„Bohemistyka" 2013, nr 1, ISSN 1642–9893

Katarzyna VACULOVÁ
Praha

Použití korpusu ve výuce češtiny – některé interferenční chyby polských mluvčích v češtině zachycené v chybové databázi CHRUP

Klíčová slova: České korpusy, žákovský korpus, studijní korpus, L2 korpus
Keywords: Czech corpora, learner corpus, L2 corpus

Abstract

Příspěvek nastiňuje situaci korpusové lingvistiky v České republice a soustřeďuje se na otázku žákovských korpusů, především na nově vzniklou databázi chyb ruských, ukrajinských a polských mluvčích. Na tomto pozadí šířeji pojednává o jedné frekventované interferenční polské chybě v českém jazyce a na jejím příkladu uvádí možné použití databáze CHRUP ve výuce češtiny.

The article outlines the situation of corpus linguistics in the Czech Republic and focuses on the issue of learners' corpora, and especially on the newly created database of mistakes of Russian, Ukrainian and Polish speakers. On the basis of this background, it deals with one frequent Polish interferential error in the Czech language and on this example it shows possible use of the CHRUP database in teaching Czech.

Vznik korpusové lingvistiky je bezprostředně spojen s vývojem počítačové techniky, která pronikla i do této vědecké oblasti a zásadním způsobem zjednodušila excerpci a analýzu lingvistických dat. Proto bývá celá dostupná škála psaných a mluvených korpusů a v nich obsaženého, bohatého jazykového materiálu předmětem častých jazykovědných výzkumů. Avšak ve slovanském prostředí bývá mnohem častěji využíván spíše badatelský potenciál korpusů před možností jejich praktického zužitkování. Korpusy tak mimo jiné zdánlivě zůstávají nedoceněnou pomůckou při výuce cizích jazyků. Z tohoto důvodu bude předkládaná studie zaměřena právě na jejich použití

v pedagogické praxi. Nejdříve bude podán stručný přehled výsledků fungování české korpusové lingvistiky, jejíž kvantum a záběr jsou výjimečné nejen ve slovanském měřítku. Následně bude představena Databáze chyb v češtině mluvčích s prvním jazykem slovanským (dále také jen CHRUP) a její přínos pro polské bohemistické prostředí. Na závěr budou prezentovány některé nejčastější chyby objevující se v textech polských studentů, které tento korpus obsahuje.

Korpusy v České republice

Rychlý vývoj korpusové lingvistiky v České republice začal rokem 1994, kdy byl na Univerzitě Karlově zřízen Ústav Českého národního korpusu. Dnes vedle tohoto ústavu na vývoji korpusů v českém prostředí pracuje i mnoho dalších institucí.

Popisujeme-li situaci v oblasti vývoje korpusů českého jazyka, je třeba podotknout, že jich existuje nebývalé množství a fungují částečně nebo zcela nezávisle na sobě. Pod pojmem Český národní korpus (dále jen ČNK) funguje pouze korpus, který je vyvíjen na již zmiňovaném ústavu a je jedním z největších národních korpusů v Evropě (Čermák, Schmiedtová 2004, s. 152). Avšak ani ČNK není jen jeden soubor textů obsažený v jednom editoru, jak bychom si mohli představovat. Mezi základní složky ČNK patří diachronní korpus (DIAKORP) s texty od konce 13. stol. do r. 1989, několik synchronních psaných korpusů řady SYN, mezi nimiž najdeme jak žánrově vyvážené korpusy, tak korpusy publicistických textů. Jedná se o několik korpusů, které se liší stářím zahrnutých textů.

Za samostatnou zmínku stojí Pražský závislostní korpus (PDT). Je to korpus češtiny, který byl vyvinut na Matematicko-fyzikální fakultě UK na bázi textů ČNK a vedle morfologické obsahuje i syntaktickou a sémantickou anotaci včetně anotace aktuálního členění věty.

Mimo to existuje zde řada dalších korpusů (některé z nich jsou součástí ČNK), na jejichž tvorbě se kromě dvou výše zmíněných institucí podílí i mnoho dalších. V tomto nepřeberném množství nelze nejmenovat Pražský mluvený korpus (PMK), Brněnský mluvený korpus (BMK), korpus neformální mluvené češtiny (ORAL), Pražský závislostní korpus mluvené češtiny (PDTSC), korpus školní komunikace (SCHOLA), korpus odborných lingvistických textů (LINK), korpus soukromé korespondence (KSK), Olomoucký korpus mluvené češtiny (OMK), Pražský fonetický korpus (PFK) a Český akademický korpus (ČAK).

V českém prostředí jsou vytvářeny i dvojjazyčné korpusy, jmenujme především rozsáhlý projekt ÚČNK InterCorp. Ten obsahuje české texty spolu s jejich překlady do 27 jazyků (nebo z těchto jazyků do češtiny), z nichž 17 je opatřených morfologickou anotací. Vedle toho existuje také rozsáhlý Pražský česko-anglický závislostní korpus (PCEDT).

Je nutno podotknout, že vedle jednojazyčných a paralelních korpusů mateřského jazyka vznikají na české půdě i cizojazyčné korpusy, což je jednoznačným důkazem úrovně a výsadního postavení české korpusové lingvistiky. Byly zde mimo jiné vytvořeny korpusy dolní a horní lužičtiny (DOTKO a HOTKO) a také Pražský arabský závislostní korpus.

Novinkou na české půdě je vznik korpusů L2, jak je zvykem je nazývat v západní tradici (viz např. Granger 2012). V českém (a také obecně slovanském) prostředí nemají tyto korpusy tradici, a tudíž ani vžitý název. Lze se setkat s pojmy studijní korpusy (Čermák, Schmiedtová 2004, s. 154) nebo žákovské korpusy (Šebesta, Škodová 2012). Jedná se o mluvené nebo psané korpusy shromažďující materiál nerodilých mluvčích. Od standardních korpusů národních jazyků se liší tím, že jejich obsahem není konkrétní jazyk, ale mezijazyk, tj. jakýsi přechodný stav mezi jazykem mateřským a cílovým (více o teorii mezijazyka viz Selinker 1987 nebo 1991).

I když první známý žákovský korpus na světě vznikl už začátkem 90. let 20. století v Belgii, ve slovanském prostředí dosud jsou tyto korpusy přítomné jen sporadicky, znám je jen slovinský korpus menšího rozsahu (Šebesta, Škodová 2010). V této souvislosti je čerstvě vytvořený žákovský korpus češtiny jako cílového jazyka (CZESL) zcela výjimečným projektem velkého významu. Spojení „cizí jazyk" v jeho názvu nefiguruje, protože materiál byl získáván i

od romských a jiných bilingvních mluvčích, pro něž čeština nemůže být považována za cizí jazyk. Korpus obsahuje psané a v menší míře i mluvené projevy na všech úrovních znalosti češtiny v rozsahu celkem přes 2 miliony slov a je tak největším neanglickým žákovským korpusem (Šebesta, Škodová 2010). Jeho část je morfo- syntakticky anotovaná, a proto umožňuje rozsáhlou kontrastivní analýzu (další podrobnosti o tomto projektu viz např. Šebesta, Ško- dová 2012).

Databáze CHRUP – chyby ruských, ukrajinských a polských mluvčích

Kromě tohoto rozsáhlého žákovského korpusu vznikají v českém prostoru menší žákovské databáze, spoluautorkou jedné z nich je i autorka tohoto příspěvku. Jedná se o Databázi chyb v češtině mluvčích s prvním jazykem slovanským konkrétně zachycující chyby ruských, ukrajinských a polských mluvčích – odtud název databáze CHRUP[1]. Tato elektronická databáze vznikala v letech 2011–2013 v rámci dvouletého projektu financovaného Grantovou agenturou Univerzity Karlovy (číslo grantového projektu 286411) a bude v nejbližší době volně přístupná na adrese: http://chrup.ff.cuni.cz. V názvu nebylo záměrně použito slovo korpus, neboť byl tento nástroj plánován jako méně rozsáhlá a nekomplikovaná databanka poskytující první pomoc učitelům a studentům češtiny, nikoliv jako rozměrný, propracovaný korpus.

CHRUP obsahuje celkem materiály od 185 osob, z toho 57 ruskojazyčných, 74 ukrajinskojazyčných a 54 polskojazyčných mluvčích. Záměrně zde nebylo použito pojmu národnost, neboť východiskem k zařazení do jednotlivých kategorií v databázi byl uvedený mateřský jazyk. Národnost je pro tyto účely pojmem přespříliš symbolickým a individuálně vnímaným, navíc zde v první řadě šlo o rozlišení respondentů z hlediska jazykové biografie. Do databáze nebyli zařazováni bilingvní mluvčí češtiny s prvním jazykem polským, ruským

[1] Tento projekt z části navazuje na fonetickou Databázi mluvené češtiny cizinců s ruštinou jako prvním jazykem.

a ukrajinským. Ve všech případech se tedy jedná o osoby, pro něž je čeština cizí jazyk, převážně na nižších úrovních její znalosti.

Samotná úroveň jazykové kompetence respondentů nebyla předmětem dotazníku, jehož vyplnění bylo předpokladem zařazení textů do projektu, jelikož jde o velice subjektivní měřítko. Místo toho byla zjišťována délka studia češtiny, která byla spolu s pohlavím, věkem, mateřským jazykem obou rodičů i partnera, s informací o tom, zda respondent žije v ČR, a o důvodech, které ho ke studiu vedly, jedním z kritérií, podle kterých je umožněno vyhledávání. Vedle toho je možné hledat podle kategorií chybové taxonomie. Ta byla rozdělena do základních pěti skupin: písmo a pravopis, morfologie, syntax, styl a lexikum, které se s výjimkou kategorie stylistických chyb dále třídí. S ohledem na maximální uživatelskou přehlednost a jednoduchost je vyhledávání omezeno na volbu konkrétní kategorie, popř. více kategorií z předem připraveného menu. Není tedy potřeba studium manuálu, značek a způsobu zadávání komplikovaných řetězců, jak je tomu např. v případě ČNK.

Databáze obsahuje kratší texty (od každého mluvčího 1 až 5) s nevědeckou, každodenní tématikou. Takto nastavená volba tématu měla redukovat případný vliv předloh a zcela odstranit možnost zahrnutí citací do databáze. Jazykový materiál není automaticky morfosyn- takticky anotován a byl podroben výlučně ruční jednoúrovňové chybové anotaci. Chybová taxonomie a anotace žákovských korpusů jsou natolik komplikované otázky, že by měly být předmětem samostatné studie. Zde pouze podotkneme, že je automatická bezchybná anotace mezijazyka velice náročná až nemožná. Dokonce automatické značkování přirozeného jazyka a regulérních textů není snadné a obsahuje chyby způsobené mimo jiné tvarovou homonymií a není 100% úspěšné (o úspěšnosti tagování korpusu viz např. Skoumalová 2011). Více informací k otázce anotace chybových korpusů lze najít v pracích týmu korpusu CZESL (viz např. Škodová, Štindlová, Hana, Rosen 2011).

Výuka češtiny pro polské mluvčí

V současné době je nabídka učebnic a publikací věnující se výuce češtiny jako cizího jazyka velmi bohatá. Přesto polští mluvčí nemají k dispozici výukový materiál, který by umožňoval výuku češtiny na pozadí polštiny jako blízce příbuzného jazyka a jsou nuceni vycházet z učebních materiálů připravených pro anglické či německé mluvčí. Jedinou dle autorčina vědomí existující učebnicí češtiny pro Poláky je publikace od Batowského z padesátých let 20. století, která je samozřejmě jak s ohledem na tematiku textů, tak používanou metodiku z dnešního hlediska zcela nepoužitelná. Taková výuka je následně mnohem méně efektivní a mnohdy pro polského studenta češtiny únavná, jelikož v ní není kladen patřičný důraz na výuku a procvičování jevů, které jsou z jeho hlediska obtížné, naopak pro něj jednodušším otázkám je věnováno přespříliš pozornosti. Pochopitelně záleží na konkrétním učiteli, jak s učebnicí nepřizpůsobenou pro polské mluvčí naloží. Avšak vyžaduje to od něj mnoho vlastní invence a práce. V této situaci se jako užitečná pomůcka v přípravě hodin výuky češtiny jeví korpus.

Pomocí korpusů a databází učitel (nebo i student) může poměrně rychle zjistit, jaké chyby jsou typické pro polské mluvčí, a shromáždit potřebné kontexty a příklady, v nichž se objevují, aby je mohl použít ve výuce nebo ve studiu. V této souvislosti budou dále uvedena základní zjištění a možný způsob práce s databází CHRUP.

Chyby v češtině polských mluvčích

Jedním ze základních zjištění, které nám CHRUP poskytl, je skutečnost, že polským mluvčím k tomu, aby byli schopni tvořit základní, srozumitelné texty v češtině, stačí poměrně krátká doba studia češtiny, neboť už po půl roce jsou schopni takové texty tvořit i v psané podobě. Avšak i v pokročilejších stádiích výuky jejich projevy nejsou zcela bezchybné. V databázi bylo zaznamenáno mnoho opakujících se chyb. Nejčastější jsou chyby v kvantitě vokálů a interpunkční chyby. Vedle toho dělá polským studentům evidentní potíže český slovosled a některé deklinační vzory, zejména měkké. Hlavní problémy se objevují rovněž v oblasti lexika a stylistiky. Zde bude pozornost věnována

jen jedné z typických chyb, a to konkrétně jevu, který nenacházíme v učebnicích, avšak v textech shromážděných v databázi se běžně objevuje.

Jde o otázku používání ukazovacího zájmena *to*. Zatímco v češtině plní toto pronominum především základní syntaktickou substituční a deiktickou funkci (Čermák 2001, s. 183), v polštině vystupuje do popředí odkazování v širším slova smyslu. Toto demonstrativum nese v polštině totiž schopnost vyjádření deiktických kategorií osoby a času jako součást složeného predikátu. Plní tedy funkci spony, kterou v češtině může zastupovat pouze sloveso *být* nebo kategoriální slovesa *provádět*, *konat* atd. (Daneš, Hlavsa, Grepl 1987, s. 23). Proto jsou v polštině zcela správné věty jako: *Wiedza to potęga*. a *Czas to pieniądz.* (Nagórko 1998, s. 283). Nahrazení slovesa *být* kopulou *to* je možné i v jiných slovanských jazycích, ale výhradně v konkrétní stylistické funkci, např. v heslech a titulech (Dalewska-Greń 2002, s. 473). V polštině jde o užití ze stylistického hlediska zcela neutrální.

Pod vlivem polských vazeb se následně v češtině objevují tyto vazby (všechny příklady jsou uvedeny přímo z databáze, tj. se zachováním veškerých ostatních chyb ve větě):

Zdravý život to zdravé jdlo.

Hanka to jeho manželka.

Nejlepší povolání to lekář.

V polštině lze dokonce obě užití (substituční a sponové) umístit ve větě v těsné blízkosti, proto se pak můžou vyskytovat chybné české věty jako např.:

Druhá věc, která se mi tady líbí to to.

Demonstrativum ve významu spony je v polštině pravidelně používáno, pokud se podmět a jmenná část predikatu liší v kategorii čísla, zatímco v jiných jazycích včetně češtiny je v takových případech forma slovesa vyjadřena v množném čísle, por. čes. *Dítě jsou starosti* a pol. *Dziecko to kłopoty* (Dalewska-Greń 2002, s. 480). Příklady do-

kazující pokusy o aplikaci tohoto pravidla v českém kontextu nacházíme opět v databázi:

Sebeovládání, silné nervy, odvaha to vlasnosti.

Tradiční rodina to také prarodiče a jiné členové.

Vady toho modelu to časté konflikty mezi členy rodiny.

Jeho přátele to lesní zvířata.

Mimo to je v polském jazyce obvykle užití vazby demonstrativum *to* + sloveso *být* v případech, kdy autor vypovědí nechce použít instrumentál, proto např. místo *Te skrzypce były kiedyś pięknym instrumentem* uvede *Te skrzypce to był kiedyś piękny instrument* (Da- lewska-Greń 2002, s. 481). Přestože je v češtině oproti polštině běžné užití nominativu ve jmenné části přísudku (o konkurenci nominativu a instrumentalu viz např. Štícha 1980), v textech polských mluvčích se vlivem mateřského jazyka příklady takové aplikace předmětného pronomina objevují:

Prarodiče jsou to lidé.

Tak Česká republika to byla pro mě země krasných míst a pohádek.

Vedle toho se objevují konstrukce, které jsou kontaminací běžného deiktického použití možného také v češtině a polských konstrukcí s *to* jako kopulí.

Televize je to „lidské okno" na svět.

Katka je to krasná holka.

Protože se tento druh chyby objevuje téměř ve všech pracích, rovněž pokročilých studentů, lze konstatovat, že by měl být na něj kladen ve výuce větší důraz. Výše uvedená exemplifikace je pouze jedním z příkladů možností vyhledávaní a interpretace polských interferenčních chyb v češtině v databázi CHRUP jako potenciálního způsobu zefektivnění studia a výuky češtiny.

Literatura

Batowski H., 1950, *Podręcznik języka czeskiego*. Książnica Atlas, Wrocław––Warszawa.

Čermák F., 2001, *Jazyk a jazykověda*. Karolinum, Praha.

Čermák F., Schmiedtová V., 2004, *Český národní korpus – základní charakteristika a širší souvislosti*. „Národní knihovna knihovnická revue", roč. 15, č. 3, s. 152–168.

Dalewska-Greń H. 2002, *Języki słowiańskie*. PWN, Warszawa.

Daneš F., Hlavsa Z., Grepl M. 1987, *Mluvnice češtiny 3, Skladba*. Academia, Praha.

Granger S. 2012, *How to Use Foreign and Second Language Learner Corpora*. In: Mackey A., Gass S. M. eds. *Research Methods in Second Language Aquisition*. *A Practical Guide*. Wiley-Blackwell, Oxford, s. 7–29.

Nagórko A., 1998, *Zarys gramatyki polskiej*. PWN, Warszawa.

Selinker L., 1987, *Interlanguage*, In: Nehls D. ed. *Studies in descriptive linguistics.* Groos, Heidelberg, vol. 17.

Selinker L., 1991, *Rediscovering interlanguage*. Longman, London.

Skoumalová H., 2011: *Porovnání úspěšnosti tagování korpusu*. In: Petkevič V., Rosen A. eds. *Korpusová lingvistika Praha 2011, sv. 3: Gramatika a značkování korpusů*. Nakladatelství Lidové noviny/Ústav českého národního korpusu, Praha, s. 199–207.

Šebesta K., Škodová S., 2010, *Žákovský korpus a jeho využití pro češtinu jako druhý jazyk.*. In: *Sborník příspěvků z Konference – 20 let vývoje didaktiky cizích jazyků. Liberec 3. prosince 2010*, Fakulta přírodovědně-humanitní a pedagogická katedra románských jazyků Technická univerzita v Liberci, Liberec.

Šebesta K., Škodová S. a kol., 2012, *Čeština – cílový jazyk a korpusy*. Technická univerzita v Liberci, Liberec.

Škodová S., Štindlová B., Hana J., Rosen A., 2011, *Víceúrovňová anotace českého žákovského jazyka*. In: Petkevič V., Rosen, A. eds. *Korpusová lingvistika Praha 2011 – 3 Gramatika a značkování korpusů*. Nakladatelství Lidové noviny, Praha, s. 208–235.

Štícha F. *Konkurence nominativu a instrumentálu přísudkového substantiva v současné spisovné češtině*. „Naše řeč", roč. 63, 1980, č. 4.