Využití elektronických korpusů ve studiu německého jazyka

Tomáš Káňa, Hana Pelouškova
Katedra německého jazyka a literatury Pedagogické fakulty Masarykovy univerzity, Brno

Klíčová slova:
korpus, jazykový korpus, korpusová lingvistika, paralelní korpus, česko-německý paralelní korpus

Anotace:
Moderní výpočetní technika otevírá donedávna omezené možnosti zkoumání jazyka v jeho přirozené podobě. Současný výzkum se provádí na elektronických korpusech – databankách textů, v nichž lze efektivně vyhledávat kýžené jazykové jevy.
Ve studiu filologických oborů však předměty seznamující s touto metodou chybějí. Článek informuje o největších národních korpusech relevantních pro studium němčiny na českých vysokých školách a o česko-německém paralelním korpusu. Představuje některé možnosti práce na korpusech, seznamuje s parametry česko-německého paralelního korpusu a navrhuje strukturu předmětu korpusová lingvistika ve studiu germanistiky.
V závěru je uveden seznam pro germanisty nejdůležitějších jazykových korpusů, dostupných prostřednictvím internetu.

„Poznání zákonitostí a pravidel stavby a užívání jazyka vyplývá jen a pouze ze studia jeho skutečného užívání (úzu), přičemž se tím myslí především užívání typické a časté.“[1]

Jedinečným zdrojem informací o jazyce je rozsáhlý, elektronicky zpracovaný autentický jazykový materiál shromážděný v jazykových korpusech. Specifický software umožňuje podle zadaných kritérií efektivně vytěžovat konkrétní výskyty jazykových jevů. Korpus se tak stává neocenitelným pomocníkem nejen vědce-lingvisty, ale také učitele a studenta daného jazyka.

Základním cílem koncepce implementace jazykových korpusů do studia učitelství německého jazyka na Pedagogické fakultě Masarykovy univerzity v Brně je vychovat kompetentního uživatele jazykových korpusů, vybaveného potřebnými znalostmi a dovednostmi umožňujícími smysluplně využít korpus vždy, kdy je potřeba. V průběhu studia se korpusy stávají samozřejmými nástroji - stejně jako gramatiky, slovníky, encyklopedie, odborná literatura, internet a další informační zdroje. Vzhledem k tomu, že cílovou skupinou jsou budoucí učitelé, předpokládáme a doufáme, že budou své kompetence předávat i svým žákům, popř. i kolegům a že celý záměr bude mít značný multiplikační efekt, který by se mohl (či měl) v konečném dopadu promítnout v kvalitnějším jazykovém „vybavení“ (a potažmo i v počítačové gramotnosti) populace. Jsme na samém počátku cesty, neboť zatímco povědomí o existenci slovníků či internetu je možno považovat za uspokojivé, povědomí o možnostech jazykových korpusů je prakticky nulové - a to i mezi učiteli!

Základní oblasti využití korpusů ve studiu korespondují s možnými oblastmi jejich využití i v mimouniverzitní praxi.

[1] Čermák (2005), s. 5.

Patří k nim především přímá podpora jazykové výuky, zdroj jazykového materiálu pro hlubší poznání jazyka (výzkum) a dále poslouží jako zdroj autentického jazykového materiálu pro sestavování učebních materiálů a testů.

Již v prvním semestru studia v předmětu **Úvod do jazykovědy** získávají studenti informace o důležitých jazykových korpusech, jejich funkcích a přístupu k nim. Jedná se zejména o jednojazyčné korpusy: **Český národní korpus**, **Mannheimský korpus** a **British National Corpus.**

Vzhledem k zaměření studia je nejvíce akcentován a v maximální míře využíván **ČNPK - Česko-německý paralelní korpus,** který vzniká na katedře německého jazyka a literatury Pedagogické fakulty Masarykovy univerzity od jara 2001 a je jediným fungujícím korpusem svého druhu na světě. Nejvýznamnějším partnerem projektu ČNPK je od počátku Ústav ČNK v Praze.
Většinu technických prací (skenování, opravy skenů, čištění textů, zarovnávání textů, vytváření metainformací) vykonávají studenti německého jazyka na Pedagogické fakultě MU. Od roku 2005 pokračuje sestavování korpusu v rámci multilingválního projektu InterCorp, který je součástí výzkumného záměru Ministerstva školství ČR „**Český národní korpus a korpusy dalších jazyků**“.[2]

Budování Česko-německého paralelního korpusu směřuje k dosažení cílových parametrů. Srovnání těchto parametrů se současným stavem korpusu zachycuje tato tabulka:

Cílové parametry	**Současnost**
Synchronní – Do korpusu jsou zařazovány texty publikované od roku 1920.	ano
Dvojjazyčný – Je vyloučena přítomnost dalšího jazyka. Poměr českých a německých originálních textů je 1:1.	ano Korpus obsahuje 63% českých originálních textů.
Dostatečně rozsáhlý – Korpus bude obsahovat minimálně 5 mil. slov v české části.	Korpus obsahuje 3,4 mil. slov v české části.
Obecný – Stanovené proporce jsou: 50% beletrie, 25% publicistika, 25% odborné texty z různých oblastí.	Zatím bylo do korpusu zařazeno 211 textů, z nichž tvoří: 63% beletrie, 17% publicistika, 20% odborné texty.
Zarovnaný (přiřazení českých a německých paralelních sekvencí)	Všechny texty jsou zarovnány na úrovni odstavce, většina z nich ještě manuálně na úroveň věty a jí odpovídající paralelní sekvence.
Lemmatizovaný v obou paralelách (přiřazení slovníkového tvaru ke každému slovu)	ano
Morfologicky značkovaný v obou paralelách (přiřazení gramatické charakteristiky ke každému slovu)	Problémem českého značení je nedostatečná desambiguace homonymních tvarů, která způsobuje cca 10% chybovosti. V německé části jsou označeny pouze slovní druhy.
Katalogizace a archivace textů na CD	ano

[2] Výzkumný záměr MŠMT ČR: MSM 0021620823.

Vnější anotace textů (metainformace) – Informace o původu a zpracování textů jsou zpracovány ke každé paralele a obsahují tyto položky: název dokumentu, název textu, zdroj, hrubou a jemnou stylistickou charakteristiku textu, médium, jméno a pohlaví autora, u překladu i jméno a pohlaví překladatele, jazyk, jazyk originálu, rok vydání, technické údaje o zpracování textu. Tyto informace lze vyvolat ke každému konkordančnímu řádku a lze podle nich vytvářet subkorpusy.	ano

Korpus je v současné době umístěn a spravován na serveru Fakulty informatiky MU a jeho uživatelským prostředím je manažer Bonito (tento manažer využívá ČNK i Slovenský národný korpus). Vzhledem k tomu, že část textů je chráněna autorskými právy, není běžně přístupný. Uvažuje se o vytvoření subkorpusu textů, které (již) autorským právům nepodléhají, a o jeho zpřístupnění nejširšímu okruhu lingvistů, učitelů a překladatelů. Verze korpusu z přelomu roku 2004/2005 existuje na CD.

Korpus může fungovat jako paralelní i jako dva samostatné jednojazyčné korpusy (český a německý). V ideálním případě by měl být on-line k dispozici pro případnou konfrontaci ve všech hodinách výuky praktického jazyka i lingvistických disciplin. To by ovšem předpokládalo přítomnost počítače, dataprojektoru a internetu ve všech posluchárnách, což je bohužel zatím hudbou (doufejme ne příliš vzdálené) budoucnosti. Zatím „přímá podpora výuky“ spočívá v tom, že studenti dostávají v hodinách drobné dílčí úkoly vyplývající z problémů či otázek nastolených přímo ve výuce. Tyto „minivýzkumy“ provedou prostřednictvím korpusu samostatně doma či v počítačové studovně a v následující hodině informují ostatní o výsledcích. Například v kapitole německé syntaxe zabývající se gerundivem pomohl korpus nalézt odpovědi na tyto otázky:

a) Jak často odpovídá adjektivu se sufixem *–telný* německé gerundivum?
b) „Žije“ české adjektivum „*očekávatelný*“ nebo je spíše vykonstruované?
c) Je prefix *an-* u slovesa *anerkennen* častěji odlučitelný nebo neodlučitelný (slovník připouští obě možnosti)?

Je potěšující, že se studenti ujímají těchto dobrovolných úkolů poměrně ochotně a samostatně odhalují taje jazyka a někdy i boří i vžité mýty tradičních učebnic: např. že německé spojce *nachdem* odpovídá v češtině nejčastěji *poté, co* (mnohem častěji je to *když*), užití spojky *nachdem* zavazuje k užití časové souslednosti (zdaleka ne vždy), německým ekvivalentem českého *si* je *sich* (jen výjimečně) atd.
Obdobným způsobem je ČNPK využíván i v semináři ke slovotvorbě, v lexikologii a v překladatelských cvičeních.

Studentům, které práce s korpusy v běžné výuce zaujala, je v jarním semestru 2006 poprvé nabízen volitelný předmět **Úvod do korpusové lingvistiky.** Cílem tohoto předmětu je prohloubit základní informace o počítačové a korpusové lingvistice a naznačit možnosti využití korpusové lingvistiky v lingvistickém výzkumu a ve výuce cizích jazyků. Základními tematickými oblastmi uvedenými v akreditačním podkladu k předmětu jsou:

1. Počítačová a korpusová lingvistika. Jazykový korpus. Typy korpusů – vstupní přednáška spojená s prezentací funkcí ČNPK.

2. Existující korpusy v ČR a v zahraničí –samostatné shromažďování a následná výměna informací z různých zdrojů – odborná literatura, internet.
3. Paralelní korpusy, jejich budování a využití – technické zpracování textů vhodných pro zařazení do paralelního korpusu (studenti se tak stávají „spolutvůrci" stávajícího korpusu).
4. Vytěžování korpusů, tvorba dotazů, konkrétní práce s jednojazyčným a paralelním korpusem – řešení zadaného dílčího výzkumného úkolu, jeho dokumentace a prezentace.
5. Využití korpusů ve výuce cizího jazyka, strategie zapojení korpusu do vyučovacího procesu, tvorba učebních materiálů a testů – návrh a dokumentace dvou cvičení (aktivit) se zapojením korpusu, jeho dokumentace a prezentace.

Předmět bude realizován kombinovanou formou. Základní podněty a instrukce získají studenti od učitele v přímé výuce, větší míra aktivity bude však požadována od studenta, který bude samostatně u svého počítače rešeršovat a řešit rozličné zadané úkoly. Souběžně s prvním zkušebním během kurzu bude vznikat interaktivní e-learningový kurz v prostředí Moodle, jehož prostřednictvím bude student komunikovat s učitelem i ostatními kolegy, získávat informace a také odevzdávat zadané úkoly. Zkušební semestr ukáže, zda navržená cesta byla vhodná.

Zájem studentů o práci s korpusy se projevuje každoročně při zadávání diplomových, bakalářských, závěrečných a ročníkových prací. Dodnes byla zpracována na základě vytěžených korpusových dat řada témat, např:

- *Syntaktische und semantische Analyse der deutschen und tschechischen Präpositionen* (série ročníkových a diplomových prací zabývajících se spojeními s předložkami *an, auf, bei, für, in, mit, durch, von, um* a jejich českými ekvivalenty)
- *Infinitivkonstruktionen als Transformationen der deutschen Nebensätze mit der Konjunktion dass und deren äquivalente Strukturen im Tschechischen* (diplomová práce)
- *Das Pronomen „Es", seine syntaktischen Funktionen und Äquivalente im Tschechischen* (diplomová práce)
- *Übersetzung von Okkasionalismen im Werk „Fimfarum"* (diplomová práce)
- *Einige tschechische Ortsnamen und ihre deutschen Äquivalente im ČNKP* (bakalářská práce)
- *Das Präfix und Präfixoid „Haupt-„ und ihre Äquivalente im Tschechischen* (bakalářská práce)
- *Suffixoide (z.B. „–werk" und „-zeug") und ihre Äquivalente im Tschechischen* (bakalářská práce)
- *Komposita mit der Basis „-maschine" und ihre Äquivalente im Tschechischen* (bakalářská práce)
- *Překladatelské postupy při řešení překladu vlastních jmen* (zatím zadaná bakalářská práce)
- *Frequenz der deutschen und tschechischen Satzbaupläne* (závěrečná práce)
- *Gründe für unterschiedliche Länge der deutschen und tschechischen Texte* (závěrečná práce)
- *Ausdruck der Vorzeitigkeit in den deutschen und tschechischen Temporalsätzen* (ročníková práce)
- *Stellung der Partikeln im Deutschen und im Tschechischen* (ročníková práce)
- *Das Subjekt im Deutschen und im Tschechischen* (ročníková práce)

Témata prací byla zpočátku zadávána živelně, na základě individuální dohody studenta s učitelem. S přibývajícími pracemi se začala rozvíjet myšlenka cíleného systematického týmového zpracovávání určitých tématických oblastí, které by mohlo vyústit ve vytvoření souhrnného kolektivního díla, např. kontrastivní gramatiky němčiny a češtiny. Prvním pokusem o metodologicky jednotné zpracování dílčích témat je série prací o německých předložkách.

Významnou oblastí využití korpusů je bezpochyby tvorba učebních materiálů.
Doposud na PdF MU vyšla dvoje skripta čerpající z korpusů („*Překladatelské semináře I - úvod a pasivní překlad*“ a „*Wortbildung - Umriss der Theorie mit Übungen*“), dokončují se skripta „*Lexikologie - Struktur und Übungen*“ a v přípravě je zcela nová koncepce skript syntaxe opírající se o korpusová data. Možnosti využití korpusů jak k sestavování učebních materiálů, tak i k výzkumné práci demonstrujme na konkrétních příkladech, z nichž budou patrny nejen klady, ale i některé nedostatky, se kterými při současném stavu těchto nástrojů musíme počítat.
Ideální stav nastane, kdy jeden korpus bude maximálně vyhovovat všem požadavkům uživatele. Od tohoto momentu jsme ještě několik let, spíš však dekád, vzdáleni. Z toho důvodu je vhodné rešerše v dostupných korpusech kombinovat.
Chceme-li např. poukázat na rozdílnost konotace stejné kontextově zapojené lexikální jednotky, nemusí paralelní korpus stačit:
Německé substantivum *Marder* se v ČNPK vyskytuje pouze 4x (z toho 1x jako proprium), vše vesměs bez příznaku.
V Mannheimském korpusu najdeme „lasiček“ 871. Zde už lze vybírat mezi negativními, pozitivními a nulovými konotacemi této šelmičky.
„*Zmrzlé muže*“ nenajdeme ani v ČNPK a ku podivu ani v ČNK. Jak se jmenují německy každý zvlášť a všichni dohromady zjistíme opět velmi rychle z Mannheimského korpusu:
(Úkol může znít např.: *Určete hierarchické vztahy lexémů „Bonifaz, Eisheilige, Pankratius, Servatius a Sopherl“*.)

K97/MAI.38743 Kleine Zeitung, 05.05.1997; "Aprilflöckchen verheißen immer Maiglöckchen":
Die Eisheiligen stehen uns noch bevor. Womit werden zwischen dem 12. und 15. Mai Pankratius, Servatius, Bonifaz und die kalte **Sopherl** uns beglücken? Ein schwacher Trost bleibt: Die bringen auch längst nicht mehr das, was ihnen nachgesagt wird. Schon wieder Die Statistiker: die haben registriert und aufgezeichnet, daß um diese Zeit die Temperatur ganz im Gegenteil steigt!

Současný ČNPK pro svou velikost zatím není vhodný k výzkumu otevřených jazykových struktur (lexikální systém, pragmatické aspekty). Potvrdila to mj. i drobná studie prezentovaná na ÖLT 2005 ve Štýrském Hradci k problematice intralingvální alonymie některých českých toponym (jednalo se o 30největším měst ČR). Z tohoto malého výzkumu vyplynul „konflikt“ dat, kdy v Mannheimském korpusu převládá užití exonyma (*Ostrau, Budweis, Iglau*), kdežto v ČNPK jsou častější česká endonyma (*Ostrava, Budějovice, Jihlava*).
Na druhé straně jsou možnosti výzkumu gramatických (a s gramatikou bezprostředně souvisejících jevů) dosud nebývalé. Například získání materiálu k celému soubor cvičení k německé slovotvorbě netrvá déle než několik hodin a studující mají možnost cvičit na autentickém materiálu (nikoliv na uměle vygenerovaných příkladech)[3]:

Ergänzen Sie die ersten Glieder der Farbkomposita..[4]
Da sah man den Thronfolger, **_____rot** im Gesicht und die Fäuste geballt, davoneilen.

[3] Káňa (2005)
[4] Zdroj: ČNPK

A teď běží odtamtud následník sám, je rudý a zatíná pěstě.
--
Das Wasser dieser Hexenkessel ist ganz verschieden gefärbt, **_____weiß**, **_____rot**, **_____blau**, **_____gelb**, oft auch hell wie Glas.
Voda těchto pekelných kotlů je různě zbarvena; bývá mléčná, ohnivě červená, blankytně modrá, sírožlutá, často také čirá jako sklo.

Ergänzen Sie fehlende Suffixe. [5]
Die **christ…**-soziale 'Partei der Arbeit' sprach **anfang…** nur von der Annexion Ostpreußens, die **bäuer…** 'Volkspartei' von der **Ausdehn…** Polens mit einer "breiten Ostseeküste" und bis zur Oder, die Nachfolgerin der 'Polnischen Sozialistischen Partei', **'Frei…, Gleich…, Unabhäng…'** , schaute mehr nach Osten als nach Westen und zielte auf **ethnograph…**, aber zugleich auf solche Grenzen, die Schutz sowohl vor dem sowjetischen Imperialismus als auch dem deutschen "Drang nach Osten" gewährten.

Ergänzen Sie den Richtigen Artikel.[6]
Manche Bewohner machten es übrigens besonders schlau: Sie ließen sich ihre neben der Bahntrasse liegenden Häuser von **....DB** teuer ablösen und kauften diese einige Jahre später…
--
Dann begann er seinen Dienst bei **....ÖBB**, war anfangs beim Brückenbau eingesetzt und in der Folge in der Werkstätte in Feldkirch tätig.

Vedle jednoduchého dotazu na **formu** (libovolně dlouhá kombinace znaků) je možné zvolit i dotaz „**tag**“ a „**lemma**“[7]. I přesto, že výsledky těchto dotazů mohou být značně nepřesné (tagging je proveden automaticky a korpus není desambiguován), hovoří rychlost z korpusu získaných dat oproti „ruční“ rešerši jednoznačně ve prospěch těchto moderních nástrojů.

Vedle korpusů psaného jazyka, o nichž zde doposud byla řeč, existují i korpusy mluvené řeči (přepisy mluvených textů). Některé z nich umožňují i poslech vybraných sekvencí (konkordancí) - pro němčinu např. *Bayerisches Archiv für Sprachsignale*[8].

Tvorbou a zpřístupněním elektronických korpusů tak paleta nástrojů výuky i výzkumu jazyka dostala zcela novou dimenzi. Záleží jen na uživateli, aby si zvolil vhodný korpus a naučil se z něj efektivně těžit.

Příspěvek vznikl v rámci výzkumného záměru MSM 0021620823.

Některé běžně dostupné korpusy:
Český národní korpus: http://ucnk.ff.cuni.cz/
Intercorp - projekt paralelních korpusů: https://trnka.ff.cuni.cz/ucnk/intercorp/
Korpora **IDS-Mannheim:** http://www.ids-mannheim.de/
Bayerisches Archiv für Sprachsignale - München (BAS):
http://www.phonetik.uni-muenchen.de/Forschung/Verbmobil/VerbDialogdeu.html
Das digitale Wörterbuch der deutschen Sprache - Berlin (DWDS): http://www.dwds.de
Wortschatz Lexikon - Leipzig: http://wortschatz.informatik.uni-leipzig.de/index.html
Slovenský národný korpus: http://www.juls.savba.sk/
British National Corpus: http://www.natcorp.ox.ac.uk/

Literatura:

[5] Zdroj: ČNPK
[6] Zdroj: IDS-Mannheim.
[7] Pelouškova -Káňa (20051,2)
[8] http://www.phonetik.uni-muenchen.de/Forschung/Verbmobil/VerbDialogdeu.html

ČERMÁK, František a kol.: Jak využívat Český národní korpus. Praha: Nakladatelství Lidové noviny, 2005.
ČERMÁK, František: Jazykový korpus: Prostředek a zdroj poznání. In: *Slovo a Slovesnost*, 56, 1995. S.119-140.
KÁŇA, Tomáš - PELOUŠKOVÁ, Hana: Tvorba, funkce a využití Česko-německého paraleleního korpusu. Praha: Ústav pro jazyk český Akademie věd České republiky, oddělení gramatiky, 2005[2]. (t.č. v tisku)
KÁŇA, Tomáš: *Wortbildung. Umriss der Theorie mit Übungen.* Vyd. 1. Brno: Masarykova Univerzita, 2005.
Kolektiv autorů: *Studie z korpusové lingvistiky*. Praha: Nakladatelství Karolinum, 2000.
PELOUŠKOVÁ, Hana - KÁŇA, Tomáš: Die Nutzung des tschechisch-deutschen parallelen Korpus. In *Grammatik und Kommunikation. Eine neue Herausforderung innerhalb des vereinigten Europas.* Vyd. 1. Trnava: Filozofická fakulta Univerzity sv. Cyrila a Metoda v Trnave, 2005[1].
PELOUŠKOVÁ, Hana - KÁŇA, Tomáš: Česko-německý paralelní korpus. In *Teoretické východiská a perspektívy vyučovania cj.* Vyd. 1. Bratislava: Retaas, s r.o., 2004.
PELOUŠKOVÁ, Hana - KÁŇA, Tomáš: Das tschechisch-deutsche parallele Korpus als effektives Mittel in der Sprachforschung und im Fremdsprachenunterricht. In *Könniggrätzer Linguistik - und Literaturtage.* Vyd. 1. Hradec Králové: Gaudeamus při Univerzitě Hradec Králové, 2003.
PELOUŠKOVÁ, Hana - KÁŇA, Tomáš: Paralelní korpus jako zdroj autentického jazykového materiálu pro výzkum i výuku jazyků. *Cizí jazyky*, Plzeň: Fraus, 2002, vol. 46, no. 2, s. 43-45.
PELOUŠKOVÁ, Hana - KÁŇA, Tomáš: Paralelní korpus jako zdroj autentického jazykového materiálu. In *Vyučovanie cudzích jazykov na základných školách.* Pedagogická fakulta Trnavskej univerzity: Pedagogická fakulta Trnavskej univerzity v Trnavě, 2002. s. 188-199. 2002.

Resumé/ Zusammenfassung

Das moderne Computerzeitalter bringt neue Möglichkeiten mit sich, die in der Sprachenforschung bisher stark begrenzt waren: Untersuchungen anhand von elektronischen Sprachkorpora.
Im philologischen Studium fehlen bis heute Fächer, die den Einblick in die Problematik der Korpuslinguistik erläuten würden. Der Artikel bringt den Entwurf einer Struktur des Faches „Korpuslinguistik für Germanisten an der Masaryk-Universität", stellt die wichtigsten elektronischen Korpora für tschechische Germanisten vor und beschreibt die Parameter des tschechisch-deutschen parallelen Korpus. Den Abschluss des Artikels bilden Beispiele für eine effektive Nutzung des Korpus im Unterricht, sowie Adressen einiger Korpora, die übers Internet zugänglich sind.

Kontakt:
PhDr. Hana Pelouśková, Ph.D.
Mgr. Tomáš Káňa, Ph.D.
Katedra německého jazyka a literatury PdF
Masarykova univerzita
Poříčí 9/11
603 00 BRNO
Tel.: +420 5 49 49 49 37
E-mail: pelouskova@ped.muni.cz; kana@ped.muni.cz