PRVNÍ ČESKÁ MONOGRAFIE O ŽÁKOVSKÝCH KORPUSECH

Karel Kučera

ŠEBESTA, Karel – ŠKODOVÁ, Svatava et al. (2012): *Čeština – cílový jazyk a korpusy.* Liberec: Technická univerzita v Liberci, 166 s. ISBN: 978-80-7372-848-9.

Nedávno publikovaná monografie dvanácti autorů *Čeština — cílový jazyk a korpusy* sleduje tři hlavní cíle, a to seznámit českou učitelskou a odbornou veřejnost s obecnou problematikou žákovských korpusů, představit první žákovský korpus češtiny CzeSL (*Czech as a Second Language*), s jehož vznikem je většina autorů publikace spojena, a naznačit možnosti využití tohoto korpusu ve výzkumu i ve výuce. CzeSL,[1] jehož koncepci a vlastnostem je věnována podstatná část knihy, je při tomto pojetí monografie představen komplexním způsobem, který umožňuje uživateli nebo jinému zájemci — specializovanému i nespecializovanému — začlenit korpus do obecných historických, lingvistických a didaktických souvislostí a vytvořit si strukturovanou představu o jeho parametrech i o povaze dat a metadat, která nabízí. Vzhledem k tomuto výrazně (zdaleka ne však výhradně) informativnímu zaměření monografie se tato recenze nevyhnutelně vyjadřuje nejen k tomu, jak se autorům podařilo korpus představit, ale do jisté míry i ke korpusu samotnému.

Obecný typologický a historický kontext, do něhož první žákovský korpus češtiny vstupuje, je přehledně vymezen v úvodních dvou kapitolách publikace, jejichž autorem je Karel Šebesta. Po stránce typologické je zde CzeSL charakterizován jako druh speciálního akvizičního korpusu (tj. korpusu orientovaného na úzus mluvčích, kteří si příslušný jazyk teprve osvojují, respektive ho neovládají na úrovni srovnatelné s rodilými mluvčími odpovídajícího věku), a to jako korpus žákovský (ekvivalent rozšířeného anglického označení *learner corpus*), tj. korpus zaměřený na užívání jazyka nerodilými mluvčími, a tedy i na proces osvojování druhého/cizího jazyka, v daném případě českého. Historického kontextu, v úvodních částech publikace naznačeného stručným přehledem historie budování akvizičních a žákovských korpusů, využívají autoři monografie v dalším textu jako pozadí pro podrobné parametrické i koncepční srovnání korpusu CzeSL s existujícími zahraničními korpusy.[2]

Je nepochybné, že CzeSL, vznikající od roku 2009, v tomto srovnání plně obstojí jak po stránce kvantitativní, tak po stránce koncepční. Rozsahem své první zveřej-

1 Korpus CzeSL je pod názvem CzeSL-plain přístupný na stránkách Českého národního korpusu (<http://korpus.cz>), kde lze najít i jeho stručnou charakteristiku. Tato jeho zveřejněná verze má rozsah přibližně 2 000 000 slovních tvarů, z nichž zhruba 80 % připadá na projevy nerodilých mluvčích a 20 % na projevy romských dětí a mládeže.

2 Za počátek historie akvizičních korpusů se obvykle považuje rok 1983, v němž byly zahájeny práce na projektu CHILDES, který dnes představuje vůbec největší existující soubor akvizičních korpusů. Počátky žákovských korpusů spadají do 90. let minulého století a jsou spojeny jednak se vznikem komerčních korpusů užívaných britskými nakladatelstvími při tvorbě příruček pro studenty angličtiny, jednak se vznikem nekomerčního korpusu ICLE (International Corpus of Learner English) na Katolické univerzitě v Lovani roku 1990.

něné verze (viz pozn. 1) se CzeSL řadí k největším žákovským korpusům a je v tomto směru srovnatelný s odpovídajícími korpusy tak velkých jazyků, jako je francouzština, němčina nebo španělština.[3] Z kvantitativního hlediska přitom CzeSL navíc vyniká i rozsáhlým (v současné době mezi žákovskými korpusy možná vůbec nejrozsáhlejším) komplexem relevantních jazykových i nejazykových informací připojovaných jak k jednotlivým slovním tvarům nebo i větším úsekům textu, tak ke každému z korpusových textů jako celku.[4]

Tomuto komplexu informací, který staví CzeSL na přední místo mezi žákovskými korpusy i z kvalitativního hlediska, je v souladu s jeho závažností věnována významná část monografie. Autoři v ní představují především dva typy přidaných informací, které mají pro pedagogické i badatelské využití žákovských korpusů zásadní význam, a to jednak informace, které se uplatňují při vnějším značkování textů, jednak stratifikované chybové informace, jimiž je obohacován přímo text. První typ informací je soustředěn do souboru 44 parametrů, z nichž 4 jsou spojeny s textem, 16 se situací, v níž text vznikl, a 24 s osobností žáka (autora textu). Pro ilustraci lze uvést, že pomocí parametrů spojených s textem se zaznamenává *médium* (mluvený text × rukopis × text psaný na počítači), *převažující slohový postup* (postup informační × popisný × vyprávěcí × úvahově-argumentační), konkrétní *téma* a *typ tématu* (téma obecné × speciální/odborné). Parametry spojené se situací vzniku textu zachycují například i to, zda pro text byl zadán rozsah či časový limit, zda text byl součástí zkoušky, zda bylo pro jeho vypracování povoleno užití slovníku, popř. jiných pomůcek atp.

Problematika druhého typu přidaných informací, tj. problematika chybové anotace, je vzhledem ke své specifičnosti i vzhledem k mimořádnému významu, který má právě v žákovských korpusech, podrobně představena ve dvou nejrozsáhlejších kapitolách monografie. V první z nich, obecnější, se její autorka Barbora Štindlová věnuje okruhu problémů souvisejících s pojmem *jazyková chyba*, který má klíčovou úlohu jak z širšího hlediska (v celém empirickém studiu osvojování druhého jazyka), tak zejména z hlediska užšího (ve studiu tzv. mezijazyka, svébytného přechodného systému zakládajícího řečový projev žáka v druhém/cizím jazyce). V souvislosti s žákovskými korpusy se pozornost v této kapitole soustřeďuje především na taxonomii chyb jako základ jakékoli systematické chybové anotace, dále na podstatu základních anotačních modelů (lineární anotace × víceúrovňová distanční anotace), na jejich výhody i nevýhody a na charakteristiku jejich konkrétních aplikací ve vybraných zahraničních korpusech. Toto obecnější srovnání a teoretické pozadí představuje vhodnou základnu pro bezprostředně následující kapitolu, na níž se podílela šestice autorů (Vladimír Petkevič, Alexandr Rosen, Barbora Štindlová, Tomáš Jelínek, Milena Hnátková a Petr Jäger) a která podává celkový obraz chybové taxonomie a anotace v korpusu CzeSL.

3 Vzhledem k širokému mezinárodnímu uplatnění angličtiny i vzhledem k délce anglické korpusové tradice nepřekvapí, že angličtina jako druhý/cizí jazyk se z takového kvantitativního mezinárodního srovnání vymyká. V současné době je zachycena v množství žákovských korpusů (resp. korpusových celků diferencovaných podle prvního jazyka žáků), z nichž největší mají rozsah přes 30 milionů slovních tvarů.

4 Zmíněná první zveřejněná verze *CzeSL-plain* neobsahuje lingvistickou anotaci; texty s morfosyntaktickou a chybovou anotací budou zpřístupněny v jiném vyhledávacím rozhraní.

Jde o kapitolu, která podle mého názoru zasluhuje dvojí ocenění, z nichž jedno patří samotnému taxonomickému a anotačnímu systému a druhé jeho zasvěcené a současně přístupné prezentaci v monografii. Taxonomii chyb i anotaci realizovanou v českém žákovském korpusu koncipovali autoři tak, aby odrážela „specifické vlastnosti češtiny jako jazyka s bohatým flektivním podsystémem a volným slovosledem" (s. 62) a současně příliš nezatěžovala anotátory „teoretickými dilematy" (s. 64).[5] Výsledkem je tříúrovinová anotace, v níž základní rovinu tvoří (přepsaný) původní text, kdežto první anotační rovina (v textu označovaná jako R1) obsahuje emendaci izolovaných slovních tvarů, kterou lze provést bez ohledu na kontext (typicky jde o přepsání/překlepy a chyby v pravopisu a morfologii), a druhá anotační rovina (R2) zahrnuje emendaci ostatních chyb, z nichž nejčastější jsou chyby ve shodě, valenci a slovosledu. V rámci roviny R1 je vyčleněno 10 konkrétních druhů chyb jako např. nesprávná flexe (např. užití *spám* místo *spím*), nesprávný slovní základ (*posvětlit* místo *vysvětlit*), neemendovatelné/„vymyšlené" slovo (je tam hodně *jinaků*), slovo z cizího jazyka (jím rád *eggs*), obecněčeský tvar/podoba (*dobrej*), knižní/nářeční/slangový/hyperkorektní tvar (s hnědými *očimi*) atp. Na rovině R2 jde o 13 druhů chyb, k nimž patří např. narušení shody (Petr *vařím* oběd), chyby v zájmenném odkazu (paní, *jenž* jsem potkal), v analytickém tvaru (Jana bude *dělá*), v negaci (mám *žádný* čas), chyba v užití gramatické kategorie (celé dopoledne *uvařím* oběd) aj. Na obou anotačních rovinách se přitom počítá s výskytem zbytkové kategorie „problémových chyb", tj. chybných vyjádření, která nebude možno jednoznačně přiřadit k žádnému z explicitně vyčleněných druhů chyb.

Jak už bylo řečeno, prezentace systému chybové taxonomie a anotace se vyznačuje zasvěceností a současně přístupností. Za zmínku stojí ještě jeden její sympatický rys, totiž realistické vědomí nemožnosti beze zbytku podchytit nepravidelnost a potencialitu přirozeného jazyka, která je v případě žákovských textů navíc rozšířena o jen stěží předvídatelnou potencialitu chybování na straně žáka, popř. i na straně anotátora textu. Jistá střízlivost, která z tohoto vědomí vyplývá a za jejíž projev lze na elementární úrovni považovat i vytvoření zmíněné zbytkové kategorie „problémových chyb", ve skutečnosti prolíná celou koncepcí zvoleného anotačního schématu. Lze ji mimo jiné vysledovat i na úrovni nejvyšší, konkrétně v realistické formulaci (a také v pořadí) principů, na jejichž základě bylo toto výsledné anotační schéma vytvořeno:

> „Anotační schéma musí [...] vyhovovat následujícím požadavkům:
> 1. schéma musí být zvladatelné pro anotátory,
> 2. taxonomie nemůže být příliš rozsáhlá, ale zároveň musí být dostatečně informativní, tj. musí umožňovat dostatečně podrobné zachycení chyb,
> 3. taxonomie by měla umožňovat budoucí rozšiřování" (s. 62).

5 Tato zdánlivě přízemní zásada má ve skutečnosti klíčový význam a je nanejvýš vhodné, že ji autoři explicitně formulují. Spolehlivost závěrů veškerého výzkumu využívajícího anotace korpusu přirozeně záleží primárně na kvalitě a konzistentnosti anotace, a v případě manuálně anotovaných korpusů je proto třeba anotátory vybavit co nejjednoznačnějším (a v rámci možností i co nejjednodušším) anotačním schématem, které minimalizuje potřebu samostatného rozhodování v komplikovaných případech.

Příznačná je v této souvislosti skutečnost, že i poté, co použitelnost anotačního schématu vytvořeného podle uvedených tří principů byla s celkově uspokojivými výsledky ověřena statistickými testy, autoři si zůstali vědomi jak potřeby dalšího zdokonalování jednotlivých složek systému (zejména instrukcí v anotačním manuálu), tak hranic, které lze v současné době sotva překročit: jak konstatují, „některé značky budou i nadále do značné míry závislé na subjektivním dojmu anotátora a vysokou míru shody mezi anotátory u nich nelze očekávat" (s. 69).

Součástí kapitoly věnované chybové taxonomii a anotaci v korpusu CzeSL je i oddíl shrnující jednotlivé fáze zpracování textů. Celý proces začíná jejich přepisem do elektronické podoby a uložením do databáze, pokračuje jejich konverzí a manuální opravou chyb s pomocí editoru *feat* a končí kontrolou těchto manuálních oprav a automatickými úpravami chybového značkování realizovanými sadou počítačových programů. Zásadou přitom je, že každý text je nezávisle zpracováván dvěma anotátory, jeho zpracování je kontrolováno nezávislým supervizorem a zjištěné rozdíly v anotaci obou anotátorů odstraňuje další pracovník, tzv. adjudikátor. Celý proces od uložení textu do databáze až po kontrolu, adjudikaci a následné zařazení do korpusu je řízen systémem Speed určeným pro správu přepsaných, zkonvertovaných a anotovaných textů.

Samostatná kapitola je v monografii věnována problematice specifické složky korpusu CzeSL, kterou tvoří subkorpus ROMi. Jde o složku výjimečnou už tím, že se orientuje na projevy romské mládeže a dětí žijících v ČR, pro něž (na rozdíl od jinojazyčných žáků, na které je zaměřen zbytek korpusu CzeSL) čeština zpravidla není druhým/cizím jazykem, čímž ROMi striktně vzato do jisté míry vybočuje z definice žákovských korpusů. Zařazení ROMi do žákovského korpusu CzeSL však přesto má své oprávnění, neboť běžná/většinová čeština sice není pro romské mluvčí druhým jazykem, ale v pravém slova smyslu není ani jejich jazykem prvním (tím je tzv. romský etnolekt češtiny znatelně ovlivněný romštinou a slovenštinou). Toto mimořádné postavení romské češtiny spolu s faktem, že velká část romských mluvčích „žije v sociokulturních podmínkách, které hraničí se sociálním vyloučením" (s. 109),[6] pak stojí v pozadí dalších specifických rysů, kterými se vyznačuje jak sám korpus ROMi, tak jeho budování.

Přestože jazyková data procházejí před zařazením do korpusu ROMi v zásadě stejným procesem, který byl výše stručně popsán v souvislosti s projevy jinojazyčných žáků osvojujících si češtinu jako druhý jazyk, bylo při sběru i při zpracování těchto dat třeba hledat odpovědi na řadu nových otázek. Autorky kapitoly Zuzanna Bedřichová a Kateřina Šormová poskytují čtenáři základní vhled zejména do problémů spojených s výběrem respondentů (např. obavy potenciálních sběračů z nařčení z diskriminace; nemožnost přímo se dotazovat žáků na jejich romskou identitu atp.) a s etickými aspekty sběru jazykového materiálu — například se skutečností, že texty se mnohdy týkají zážitků nebo situací, které jsou pro většinovou společnost vzdálené nebo těžko představitelné a které by při určitém způsobu citace korpusových dokladů mohly vést ke konstruování zkresleného obrazu dané skupiny mluvčích. Spe-

6 Ke korpusu ROMi se proto v textu monografie odkazuje i obecněji jako ke korpusu mluvčích ohrožených sociálním vyloučením.

cifické problémy se však objevily také v souvislosti se získáváním metadat (v zájmu co největší spontánnosti projevů se často ukázalo jako vhodné nebo i potřebné omezit počet zaznamenávaných parametrů spojených s osobou respondenta aj.) a v souvislosti s přepisem získaných textů (nesnáze spojené s obtížnou čitelností psaných textů a s nízkou kvalitou nahrávek pořizovaných — opět v zájmu co největší spontánnosti projevu — v ne zcela vhodném prostředí). Pragmatická řešení naznačených problémů vyústila v některá konkrétní opatření, jimiž se podařilo budování korpusu ROMi podstatně zefektivnit (např. získání individuálních sběračů jazykových dat z řad romských pedagogických asistentů nebo z nevládních organizací spolupracujících s Romy). Vzhledem k naznačené zvýšené náročnosti práce na tomto korpusu je třeba zvlášť ohodnotit, že jeho materiálová základna (tj. souhrn jak psaných a mluvených projevů již zpracovaných a zveřejněných, tak projevů dosud nezpracovaných) svým rozsahem přes dva miliony slovních tvarů zhruba čtyřnásobně převýšila původní plán. Korpusu ROMi tak v rámci projektu CzeSL — ale i mimo něj — přísluší významné místo nejen proto, že je prvním žákovským korpusem orientovaným na češtinu romských dětí a mládeže, ale také proto, že se mezi našimi i zahraničními žákovskými korpusy řadí k největším. Dodejme, že ROMi má ještě jeden pozoruhodný rys: na rozdíl od běžných žákovských korpusů v jeho materiálové základně výrazně převažují mluvené projevy (přibližně 1 600 000 slovních tvarů) nad texty psanými (přibližně 450 000 slovních tvarů).

Vedle pěti kapitol, které jsou věnovány výše naznačené celkové charakteristice korpusu CzeSL, chybové taxonomii, anotaci a subkorpusu ROMi, obsahuje recenzovaná monografie ještě tři kapitoly zaměřené pedagogicky. První z nich, převážně teoretická, zčásti navazuje na výklad kapitol o chybové taxonomii a anotaci (předkládá mj. i jednu z možností elementární klasifikace morfologických tvarů na systémové, nesystémové a defektní), v centru pozornosti jejího autora Milana Hrdličky však stojí především sám pojmem *jazyková chyba* (včetně jeho vztahu k pojmům jako *jazyková správnost*, *norma*, *kodifikace* a *úzus*) a problematika práce s jazykovými chybami ve výuce cizího a zčásti i mateřského jazyka. Ostatní dvě kapitoly jsou zaměřeny výrazněji aplikačně. Pavlína Vališová představuje v kapitole *Využití korpusových dat při výuce češtiny jako cizího jazyka* několik typů induktivních cvičení, která byla vytvořena na principu tzv. Data Driven Learning a jsou určena k posílení aktivní role žáka při studiu cizího jazyka (konkrétní ukázky těchto cvičení spočívají ve vyhledávání jazykových informací v synchronních korpusech Českého národního korpusu). Kapitola *Nástin využití žákovských korpusů pro jazykové vyučování*, jejíž autorkou je Svatava Škodová, se zabývá zejména potenciálem a limity žákovských korpusů a podává stručný přehled dosavadních výsledků využívání žákovských korpusů v zahraničí. Lapidární konstatování, k němuž autorka dospívá v závěru kapitoly, lze podle mého názoru považovat za realistickou charakteristiku současného postavení žákovských korpusů a stavu jejich využití vůbec: „Časné publikace týkající se výzkumu žákovských korpusů explicitně zdůrazňovaly jejich kladný potenciál pro jazykové vyučování […], ale přímé využití korpusů v hodinách cizích jazyků je neobvyklé a dopad korpusové lingvistiky na sylaby nebo design materiálů je i ve světovém kontextu minimální. Právě tyto oblasti se jeví jako jedny z nejdůležitějších v následujícím bádání na poli aplikované lingvistiky opřené o data žákovských korpusů“ (s. 138).

Celkově lze bez výhrad konstatovat, že monografie *Čeština — cílový jazyk a korpusy* splnila cíle, které si kolektiv jejích autorů vytkl, a představuje tak nejen koncepční a uživatelskou charakteristiku korpusu CzeSL, ale i přehledné a s nadhledem zpracované uvedení do problematiky budování — a z jisté části i využívání — žákovských korpusů vůbec. Vzhledem k tomu, že jde o první takto souborně zaměřenou monografii u nás a současně o první žákovský korpus orientovaný na český jazyk, nezbývá než si v návaznosti na výše citovaná slova Svatavy Škodové přát, aby tato publikace našla své adresáty nejen mezi korpusovými lingvisty a odborníky zabývajícími se teorií výuky a osvojování druhého jazyka (popř. spisovné podoby jazyka prvního ve specifických případech typu romských žáků), ale především přímo mezi pedagogy a žáky, jimž korpus CzeSL nabízí nemalé možnosti využití při tvorbě učebních materiálů i přímo ve výuce.

Karel Kučera | Ústav Českého národního korpusu FF UK v Praze
karel.kucera@ff.cuni.cz