které jsou v korpusu jak v češtině, tak ve slovenštině, ale i angličtině. S množstvím zvolených jazyků velikost dat pochopitelně klesá. Orientačně jsou v tabulce 2 uvedeny velikosti paralel, které by mohly být užitečné pro výzkum a ve výuce na FAJ EU.

Tabulka 2 Velikosti kombinací InterCorpu ve verzi 6

	sk	**en**	**de**	**hu**	**es**	**ru**
sk - en	33 mil.	35 mil.				
sk - de	33 mil.		36 mil.			
sk - de - en	32 mil.	38 mil.	36 mil.			

sk - hu	33 mil.			32 mil		
sk - de (bel.)	0,9 mil		1 mil.			
sk - en (bel.)	1,1 mil	1,4 mil				
sk - de - en - hu	269 tis	245 tis	317 tis	268 tis		
sk - es	30 mil.				30 mil.	
sk - ru	432 tis.					435 tis.

V tabulce 3 jsou pro srovnání uvedeny velikosti slovenských paralel v paralelních korpusech na JUĽŠ SAV, Bratislava.

Tabulka 3 Velikosti slovenských paralel paralelních korpusů SNK

Slovensko-anglický paralelný korpus	184 mil.
Slovensko-bulharský paralelný korpus	78 mil.
Slovensko-český paralelný korpus	10 mil.
Slovensko-francúzsky paralelný korpus	??
Slovensko-latinský paralelný korpus	0,75 mil.
Slovensko-maďarský paralelný korpus	1,5 mil.
Slovensko-ruský paralelný korpus	2 mil.

Zdroj: http://korpus.juls.savba.sk/par.html

Ze srovnání je zjevné, že pro kontrastivní práci např. mezi slovenštinou a angličtinou je jednoznačně výhodnější pracovat s korpusem u SNK, protože obsahuje mnohem víc dat než průnik české, slovenské a anglické paralely. Ovšem např. slovensko-maďarská data jsou mnohem bohatší v InterCorpu.

Všechny údaje v tabulkách jsou v počtech tokenů (počítačově segmentovatelných „slov“). Rozdíly v počtech (např. 33 mil. slovenských slov oproti 36 mil. německých slov) jsou dány rozdíly v typologii jazyků.

Během psaní tohoto článku byla spuštěna verze 7 InterCorpu, v níž přibyly další texty a především velké množství dat z volně přístupných filmových titulků. Tyto „texty“ jsou pro práci s jazykem cenné především pro svoji blízkost k mluvené řeči.

Výhoda InterCorpu spočívá také v tom, že lze volit mezi několika jazyky najednou. Jak může vypadat výsledek dotazu, naznačuje tabulka 4. Byl položen dotaz na základní tvar (lemma) slova *zahrada* (tedy *zahrady*, *zahradě* atd.). Z několika desítek konkordancí (dokladových řádků) byla vybrána jen jedna – z knihy Rowling, J. K.: Harry Potter a kámen mudrců.

Tabulka 4: Paralelní konkordance dotazu na lemma *zahrada*

Teď seděla na zídce jeho zahrady .	Sie saß jetzt auf seiner Gartenmauer.	It was now sitting on his garden wall.	Az állat most a kertet szegélyező falon üldögélt.	Teraz sedela na záhradnom múriku.

Doklady z korpusu lze vždy uložit do počítače v různých formátech (txt, xls, doc). Práce s InterCorpem je podrobně popsána v češtině na http://wiki.korpus.cz/doku.php. Anglická verze se plánuje, dosud (leden 2015) spuštěna není. Podrobný popis v němčině je také v publikaci Káňa, 2014.

Využití InterCorpu

Jak již bylo naznačeno, paralelní korpus InterCorp je vhodný jako zdroj autentického jazykového materiálu ve výzkumu a výuce cizích jazyků (srov. také Pelouškova 2013).

Ve výzkumu se přímo nabízí rozšíření kvalitativních analýz o kvantitativní rozměr. Podíváme-li se na jednu z posledních statí k problematice frazémů v ekonomických textech (Lišková, 2014) a zvolíme kterýkoliv frazém, můžeme na korpusových datech celkem rychle zjistit také frekvenci frazémů (ty se zpravidla neuvádějí). Dále můžeme celkem snadno ověřit i jejich relevanci pro běžnou řeč a v neposlední řadě i další možnosti ekvivalentů v cílovém jazyce. V uvedené publikaci D. Liškové se jedná o excerpty frazémů ze tří odborných ekonomických publikací (3, 2015, s. 6). Autorka k nim uvádí slovenské ekvivalenty odborného ekonomického stylu a vyznačuje míru formálně-sémantické shody frazémů. Exemplárně jsou zde rozvedeny tři frazémy: *Bände sprechen*, *auf die Beine kommen* a *in den Wind schlagen.*

Bände sprechen

„‚*Das spricht Bände* über den Zustand der Wirtschaft (Roubini, N. 2010. *Weltwirtschaft und ihre Zukunft*, S. 12.)
das spricht Bände – to hovorí za všetko/to veľa hovorí“ (3, 2014, s. 7)

V následující tabulce (5) jsou uvedeny ukázkové konkordance z paralelního korpusu. Frazém má v porovnání s ostatními dvěma střední frekvenci a podobnou jako např. *ins Stocken geraten*. Doklady pocházejí vesměs z politicko-ekonomických nebo čistě publicistických/zpravodajských textů.

Tabulka 5 Doklady frazému *Bände sprechen* v německo-slovenském kontrastu

Ein Jahresbericht kann über das Erreichte (…) **Bände sprechen**.	Výročná správa **môže hovoriť veľa** o dosiahnutých výsledkoch …
Seine engsten Freunde auf internationaler Ebene sind Diktatoren wie Castro, Lukaschenko und Ahmadinedschad, **was Bände spricht**.	Jeho najbližšími priateľmi na medzinárodnej scéne sú diktátori Castro, Lukašenko a Ahmadínedžád, čo **hovorí mnoho**.
Es spricht Bände, dass dieses Gedenken im Anschluss an eine Aussprache zu einem Bericht mit dem Titel " Frauenarmut in der Europäischen Union " stattfindet .	**Je veľavravné**, že táto spomienka prichádza po rozprave o správe s názvom Podoba chudoby žien v Európskej únii.

in den Wind schlagen

„*etw. in den Wind schlagen* – pustiť niečo do vetra/brať niečo na ľahkú váhu“ (3, 2014, s. 15)

Tabulka 5 uvádí u tohoto relativně velmi slabě frekventovaného frazému téměř všechny doklady z korpusu. Z její pravé poloviny je jasně vidět, že nemá ve slovenštině přímý ekvivalent.

Tabulka 6 Doklady frazému *in den Wind schlagen* v německo-slovenském kontrastu

… und **schlagen** fröhlich **die Ergebnisse** von Referenden **in den Wind**,.	…, pričom s radosťou **odsunieme nabok** výsledky referenda,
… indem sie **unsere Angebote** für Entwicklungshilfe einfach **in den Wind schlagen.**	Niektoré vlády v Afrike možno prerušia byrokraciu Komisie tým , že **odignorujú našu ponuku** na poskytnutie rozvojovej

	pomoci.
Leider haben die Finanzminister (…) **diese Vorschläge in den Wind geschlagen** -	S ľútosťou musím povedať, že ministri financií (…) **odmietli tieto návrhy** …
Kann es sich die Kommission (…) tatsächlich erlauben, **die Stellungnahme** des Rates , des Parlaments (…) **in den Wind zu schlagen**?	Môže si (…)Komisia skutočne dovoliť **pohŕdať** (…) **stanoviskom** Rady, Parlamentu…?
Die Einwände (…) müssen, koste es, was es wolle , **in den Wind geschlagen.**	… **je potrebné prekonať a prestať si všímať** akékoľvek námietky …

Stejný frazém lze porovnávat i trojjazyčně, pokud jsou dostatečná data. Jedná se první řádek tabulky 6, rozšířený o anglickou paralelu.

Tabulka 6a Doklad frazému *in den Wind schlagen* v německo-anglicko-slovenském kontrastu

…, denn wir wenden dieselben Standards ja auch innerhalb unserer eigenen Grenzen an und **schlagen** fröhlich die Ergebnisse von Referenden **in den Wind**, wenn wir der Meinung sind, dass sie den falschen Ausgang genommen haben.	because, of course , we apply the same standards within our own borders, happily **swatting aside** referendum results when we deem them to have gone the wrong way .	Rovnaké normy, prirodzene, uplatňujeme aj v rámci vlastných hraníc, pričom s radosťou **odsunieme nabok** výsledky referenda, ak ich považujeme za nevhodné .

auf die Beine kommen

„‚Viele dieser Menschen werden finanziell nicht mehr **auf die Beine kommen**.' (Wagenhofer, (2010). Let´s make Money S. 80.)
auf die Beine kommen – postaviť sa na nohy" (3, 2014, 7)

Poslední frazém je ze všech tří zde sledovaných nejfrekventovanější. Z dvaceti dokladů (což v korpusovém přístupu jistě není mnoho, a tudíž se považuje jen za podmíněně platné) můžeme odhadnout, (1) jak se německý frazém „chová" a (2) jaké má ekvivalenty ve slovenštině i angličtině:
ad 1) *auf die Beine kommen* se vyskytuje většinou v pořadí typickém pro vedlejší větu:
damit Europa wieder auf die Beine kommt.
…wieder auf die Beine zu kommnen.
Toto zjištění má význam především pro výuku syntaxe. Nelze si také nevšimnout, že se tento frazém vyskytuje v typickém spojení (chunk) *wieder auf die Beine zu kommen*.

Dalším zjištěním je fakt, že se frazém vyskytuje převážnou většinou v publicistických/ekonomických textech. K tomuto tvrzení je nutné dodat, že převážná část paralelních textů je sice z oblasti publicistiky, jsou mezi nimi však i texty beletristické. I v nich se toto spojení objevuje:

Tabulka 7 Doklady spojení *auf die Beine kommen* v beletristických textech

… und eine ganze Weile überhaupt nicht mehr **auf die Beine kommen** konnten.	… and could not **get on their legs** again for some time.	Predháňali sa v ďakovaní a klaňali sa mu až po zem, hoci pritom od únavy padali a

		dlhšie potom nevládali **vstať**. (Tolkien-Hobit)
»Nein!«, rief Teabing schrill und versuchte vergeblich, **auf die Beine zu kommen**.	"No!" Teabing cried out, trying in vain to **stand**.	"Nie!" vykríkol Teabing, márne sa pokúšajúc **vstať**. (brown-sifra)

Doklad v prvním řádku je z Tolkienova Hobita, v druhém řádku pak z Brownovy Šifry mistra Leonarda. Příklady ukazují, že v beletristických textech se toto spojení užívá v původním významu, není to ovšem spojení těsné, chybí zde ona metaforičnost (personifikace), která se projevuje v ekonomických textech.

ad 2) Typické protějšky německého frazému v ekonomických textech tedy jsou tyto:

get back to one´s feet; recover	*zotaviť sa*
get going again	*naštartovať (hospodárstvo)*
get on one´s legs; get moving again	*pohnúť sa vpred;*

Závěr

Z předchozích řádků plyne, že korpusy, v nichž jsou pro daný výzkum uloženy vhodné texty, mohou zásadním způsobem přispět k plastičnosti pohledu na jazyk i k revizi slovníků, gramatik. Do žádného slovníku ani gramatiky totiž nelze napsat všechno a už vůbec ne všechny ekvivalenty v jiném jazyce. Korpus může tuto nutnou selekci kompenzovat.

Jak bylo uvedeno již výš, je možné korpusy InterCorp použít k výuce a procvičování gramatiky, lexika, slovotvorby, překladů. Ke všem těmto oblastem byly sestaveny různě rozsáhlé učební pomůcky pro posluchače s L1 češtinou. Jejich představení a funkce včetně zkoušecího portálu budou náplní dalších statí.

Seznam bibliografických odkazů

1. ČERMÁK, F. 1997. *Jazyk a jazykověda*. Praha: Pražská imaginace, s. 135.
2. KUMARAVADIVELU, B. 2006. *Understanding Language Teaching. From Method to Postmethod*. Mahwah (New Jersey)/London: Lawrence Erlbaum Associates, Publishers, s. 83 – 160.
3. LIŠKOVÁ, D. 2015. *Phraseologismen in den deutschen populärwissenschaftlichen Texten mit Fokus auf die Wirtschaft. (Eine komparative Analyse).* Eisenstadt.

Literatura

ČERMÁK, F. 1997. *Jazyk a jazykověda*. Praha: Pražská imaginace.

BENKO, V. 2014. Aranea: Yet Another Family of (Comparable) Web Corpora. In: *SOJKA, P., HORÁK, A., KOPEČEK, I., PALA, K. Text, Speech and Dialogue.* 17th International Conference, TSD 2014. Brno, Czech Republic, September 8 – 12, 2014. Proceedings. Springer International Publishing Switzerland, s. 257 – 264.

DOVALIL, V., KÁŇA, T., PELOUŠKOVÁ, H., ZBYTOVSKÝ, Š., VAVŘÍN, M. 2013, 2014. *Korpus InterCorp – němčina, verze 6 z 8. 4. 2013 a verze 7 z 19. 12. 2014.* Praha: Ústav Českého národního korpusu FF UK.

KÁŇA, T. 2014. *Sprachkorpora in Unterricht und Forschung DaF/DaZ.* Brno: Masarykova univerzita.

KUMARAVADIVELU, B. 2006. *Understanding Language Teaching. From Method to Postmethod*. Mahwah (New Jersey)/London: Lawrence Erlbaum Associates, Publishers.

KUNKEL-RAZUM, K., MÜNZBERG, F. 2005. *Duden: die Grammatik. Unentbehrlich für richtiges Deutsch*. Mannheim: Dudenverlag.

LIŠKOVÁ, D. 2015. *Phraseologismen in den deutschen populärwissenschaftlichen Texten mit Fokus auf die Wirtschaft. (Eine komparative Analyse).* Eisenstadt (v tisku).

PELOUŠKOVÁ, H. 2013. Das Projekt InterCorp und seine Rolle in der Deutschlehrerausbildung und Forschung. In: *Slowakische Zeitschrift für Germanistik 5/2*. Banská Bystrica: Verband der Deutschlehrer und Germanisten der Slowakei, s. 55 – 64.

Kontakt
Mgr. Tomáš Káňa, PhD.
Masarykova univerzita
Pedagogická fakulta
Katedra německého jazyka a literatury
Poříčí 9, 603 00 Brno
Česká republika
Email: kana@ped.muni.cz