TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta přírodovědně-humanitní a pedagogická

Katedra českého jazyka a literatury

Karel Šebesta – Svatava Škodová
a kolektiv

Čeština – cílový jazyk a korpusy

Liberec 2012

Tato publikace vznikla v rámci projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk* (CZ.1.07/2.2.00/07.0259) podpořeného z Evropského sociálního fondu a státního rozpočtu České republiky. Projekt se uskutečnil v letech 2008–2012 na Technické univerzitě v Liberci, partnery projektu byla Univerzita Karlova v Praze a Asociace učitelů češtiny jako cizího jazyka.

Recenzenti: doc. PhDr. Marie Hádková, Ph.D., PhDr. Jiří Hasil, Ph.D.

ISBN 978-80-7372-848-9

Obsah

1. Cesty k žákovským korpusům

Karel Šebesta

Žákovské korpusy, resp. akviziční korpusy obecně, jsou v aplikované lingvistice a didaktice jazyka stále ještě nástrojem relativně novým. Akviziční korpusy (zahrnující v různé míře i data nerodilých mluvčích) začaly vznikat přibližně v polovině 80. let minulého století, samostatné korpusy žákovské (tj. specializované akviziční korpusy jazyka nerodilých mluvčích) přibližně od let devadesátých. Jejich využití však přináší natolik přesvědčivé výsledky, že je více než zřejmé, jak výrazné změny nejen ve studiu osvojování jazyka a pozdějšího jazykového vývoje žáků, ale také v jazykovém vyučování, v tvorbě jazykových slovníků, gramatik a učebních materiálů a dalších didaktických nástrojů jsou s nimi spojeny.

Akviziční korpusy můžeme vymezit především jejich funkcí. Jsou to korpusy, které *slouží primárně studiu procesů osvojování jazyka, včetně tzv. pozdějšího jazykového vývoje a užívání jazyka mluvčími, kteří (daný) jazyk neovládají na úrovni odpovídající úrovni dospělého rodilého mluvčího*. Sekundárně mohou plnit a plní řadu důležitých funkcí v didaktice jazyka: jsou významným zdrojem dat při tvorbě učebnic a učebních pomůcek, jako jsou slovníky nebo gramatiky, při přípravě testů a jazykových cvičení různého typu a uplatňují se i přímo jako didaktický nástroj v jazykové výuce.

Žákovské korpusy jsou jejich podtypem – zachycují *užívání jazyka nerodilými mluvčími*, a uplatňují se tedy při studiu procesů osvojování a užívání druhého/cizího (souhrnně cílového) jazyka, resp. v jeho didaktice.

Akviziční korpusy se opírají o dvojí tradici. Tu mladší představuje tradice korpusové lingvistiky, sahající přibližně do 60. let minulého století. Akviziční korpusy a jejich tvůrci využívají zejména technické a programové nástroje vyvinuté v korpusové lingvistice a sloužící zpracování jazykových dat a jejich prohledávání, v jisté míře i některé metodologické postupy při analýze dat a interpretaci výsledků.[1]

[1] Využitelnost korpusů pro jazykové vyučování byla badatelům a institucím působícím v oblasti jazykového vyučování zřejmá už při samém jejich vzniku. Již první korpus americké angličtiny, vytvořený H. Kučerou a N. Francisem koncem 60. let minulého století (tzv. *Brown corpus*), vyvolal zájem bostonského nakladatelství Houghton–Mifflin a byl využit při sestavování slovníku *American Heritage Dictionary*, prvního slovníku využívajícího korpusová data. Sepětí korpusové lingvistiky s praktickým využitím korpusů v aplikacích různého druhu v dalších letech nesláblo, spíše naopak. V 80. letech sehrál v tomto směru významnou roli kolektiv badatelů Birminghamské university a jejich projekt COBUILD, z něhož vzešla celá řada didakticky využitelných slovníků, gramatik a dalších příruček.

Vedle toho se však mohou opřít o více než dvousetletou tradici získávání a využívání záznamů řeči osob při studiu osvojování jazyka.[2] Ta nabízí badatelům jednak bohaté zkušenosti se získáváním jazykového materiálu a kontextových proměnných k němu i s jeho zpracováváním a analýzou, včetně ucelených a ověřených, byť z různých úhlů pohledu kritizovaných metodologických postupů, jako je kontrastivní studium jazyka nebo chybová analýza, jednak obecné teoretické rámce potřebné pro vyhodnocování a interpretaci výsledků.

Povaha dat zaznamenávaných badateli v oblasti osvojování jazyka a způsob jejich záznamu se měnily v závislosti na řadě faktorů, v neposlední řadě na dostupnosti technických prostředků pro záznam (neuvažujeme zde o datech experimentálně elicitovaných – k tomu viz kapitola 2). Zpočátku dominovaly záznamy deníkové; od 60. let s nástupem kvalitnější nahrávací techniky a s novým pohledem na osvojování jazyka, který přinesla generativní gramatika (především se zaměřením na jazykové univerzálie, na hledání systematičnosti ve formování jazykové kompetence dětí atd.), vznikají magnetofonové nahrávky založené na soustavném, dlouhodobém sledování jazykových projevů dětí a jejich vývoje.

Zpravidla šlo o sbírky s jednorázovým využitím; v posledních letech před vznikem akvizičních korpusů se však objevily nahrávky a přepisy, které svým rozsahem i způsobem využití už připomínají korpusy elektronické. Jejich autor, Roger Brown, nepoužil získaná a přepsaná jazyková data pouze pro svůj vlastní výzkum, ale rozmnožil je a dal k dispozici větší komunitě badatelů.[3] To byl významný impuls pro vytvoření první velké, mezinárodní elektronické databanky jazyka dětí – CHILDES; Brownovy nahrávky se po převedení do elektronické podoby ve standardizovaném formátu staly její první součástí. Databáze CHILDES dnes představuje zdaleka největší soubor akvizičních korpusů na světě, jak co do celkového objemu jazykových dat, tak co do počtu zastoupených jazyků a typů zaznamenaných textů.[4]

V posledních desetiletích se v didaktické oblasti začínají stejnou měrou uplatňovat korpusy akviziční, včetně korpusů žákovských.

[2] K tomu viz Šebesta, 2010, který uvádí i tradici domácí; k zahraniční, zvl. anglosaské tradici viz např. Ingram (1989), Aijmer (2009), Behrens (2008) aj.

[3] R. Brown z nich čerpal především pro svou zásadní práci z r. 1973 *A First Language: The Early Stages*. K šíření jeho přepisů a jejich využívání dalšími badateli uvádějí např. J. L. Sokolov a C. E. Snowová (Sokolov, Snow, 1994, s. 2): *Roger Brown made copies of transcripts from his study available to other researchers, using the state-of-the-art technology of the times, mimeography. The transcripts were typed, not onto paper, but onto mimeo masters, and as many copies as possible were then run off. Some masters generated more copies than others, and Roger with characteristic generosity had given copies away so freely that by 1983, when we went to collect a full set for inclusion in CHILDES, several sessions were down to the last copy, that one often embellished with marginal notes on negation by Ursula Bellugi, on morphological markers by Courtney Cazden, or other checks, codes, and analyses in unrecognized hands.*

[4] Práce na projektu CHILDES B. McWhinneyho a C. Snowové byly zahájeny v r. 1983. Dnes tato databáze zahrnuje řadu dílčích subkorpusů s celkovým objemem téměř 45 milionů slov od dětí mluvících 28 různými jazyky (uváděno k r. 2008). Pro srovnání – tento objem dat je téměř pětkrát větší, než je velikost druhého největšího mluveného korpusu (korpusu mluvené dánštiny, ten měl ve stejné době pouze 9 milionů slov; tamtéž), a devětkrát větší než objem třetího největšího, mluvené složky Britského národního korpusu, ten k uvedenému roku vykazoval rozsah 5 milionů slov (MacWhinney, 2008, s. 165–166).

Stejné metody studia osvojování jazyka i sběru a zaznamenávání dat se využívaly a využívají jak u jazyka prvního, tak i druhého/cizího. To platí i o akvizičních korpusech, které od počátku zahrnovaly a zahrnují jak jazyková data rodilých mluvčích, tak často v různé míře i mluvčích nerodilých.

Typickým dokladem je např. *Arizona Corpus of Elementary Student Writing*, vybudovaný na Northern Arizona University, který obsahuje více než 5000 esejů sebraných ze 40 tříd v 15 městech Arizony a napsaných studenty tří různých jazykových komunit: anglické, španělské a komunity Navajo (Biber et al., 1998), stejně jako např. *Michigan Corpus of Academic Spoken English* (MICASE), který zahrnuje přepisy mluvených projevů vzniklých v akademickém prostředí a obsahuje jak projevy mluvčích prvního jazyka, tak (v menší míře) mluvčích nerodilých, nebo analogický *British Academic Spoken English* (BASE*) corpus*, obsahující celkem 200 záznamů z různých kateder dvou britských univerzit a založený rovněž jak na angličtině jako prvním jazyku, tak v menší míře i na datech nerodilých mluvčích. Obecně akviziční zaměření má i projekt AKCES, *Akviziční korpusy češtiny* (viz dále).

Žákovské korpusy (learner corpora)[5] jako samostatný subtyp korpusů akvizičních se zformovaly jen o několik málo let později, v 90. letech minulého století.[6] Poprvé se termín *learner corpus* objevil v komerční sféře – začalo s ním pracovat nakladatelství Longman při tvorbě jazykových slovníků pro studenty angličtiny; na počátku éry akvizičních korpusů L2 byly tedy korpusy komerční.[7]

[5] Korpusy jazyky nerodilých mluvčích bývají vcelku jednotně označovány termínem *learner corpus* nebo jeho ekvivalentem v příslušném národním jazyce; méně často se setkáváme s alternativním označením *interlanguage corpus*. U nás se zpočátku pracovalo s termínem *korpus studijní* (Čermák, Schmiedtová, 2004), od r. 2009 se v souvislosti s budováním prvního korpusu tohoto typu pro češtinu začal uplatňovat termín *korpus žákovský*. Akviziční korpusy zaměřené na osvojování prvního jazyka jednotné označení postrádají. Někdy se setkáváme s termínem *korpus vývojový* (srov. McEnery, Xiao, Tono, 2006, s. 65 – *The term* ***learner corpus*** *is used here as opposed to a* ***developmental corpus****, which consists of data produced by children acquiring their first language)*, které je ovšem nepřesné, jindy s popisným pojmenováním *korpus jazyka dětí, korpus jazyka mládeže* apod. V této práci užíváme termín *akviziční korpusy*, pojednáváme-li o obecných charakteristikách korpusů tohoto typu; pro potřeby rozlišení pak užíváme termíny *akviziční korpus L1*, resp. *L2*, v druhém případě alternativně rovněž termín *korpus žákovský*. K užívané terminologii v této oblasti podrobněji viz Šebesta (2010).

[6] Nepočítáme-li data získaná od nerodilých mluvčích a zařazená do obecně zaměřených akvizičních korpusů, jako je např. CHILDES.

[7] Vedle korpusu *Longman Learner's Corpus* patří k nejznámějším komerčním korpusům tohoto typu korpus nakladatelství Cambridge University Press CLC (*Cambridge Learner*

Prvním známým nekomerčním korpusem zaměřeným výlučně na jazyk nerodilých mluvčích byl ICLE, *International Corpus of Learner English*, vytvářený od r. 1990 v Centre for English Corpus Linguistics (CECL) na Katolické univerzitě v Lovani, který od 90. let inspiroval řadu následovníků.

ICLE, budovaný pod vedením S. Grangerové, zahrnuje úvahové a argumentativní nebo literární eseje[8] o minimálním rozsahu 500 a maximálním 1000 slov napsané studenty angličtiny jako cizího jazyka ve třetím nebo čtvrtém ročníku vysokoškolského studia. Skládá se z dílčích subkorpusů pocházejících od studentů z různých zemí, resp. s různými prvními jazyky, velikostně vyrovnaných – každý má rozsah 200 000 slov. Stanoven byl rovněž minimální počet studentů pro každý subkorpus (200) a maximální velikost příspěvku každého z nich (1000 slov). Eseje mohli studenti psát ve svém volném čase nebo jako součást zkoušky, přípustné bylo použití slovníků nebo jiných příruček, např. mluvnic, nikoli však pomoc rodilých mluvčích angličtiny.

V první publikované verzi zahrnoval ICLE eseje získané od studentů z 11 zemí s různými prvními jazyky (včetně češtiny); v druhé zveřejněné verzi přibylo dalších 5 zemí a celkový objem jazykových dat vzrostl z 2,5 milionu na 3,5 milionu slov (v počtu esejů je to nárůst z 3640 na 6085 textů).

Projekt ICLE vedla S. Grangerová; ta vede i většinu lovaňských projektů navazujících na ICLE nebo se na jejich řešení alespoň podílí. Dnes uvádí CECL celkem 5 korpusových projektů: FRIDA, LINDSEI, LONGDALE, VESPA, TeMa, LOCNEC, dále jinak zaměřené korpusy PLECI a MULT-ED.

Korpus FRIDA (*French Interlanguage Database*) obsahuje texty psané ve francouzštině a je rozdělen do tří subkorpusů podle toho, zda je prvním jazykem žáků, od nichž texty pocházejí, angličtina, dánština nebo nějaký jiný jazyk.[9] LINDSEI (*Louvain International Database of Spoken English Interlanguage*) představuje mluvený protějšek korpusu ICLE, obsahuje tedy mluvený jazyk pokročilých studentů angličtiny s různými prvními jazyky. Jeho budování bylo zahájeno r. 1995 vytvořením souboru přepisů 50 rozhovorů se studenty angličtiny s francouzštinou jako prvním jazykem o celkovém objemu 100 000 slov. Dnes se uvádí celkem 11 takových komponent s různými prvními jazyky dokončených a zveřejněných a ještě větší počet ve fázi zpracování, všechny se stejnou strukturou: rozhovor na zadané téma, volná diskuse a popis obrázku. Srovnávacím korpusem pro LINDSEI je korpus rozhovorů s rodilými mluvčími angličtiny LOCNEC.[10]

Projekty LONGDALE (*Longitudinal Database of Learner English*) a VESPA byly zahájeny o 3 roky později než LINDSEI. Zatímco ICLE a LINDSEI jsou korpusy průřezové, LONGDALE je databáze založená na dlouhodobém sběru anglických projevů týchž mluvčích (s různými prvními jazyky) po dobu minimálně tří let, od prvního do třetího ročníku jejich studia na lovaňské univerzitě. Sběr materiálu probíhá minimálně jednou ročně a týká se textů různého druhu; v prvních třech letech sběry zahrnovaly pouze psané argumentativní eseje; studenti mohli volit ze čtyř různých témat a eseje měly stanovený rozsah od 500 do 700 slov.

Cílem projektu VESPA (*The Varieties of English for Specific Purposes Database*) bylo vytvořit databázi anglických textů různých žánrů (články, zprávy, disertační práce) vztahujících se k různým oborům (lékařství, biologie, ekonomie, jazykověda, právo atd.) a vytvořených nerodilými mluvčími s různými prvními jazyky a s různou úrovní zkušeností v psaní anglických textů tohoto typu, od studentů prvního ročníku univerzity po studenty doktorského studia.[11]

Další korpusy CECL už nelze označit jako korpusy žákovské, třebaže některé z nich vztah k jazykovému vzdělávání mají; především to platí o korpusu TeMa, zahrnujícím jazykový materiál učebnic angličtiny jako cizího jazyka o celkovém rozsahu více než 724 tisíc slov.[12]

ICLE předznamenal etapu budování žákovských korpusů i na dalších univerzitách v různých zemích světa, převážně v Evropě a na Dálném východě. Naprostá většina těchto korpusů zachycuje jako cílový jazyk angličtinu. Přesné údaje o světových žákovských korpusech je obtížné získat, protože jde o oblast velmi dynamickou, rychle se rozvíjející, a ne všechny existující korpusy jsou zveřejněny. Určitou představu o existujících žákovských korpusech si lze učinit na základě přehledu světových žákovských korpusů na stránkách CECL.[13] Z celkového počtu cca 90 tam uváděných korpusů je přibližně 60 věnováno angličtině jako cílovému jazyku a z celkového objemu cca 100 milionů slov (údaje o velikosti ovšem nejsou u všech korpusů uváděny) připadá na angličtinu více než 95 %.

Větší žákovské korpusy (zpravidla do 1 milionu slov) najdeme ještě u některých dalších větších indoevropských jazyků – němčiny, francouzštiny, španělštiny, italštiny. Jazyky střední a menší velikosti (počtem mluvčích) jsou zastoupeny žákovskými

Corpus). Angažovanost významných nakladatelství, která jsou zaměřena alespoň částí své produkce na vydávání učebnic, slovníků, mluvnic a dalších materiálů určených pro jazykovou výuku, v tvorbě akvizičních korpusů dokládá zřetelně jejich praktickou užitečnost.

[8] Literární eseje nemají tvořit více než čtvrtinu celkového objemu každého subkorpusu (vymezeného prvním jazykem studentů). Mimo okruh zájmu tvůrců ICLE jsou texty popisné či vyprávěcí a texty na technická témata. (Jak se uvádí v pokynech, není tedy vhodné zadat téma *The British Electoral System*, ale např. téma *The British Electoral System is no guarantee of democracy.*) Omezení na argumentativní a literární eseje se přeneslo i do některých dalších korpusů, které se metodicky o ICLE opírají; korpus češtiny toto omezení nepřebírá.

[9] Viz http://www.uclouvain.be/en-cecl-frida.html.

[10] Viz http://www.uclouvain.be/en-cecl-lindsei.html; http://www.uclouvain.be/en-cecl-locness.html.

[11] Viz http://www.uclouvain.be/en-cecl-vespa.html.

[12] Viz http://www.uclouvain.be/en-cecl-tema.html.

[13] Viz http://www.uclouvain.be/en-cecl-lcWorld.html. Pokud není odkázáno na jiný pramen, opírají se všechny číselné údaje v tomto textu o tento zdroj.

korpusy s významnějším objemem jazykových dat jen ojediněle; v citovaném přehledu jsou to např. švédština (The ASU corpus – *Andraspråkets strukturutveckling*), estonština (EIC – *The Estonian Interlanguage Corpus*) a nyní rovněž čeština (korpus CzeSL – *Czech as a Second Language*). Ze slovanských jazyků se dosud uvádí vedle českého korpusu pouze slovinský PiKUST o objemu 35 000 slov.

Repertoár prvních či výchozích jazyků je poněkud pestřejší, především proto, že některé korpusy, zvl. lovaňský a rovněž oba uvedené korpusy komerční, zahrnují jazyková data studentů s prvními jazyky různými. Jednoznačnou převahu však mají mluvčí jazyků dálněvýchodních, zejména čínštiny, s velkým odstupem pak japonštiny a korejštiny.

Od čínských mluvčích pochází téměř třetina dat v dosud známých nekomerčních žákovských korpusech. Převaha čínštiny jako prvního jazyka je dána především díky velkému korpusu angličtiny čínských studentů HKUST (*Hong Kong University of Science and Technology Learner Corpus*) o udávané velikosti 25 milionů slov, ale i řadě korpusů dalších, méně objemných, jako je SWECCL (*The Spoken and Written English Corpus of Chinese Learners*, cca 2 miliony slov), CLEC (*Chinese Learner English Corpus*, cca 1 milion slov), MSEE (*Corpus for Middle School English Education*, 2,3 milionu slov), TLCE (*The Taiwanese Corpus of Learner English*, cca 2 miliony slov) aj.

Situace na poli žákovských korpusů se samozřejmě rychle mění a přehled, o nějž tyto informace opíráme, nemusí být zcela spolehlivý; masivní převaha angličtiny jako jazyka cílového a čínštiny jako jazyka výchozího je však faktem těžko zpochybnitelným.

Čeština byla od 90. let v žákovských korpusech zastoupena pouze jako první, výchozí jazyk v jednom ze subkorpusů lovaňského ICLE s angličtinou jako jazykem cílovým (v obvyklém rozsahu 200 000 slov).

Situace se začala měnit v r. 2005, kdy byl na Univerzitě Karlově v Praze zahájen projekt AKCES/CLAC (*Akviziční korpusy češtiny/Czech Language Acquisition Corpora*), směřující k vybudování souboru akvizičních korpusů českého jazyka, včetně korpusu češtiny nerodilých mluvčích. AKCES je koncipován jako relativně volný komplex korpusů sloužících primárně potřebám studia osvojování jazyka v jeho třech základních podobách: (a) **osvojování prvního jazyka/prvních jazyků** v raném věku, (b) tzv. **pozdějšího jazykového vývoje** ve věku školním a (c) **osvojování jazyka druhého/cizího** (souhrnně cílového), ale také (d) potřebám studia **oslabování/rozpadu jazyka**, prvního i druhého.[14]

[14] Procesy spojené s oslabováním/rozpadem jazyka jsou – jako protějšek procesů jeho osvojování – významným zdrojem poznání hybných faktorů vývoje jazykové a komunikační kompetence člověka. Korpusy takto zaměřené jsou však zatím ve světovém měřítku jen velmi řídké (Yoshitomi, 2007).
Uvedené čtyři oblasti zájmu AKCESu jsou přirozeně ještě dále strukturovány; pozornost je např. věnována nejen osvojování češtiny jako cílového jazyka, ale rovněž fungování češtiny

Prvním zveřejněným korpusem AKCES byl korpus SCHOLA2010 (přepisy nahrávek vyučovacích hodin na českých základních a středních školách), zveřejněný na adrese http://ucnk.ff.cuni.cz/schola.php)[15] a korpus SKRIPT2012 (přepisy slohových prací českých žáků na různých úrovních školní docházky),[16] zatím ve fázi závěrečných úprav; oba vznikly s finanční podporou VZ MSM 21620825, vedeného Z. Starým.

Ve vazbě na AKCES byla zpracována koncepce prvního žákovského korpusu češtiny CzeSL (*Czech as a Second Language*), který vzniká od r. 2009 jako jeden z výstupů projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk* (číslo CZ.1.07/2.2.00/07.0259) v rámci OP Vzdělávání pro konkurenceschopnost, s finanční podporou Strukturálních fondů EU (ESF) a státního rozpočtu České republiky a ve spolupráci Technické univerzity v Liberci, Univerzity Karlovy v Praze a Asociace učitelů češtiny jako cizího jazyka.[17]

Předkládanou monografií se kolektiv autorů spjatých s uvedeným projektem snaží seznámit podrobněji českou veřejnost odbornou a zejména učitelskou s problematikou žákovských korpusů a v tomto širším kontextu představit rovněž první žákovský korpus češtiny CzeSL, jeho zaměření, parametry, vybavenost lingvistickými a zejména chybovými anotacemi a ukázat na možnosti jeho využití, v neposlední řadě ve výuce.

O specifických vlastnostech žákovských korpusů v porovnání s obecnými lingvistickými korpusy synchronními pojednává druhá kapitola, zároveň se v ní uvádějí některé podstatné charakteristiky světových žákovských korpusů a na tomto obecném pozadí jsou stručně představeny základní parametry CzeSL.

Následující čtyři kapitoly se věnují význačnému rysu jazyka nerodilých mluvčích, totiž jeho chybovosti, a to jednak z hlediska žákovských korpusů a práce s chybami při jejich budování, přesněji z hlediska systémů chybové anotace korpusů tohoto typu (tedy identifikace a emendace chyb, v omezenější míře jejich deskripce, vzác-

jako prvního jazyka při osvojování jazyků cizích, jazykovému vývoji mládeže ze sociálně a kulturně znevýhodněných komunit, dětskému bilingvismu apod.

[15] Korpus byl vytvořen na Filozofické fakultě Univerzity Karlovy v Praze, projekt probíhal s podporou výzkumného záměru MSM 0021620825 (Jazyk jako lidská činnost, její produkt a faktor); vedoucí projektu K. Šebesta, koordinátorka projektu H. Goláňová.

[16] Korpus vzniká rovněž na Filozofické fakultě Univerzity Karlovy v Praze, vedoucí projektu K. Šebesta, na koordinaci prací se podílely především J. Letafková a B. Jelínková, v závěrečné fázi též H. Goláňová a E. Hlaváčková.

[17] Příjemcem podpory je TU v Liberci, na řešení se partnersky podílí UK v Praze a Asociace učitelů češtiny jako cizího jazyka. Řešení se účastní několik pracovišť z obou univerzit: z TUL je to KČL Fakulty přírodovědně-humanitní a pedagogické, z UK ÚJOP a několik ústavů FF UK: ÚČJTK, ÚTKL, ÚBS a ÚČNK, dále řada studentů doktorského, magisterského i bakalářského studia. Velkou zásluhu na vzniku korpusu mají i četná pracoviště neakademická, především základní a střední školy z různých regionů ČR, občanská sdružení a řada individuálních spolupracovníků.

ně i evaluace), jednak z hlediska možností pracovat s chybami konkrétních skupin mluvčích ve výzkumu a v jazykovém vyučování.

Chybovou anotací ve světových žákovských korpusech se zabývá kapitola 3. Chybová anotace žákovských korpusů byla zatím uplatněna převážně u jazyků s poměrně chudou flexí a pevným slovosledem. Při chybové anotaci českých jazykových dat byl řešitelský tým nucen vyrovnat se se specifickými problémy, které s sebou nesou typologické charakteristiky češtiny a také povaha materiálu, který je do CzeSL zařazen (vysoce chybové texty začátečníků). Unikátní model vícerovinné chybové anotace, který byl s ohledem na tyto faktory vyvinut, zvolená taxonomie chyb, proces chybové anotace a jeho programové zajištění jsou představeny v kapitole 4.

Kapitola 5 pojednává obecněji o pojetí chyby v české lingvistické tradici a o hodnocení chyby v projevech rodilých a nerodilých mluvčích; chybovostí jedné skupiny mluvčích (dětí ze sociokulturně handicapovaných komunit) se zabývá kapitola 6, která představuje v této souvislosti jeden ze subkorpusů CzeSL – ROMi.

Poslední dvě kapitoly jsou věnovány využití žákovského korpusu v jazykovém vyučování; kapitola 7 podává nástin využití žákovských korpusů pro jazykové vyučování, kapitola 8 představuje specifičtěji možnosti využití korpusových dat při výuce češtiny jako cizího jazyka.

CzeSL bude zpřístupněn výzkumnému i pedagogickému využití v r. 2012. Věříme, že se stane dobrým a užitečným pomocníkem badatelům, studentům i učitelům.

2. Parametry žákovských korpusů a CzeSL

Karel Šebesta

Žákovské korpusy a akviziční korpusy obecně[18] patří mezi tzv. korpusy speciální. J. Sinclair, který s tímto termínem pracuje, je vymezuje jako takové korpusy, které nesplňují některý či některé z parametrů u obecných lingvistických korpusů synchronních očekávaných, a které proto nemohou sloužit jako zdroj dat pro popis normálního užití daného jazyka.[19] Tyto odlišnosti se týkají řady rysů: velikosti korpusu, výběru jazykového materiálu, zvláště jeho autenticity a reprezentativnosti ve vztahu k jazyku, repertoáru zaznamenávaných metadat atd. Ve všech těchto a některých dalších relevantních parametrech vykazují žákovské korpusy jiné hodnoty než obecné lingvistické korpusy synchronní.

V této kapitole si postupně postupně zejména (a) velikosti žákovských korpusů, (b) povahy sbíraných dat, zvl. toho, že jde o data mezijazyka (interlanguage), a jejich autenticity, (c) sledovaných metadat, tj. metadat vztahujících se k textu, k situaci jeho vzniku a sběru a k jeho autorovi, dále (d) specifik v zaznamenávání a dalším zpracování jazykových dat pro žákovské korpusy. Na tomto pozadí pak představíme příslušné parametry CzeSLu, korpusu češtiny nerodilých mluvčích.

[18] V dalším textu se budeme zabývat výlučně korpusy žákovskými a žákovským korpusem češtiny CzeSL. Naprostá většina uváděných skutečností se však týká ve větší nebo menší míře akvizičních korpusů obecně.

[19] Sinclair, 1996: *„a corpus is assumed to have certain characteristics attached, with default values. Unless stated, these characteristics are attributed to anything called a corpus. A corpus which has one or more non-default values for these characteristics is termed a special corpus: its title should specify its deviations from the assumptions… The special corpora are those which do not contribute to a description of the ordinary language, either because they contain a high proportion of unusual features, or their origins are not reliable as records of people behaving normally… Corpora of the language of children, geriatrics, non-native speakers, users of extreme dialects and very specialised areas of communication (like the heraldic blazon or the knitting pattern, or the auctioneer's pattern) should also be designated special corpora because of the unrepresentative nature of the language involved.“*

2.1. Velikost žákovských korpusů; korpusy průřezové a vývojové

Objemem jazykových dat se žákovské korpusy od obecných korpusů lingvistických liší velmi nápadně. Charakteristické je, že se jejich velikost vyjadřuje jen zřídka v milionech slov, většinou jen ve statisících nebo desetitisících, a počítají se dokonce i jednotlivé tisíce slov (*Pilot Arabic Learner Corpus*, korpus angličtiny jako druhého jazyka arabských mluvčích, např. vykazuje velikost 9000 slov, PiKUST, korpus slovinštiny nerodilých mluvčích, 35 000 slov apod.). V přehledu světových žákovských korpusů, který na svých webových stránkách zveřejňovalo donedávna CECL, měly ze 67 korpusů, u nichž byla uvedena velikost, pouze tři objem větší než 10 milionů slov.[20]

Žákovské korpusy milionové nebo několikamilionové jsou známy téměř výlučně pouze pro angličtinu jako cílový jazyk, korpusy neanglické se této velikosti přibližují zatím jen výjimečně. Pokud bychom tyto počty porovnávali s miliardovými počty slov ve velkých lingvistických korpusech obecných, musili bychom konstatovat, že jsou to hodnoty zcela zanedbatelné (např. nereferenční korpus SYN v rámci Českého národního korpusu uvádí velikost 1 300 milionů slov).

Malé objemy žákovských korpusů souvisí s velkou obtížností sběru dat od nerodilých mluvčích a náročností jejich zpracování. Sběr dat je obtížný nejen proto, že celkový objem jazykových projevů nerodilých mluvčích je sám o sobě ve srovnání s projevy rodilých mluvčích malý, ale také proto, že jsou tyto projevy většinou obtížněji dostupné a sbírají se v malých množstvích. To přirozeně žákovské korpusy znevýhodňuje ve srovnání s obecnými lingvistickými korpusy, které mohou získávat velká množství dat už v elektronické či oskenované podobě. Jejich získávání přirozeně rovněž naráží na četné překážky (srov. Čermák, 2011, s. 18), ve srovnání se situací při tvorbě žákovských korpusů znamená však tato možnost obrovskou výhodu.

K tomu musíme připočítat skutečnost, že projevy nerodilých mluvčích se sbírají zpravidla jako nahrávky nebo v rukopisné podobě, že se tedy musí manuálně přepisovat nejen materiál mluvený, ale i psaný (mnohé žákovské korpusy se textům psaným na počítači vyhýbají, protože by mohl být výsledek zkreslen automatickými opravami, nebo je alespoň doplňují texty rukopisnými). Náročnost sběru zvyšuje i potřeba zaznamenávat u každého mluvčího a textu rozsáhlý soubor sociologických a didaktických informací, které jsou pro využití korpusu relevantní; u obecných lingvistických korpusů jsou tyto údaje podstatně skromnější. (Při hledání v korpusu SYN2009pub můžeme např. zjistit, v kterém médiu a který den byl určitý text zveřejněn, ale nedostáváme ani neočekáváme informaci o autorovi nebo okolnostech vzniku textu, o tom, zda prošel korekturou či korekturami, zda bylo autorovi téma zadáno apod. Omezenější jsou i informace u korpusů mluvených, třebaže u nich už jsou zaznamenána základní data o mluvčím, jako je věk, vzdělání nebo regionální příslušnost.)

Větší velikosti zpravidla dosahují ty korpusy, které mají usnadněnu spolupráci se školami, jež jim jazykovou produkci nerodilých mluvčích mohou poskytnout. Zatím všechny známé žákovské korpusy, které přesáhly minimální rozsah a vykazují nyní více než 10 milionů slov, jsou založeny plně nebo z velké části na textech vytvořených jako součást jazykových zkoušek a jsou podporovány školami, které takové zkoušky organizují (dobrým příkladem je komerční korpus CLC, *Cambridge Learner Corpus*, o rozsahu cca 35 milionů slov, který zahrnuje texty ze zkoušek)[21], popř. jejich sestavovatelé se školami spolupracují na komerčním či jiném základě (např. druhý velký komerční korpus LLC, *Longman Learners' Corpus*, vzniká z esejů a textů ze zkoušek, které školy zasílají výměnou za slovníky z produkce Longman). Na textech vzniklých v rámci oficiálních zkoušek jsou založeny i větší korpusy nekomerční. Písemné práce vzniklé v rámci maturitních zkoušek jsou např. součástí jednoho z největších korpusů HKUST, *Hong Kong University of Science & Technology learner corpus.*

Rovněž větší žákovské korpusy mluvené se opírají o data vzniklá při školním testování mluveného projevu v angličtině – např. mluvená složka korpusu BICCEL, *Bilingual Corpus of Chinese English Learners*, čerpá z nahrávek získaných při národním testu mluvené angličtiny, korpus NICT JLE, *Japanese Learner English*, je rovněž založen na testech mluvené angličtiny a stejně tak i korpus SWECCL, *Spoken and Written English Corpus of Chinese Learners.*

Malý objem jazykových dat přirozeně znamená pro tvůrce korpusů významné omezení. Žákovské korpusy proto bývají často zaměřeny pouze na jazyk omezeného okruhu mluvčích, např. pouze jednoho prvního jazyka, pouze jedné či dvou úrovní ovládání cílového jazyka, omezují se na projevy malého počtu žánrů atp.

To zároveň limituje i možnosti badatelské práce s nimi a klade značné nároky na výběr vhodného segmentu nerodilých mluvčích a jejich jazykových projevů, na něž se plánovaný žákovský korpus zaměří, primárně přirozeně se zřetelem ke konkrétnímu badatelskému záměru, ale i s výhledem na navazující možnosti dalšího badatelského či didaktického využití.

Klíčovým rozhodnutím je především volba mezi korpusem *průřezovým* (transverzálním), tedy takovým, který zachycuje projevy různých žáků v jedné etapě jejich jazykového vývoje, *longitudinálním*, který zachycuje projevy téhož žáka nebo týchž žáků v různých etapách jeho či jejich jazykového vývoje, a *pseudolongitudinálním*,

[20] Data uváděna podle přehledu světových žákovských korpusů na adrese http://www.uclouvain.be/en-cecl-lcWorld.html.

[21] Na oficiálních webových stránkách (http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB) se uvádí, že CLC v současné době zahrnuje soubor více než 135 tisíc textů vzniklých jako součást některé ze zkoušek ESOL (celkem v deseti úrovních, od úrovně odpovídající úrovni A2 referenčního rámce). Tento objem odpovídá 135 tisícům studentů ze 190 různých zemí se 130 různými prvními jazyky.

který zachycuje projevy různých žáků v různých etapách jejich jazykového vývoje (oba posledně uvedené typy korpusů lze označit souhrnně jako *vývojové*).

Mezi dnes existujícími žákovskými korpusy jsou zastoupeny v zásadě všechny uvedené typy. Relativně řidší jsou korpusy longitudinální, které vyžadují pravidelné sledování týchž žáků po delší souvislou periodu, u cizinců to nebývá snadné. Longitudinální data uvádějí hlavně korpusy spojené se sledováním cizího jazyka u školních dětí nebo vysokoškolských studentů, jejichž dlouhodobější sledování je snazší. Např. BELC (*Barcelona English Language Corpus*) obsahuje data získaná od dětí a mládeže – písemné práce, ústní vyprávění, rozhovory a hraní rolí; data školních dětí obsahuje CYLIL (*Corpus of Young Learner Interlanguage*), o písemné i ústní projevy studentů se opírá LONGDALE (*LONGitudinal DAtabase of Learner English*). Longitudinální povahu mají také data obsažená v *Telecollaborative Learner Corpus of English and German Telekorp*, u něhož bylo dlouhodobé sledování usnadněno zvoleným médiem – jde o bilingvální korpus zahrnující počítačem zprostředkované výměny mezi 200 Američany a Němci.

Většina korpusů však má povahu korpusů průřezových nebo pseudolongitudinálních a smíšených. U nich je podstatné, které fáze jazykového vývoje nerodilých mluvčích zachycují. Pro studium vývoje jazykové kompetence žáků je důležité, aby osvojování jazyka bylo korpusovými daty pokryto pokud možno v co největší míře. U angličtiny nebo dalších jazyků vybavených větším počtem žákovských korpusů je to snazší, jejich souhrn pokrývá přinejmenším celé spektrum ovládání jazyka, od začátečníků po ty nejpokročilejší.[22] U jazyků menších, jako je čeština, kde nelze dostatečně velký počet žákovských korpusů v dohledné době očekávat, je účelné usilovat o co největší pokrytí alespoň jazykového vývoje od počátku.

2.2. Jazyková data v žákovských korpusech

Tvůrci obecných lingvistických synchronních korpusů vycházejí z více či méně zřetelně vyjádřeného předpokladu, že tyto korpusy, jsou-li adekvátně sestaveny, reprezentují daný jazyk jako celek nebo k tomu směřují. Reprezentativnost obecného lingvistického korpusu vůči celku národního jazyka je požadavkem obecně přijímaným jako cíl, k němuž tvůrci korpusu směřují, nikoli jako cíl, jehož by bylo v současné době možno dosáhnout. Srov. např. vyjádření F. Čermáka (2011, s. 16), který za lingvistický pokládá takový korpus, „který umožňuje vyvážené a reprezentativní zkoumání relativně celého jazyka, nikoliv jeho části, často v nezdůvodněných proporcích".

[22] Úplný obraz angličtiny jako druhého/cizího jazyka přirozeně nepodávají, ale takovou aspiraci žákovské korpusy ani nemají.

Podstatné je přitom nejen slovo „relativně", ale i poukaz na „(ne)zdůvodněné proporce". Kritéria, podle nichž se stanovují náležité proporce jednotlivých typů textů zařazených do korpusu, mohou být různá. U psané složky Českého národního korpusu je to zřetel k recepci, tedy k čtenosti jednotlivých typů textů (Čermák, Králík, Kučera, 1997). U korpusů mluvených je ovšem obtížné takové kritérium uplatnit, ale odpovídá mu snaha získat pro mluvený korpus pokud možno jazyk prototypicky mluvený.

2.2.1. Žákovské korpusy jako korpusy mezijazyka

Žákovské korpusy takovou ambici, přispívat ke zkoumání národního jazyka, nemají. Dalo by se dokonce říci, že nesměřují ani k studiu jeho, třeba nedokonalého užívání nerodilými mluvčími.

Současné koncepce osvojování cílového jazyka se sice v řadě bodů liší, shodují se však v tom, že jazykový systém zakládající řečový projev žáka, který si cílový (cizí/druhý) jazyk osvojuje, je samostatný, svébytný útvar, odlišný od jazyka cílového, velmi variabilní a dynamický, jehož vývoj i užívání se řídí zřetelnými, rozpoznatelnými zákonitostmi a závisí na celé řadě vnějších i vnitřních faktorů. Aby se vyjádřil jeho přechodný charakter, bývá tento systém v dnešních teoriích osvojování a užívání druhého jazyka označován jako **mezijazyk** (interlanguage).[23]

Předmětem zájmu badatelů při výzkumu osvojování druhého jazyka a také předmětem zájmu tvůrců žákovských korpusů je právě tento mezijazyk, jeho vývoj, užívání a prožívání, jeho proměnlivost a faktory, které ji ovlivňují. Proto také bývají konkrétní žákovské korpusy někdy alternativně označovány jako **korpusy mezijazyka** (srov. ARIDA – *Arabic Interlanguage Database*; FRIDA – *French Interlanguage Database*; CYLIL – *The Corpus of Young Learner Interlanguage*; ICCI – *The International Corpus of Crosslinguistic Interlanguage*; NICKLE – *The Neungyule Interlanguage Corpus of Korean Learners of English*; EIC – *The Estonian Interlanguage Corpus*). Rovněž jedna z nejrozšířenějších analýz jazykových dat založená na využití žákovských korpusů nese název **kontrastivní analýza mezijazyka** či mezijazyků (CIA – Contrastive Interlanguage Analysis).

Primární funkcí žákovských korpusů je sloužit jako zdroj dat pro empirické studium mezijazyka v jeho vývoji a v jeho variabilitě (k tomu viz dále, oddíl 2.3). Nepředpokládá se, že žákovský korpus bude reprezentativní vůči cílovému jazyku nerodilých mluvčích nebo jejich mezijazyku jako celku, to při obrovské variabilitě mezijazyka není dost dobře možné ani by to nebylo užitečné, ale že umožní studovat jeho proměnlivost ve vztahu k co největšímu počtu sledovatelných a pečlivě zaznamenávaných faktorů.

[23] Termín *interlanguage* poprvé použil H. Selinker počátkem 70. let; přibližně tomuto pojmu odpovídá termín *idiosynkratický dialekt* P. Cordera; podrobněji k tomu C. James (1998).

To také znamená, že uvažujeme-li u žákovských korpusů o vyváženosti, je tato **vyváženost vázána** nikoli na recepci, jak je tomu u obecných lingvistických korpusů synchronních, ale **na produkci** jazyka. To je další z jejich podstatných odlišností od obecných korpusů lingvistických. Důležitým kritériem kvality žákovských korpusů je pak míra zastoupení různých externích proměnných, které variabilitu a užívání mezijazyka ovlivňují, přirozeně v závislosti na tom, jakému výzkumu má žákovský korpus sloužit.

2.2.2. Autentičnost dat

Verbální produkce žáků je pouze jedním ze tří typů dat užívaných ve výzkumu osvojování jazyka obecně (tedy nikoli pouze vývoje mezijazyka). Vedle toho se pracuje ještě s **performančními daty**, která nezahrnují verbální produkci (lze je získat např. zaznamenáváním neverbálních reakcí prokazujících porozumění, měřením reakčního času, testováním úsudku žáků o gramatičnosti vyjádření apod.), a se zprávami žáků o učení se jazyku a jeho užívání, založenými na **introspekci** (Ellis, 2008, s. 912).

Jazyková produkce je zdroj využívaný ve výzkumu tradičně. Pracuje se přitom s produkcí značně rozrůzněnou, zejména podle míry její spontaneity a volnosti, resp. naopak řízenosti. K tomu se vztahuje další důležitý pojem užívaný v korpusové lingvistice – **autentičnost** jazykových dat. Obecné lingvistické synchronní korpusy jsou založeny, jak se uvádí, na autentických jazykových datech, tedy na datech, která vznikla v reálné, autentické komunikační situaci; data neautentická nejsou do takových korpusů zařazována (Sinclair, 1996).

Výzkum osvojování a užívání jazyka žáky pracuje s členěním jemnějším. Základní rozlišení na texty vzniklé ze skutečné komunikační potřeby a texty ostatní se vnímá jako příliš hrubé, protože nedovoluje přihlédnout k různým podmínkám, za nichž text vzniká a které mohou míru jeho autenticity ovlivnit. V pojetí užívaném v studiu osvojování a užívání jazyka žáky je autenticita rozložena na škále, na jejímž jednom pólu stojí **přirozené, volné, „neřízené" vyjadřování** v autentických situacích reálného života, ideálně jejich vernakulární projevy, na druhém pólu vyjadřování, které je jasně řízeno jinou osobou než mluvčím s primárním zřetelem k (jazykové) formě sdělení (**experimentálně elicitované** projevy).

V širokém pásmu mezi oběma póly se nacházejí projevy, které byly rovněž elicitovány jinou osobou (rodičem, učitelem, badatelem, sběračem), ale s pozorností zaměřenou na obsah či funkci sdělení, nikoli na jeho jazykovou formu. Tyto projevy bývají označovány jako projevy **klinicky elicitované** (Ellis, Barkhuizen, 2005, s. 23). Do této kategorie náleží mj. rovněž eseje a slohové práce vytvářené pro školní účely. Míra jejich řízenosti může být různá v závislosti na řadě faktorů, především na detailnosti a způsobu zadání, na povaze činností organizovaných před vlastní verbální produkcí, na charakteru, síle a frekvenci intervenčních zásahů učitele nebo výzkumníka při produkci atd.

Žákovské korpusy mohou jen velmi zřídka čerpat z dat skutečně autentických, zejména pokud se neomezují na produkci žáků na nejvyšších úrovních ovládání cílového jazyka („přirozená" produkce žáků v cílovém jazyce je zejména na nižších úrovních jeho ovládání značně omezena nejen kvantitativně, ale i funkčně, tematicky a žánrově) a nezískávají ji v tzv. přirozeném prostředí (pokud se žák učí cizí jazyk mimo příslušné jazykové prostředí, užívá ho aktivně téměř výlučně pouze při výuce).[24] Zaměřují se většinou na sběr dat klinicky elicitovaných, získaných v typicky školních situacích – jak jsme uváděli, jde většinou o eseje psané v rámci zkoušek, o ústní projevy při zkouškách, dále o eseje psané přímo pro korpus či interview namluvená rovněž pro korpus.

Někteří badatelé, snad pod vlivem požadavku autenticity jako obligatorního pro jazykové korpusy (Sinclair, 1996), pokládají i tento typ dat za data autentická,[25] s tím, že např. psaní esejů jako součást výuky je rovněž autentická aktivita – ve třídě. Takový přístup však stírá principiálně podstatný rozdíl mezi daty vzniklými v přirozených situacích mimo výukový kontext a daty spjatými ve větší nebo menší míře s výukou (u nich můžeme přinejmenším očekávat tlak prostředí směrem k užívání mezijazyka co nejbližšího jazyku cílovému) a vede k tomu, že se někdy i texty čtené nahlas mohou pokládat za autentická mluvená data.[26]

Čtené texty a čtené seznamy izolovaných, spolu nesouvisejících slov jsou zařazovány jako jedna ze složek do řady mluvených korpusů. Souvisí to s potřebou získat materiál pro studium zvukové stránky mezijazyka a její proměnlivosti, zvl. v závislosti na péči (míře pozornosti), kterou mluvčí zvukové stránce svého projevu věnuje. Tak např. ISLE (*Interactive Spoken Language Education*) *speech corpus* uvádí jako zdroj svých dat mj. čtení jednoduchých vět a minimálních párů; korpus ESCCL (*English Speech Corpus of Chinese Learners*) hlasité čtení dialogů; LeaP Corpus (*Learning Prosody in a Foreign Language*) čtení seznamu izolovaných slov a čtení krátkého příběhu (vedle převyprávění a volného mluveného projevu při rozhovoru), *Learners' Corpus of Reading Texts* nepřipravené čtení anglických textů – krátkých výtahů z krásné literatury nebo umělých dialogů.

Data založená na čtených textech či izolovaných slovech jsou přirozeně v mluvených korpusech, zejména foneticky zaměřených, velmi užitečná, jejich označení jako

[24] *Even the most authentic data from non-native speakers is rarely as authentic as native speaker data, especially in the case of EFL learners, who learn English in the classroom. We all know that the foreign language teaching context usually involves some degree of „artificiality" and that learner data is therefore rarely fully natural.* (Granger, 2002, s. 8)

[25] S. Grangerová (Granger, 1998, s. xxi) např. uvádí, že výzkum opřený o žákovské korpusy *„uses the methods and tools of corpus linguistics to gain better insights into authentic learner language"*; srov. i tamtéž, s. 4–5.

[26] Srov. podmíněné vyjádření S. Grangerové (Granger, 2002, s. 16): *In as far as essay writing is an authentic classroom activity, learner corpora of essay writing can be considered to be authentic written data, and similarly a text read aloud can be considered to be authentic spoken data.*

dat autentických je však zavádějící – srov. též stanovisko F. Čermáka k zařazování např. rozhlasových pořadů do mluvených lingvistických korpusů (Čermák, 2011, s. 16).

Obecně lze říci, že dnes budované a užívané žákovské korpusy jsou založeny v naprosté většině na datech klinicky elicitovaných s různou mírou řízenosti. Výjimkou jsou korpusy vytvořené z textů vzniklých např. jako kvalifikační práce, anotace výzkumného projektu, výzkumná zpráva apod. U takových korpusů se ovšem obtížně sledují všechny faktory, které mohly výslednou podobu textu ovlivnit, např. vliv vzoru, konzultace s rodilým mluvčím, revize textu korektorem apod.; je proto potřeba s nimi pracovat samostatně a formulovat závěry na jejich základě zdrženlivě.

2.3. Variabilita mezijazyka a zaznamenávané parametry

Jedním z charakteristických znaků mezijazyka je jeho mimořádná variabilita, větší než variabilita jazyka prvního (Selinker, Gass, 2008, s. 259). Výzkum této variability se stal významnou součástí studia osvojování a užívání cílového jazyka přibližně od 70. let, tedy od změny paradigmatu v pohledu na osvojování druhého jazyka, signalizované formálně užíváním pojmu mezijazyk a s ním spojenými koncepty a spojené s poznáním, že žáci s různými prvními jazyky osvojující si týž druhý/cizí jazyk procházejí v zásadě týmiž vývojovými fázemi bez ohledu na podmínky jeho osvojování (Romaine, 2003, s. 409–410). To pak vedlo spolu s rozvojem variační sociolingvistiky[27] k podrobnému zkoumání variability v užívání druhého/cizího jazyka a faktorů, které ji ovlivňují, mj. i ve srovnání s variabilitou prvního jazyka dětí a mládeže z různých sociálně význačných prostředí, s vývojem jazyků kreolských, pidžinizací jazyka apod. Dnes patří proměnlivost žákovského mezijazyka a její možné příčiny mezi centrální témata výzkumu jeho osvojování a užívání nerodilými mluvčími.

Současné koncepce variability užívání jazyka rozlišují několik jejích typů; podrobnou typologii jazykové variability lze nalézt např. u R. Ellise (2008, s. 129). Z hlediska žákovských korpusů a jejich vytěžování je účelné především odlišení tzv. variability volné a variability systematické. Oba typy se obvykle spojují s různými fázemi osvojování daného jazykového jevu (srov. např. Selinker, Gass, 2008, s. 277); volná variabilita se objevuje ve fázi druhé, kdy si žák osvojil např. dvě formy pro vyjádření dvou významů, ale používá je záměnně, aniž by bylo možné shledat v jejich užívání nějakou pravidelnost, ať už by byla dána čímkoli. O systematické variabilitě mluvíme tehdy, vykazuje-li užívání dvou různých forem více či méně zřetelnou pravidelnost, podmíněnou nějakými vnějšími faktory; tento typ variability se vyvíjí později. Systematická variabilita přitom může záviset na faktorech povahy jazykové, psycholingvistické i sociolingvistické.

Právě pro odlišení variability volné a systematické a pro studium různých jejích druhů je podstatné zaznamenávat v žákovských korpusech co nejširší okruh vnějších okolností, které žákovskou produkci doprovázely. Těchto okolností, faktorů, které užívání jazyka řídí a ovlivňují jeho variabilitu, je celá řada. Na prvním místě jsou to přirozeně okolnosti spojené s vlastním **textem** (psaný vs. mluvený, žánr, téma apod.), dále okolnosti spojené se **situací** jeho vzniku a sběru (longitudinální vs. průřezové záznamy, stupeň připravenosti, povaha přípravy, způsob zadání, možnost využít pomůcky, časové omezení, omezení rozsahu, sběrač apod.) a okolnosti spojené s **mluvčím** (jeho věk, úroveň znalosti cílového jazyka, první jazyk, další jazyky, které ovládá, způsob učení se cílovému jazyku apod.). Při plánování korpusu se tyto faktory berou v úvahu jako kritéria určující výběr materiálu a způsob jeho sběru, při jeho budování jsou jejich parametry zaznamenávány jako metadata, při práci s korpusem jsou to základní proměnné, o něž se může výzkum opřít, především výzkum zaměřený na studium variability mezijazyka, resp. užívání jazyka nerodilými mluvčími.

Za ideálních podmínek by žákovské korpusy měly poskytovat dostatečná data pro studium všech typů variability jazyka žáků, resp. měly by výzkum této variability a její vazbu na jazykové, sociální a psychologické faktory usnadnit tím, že zahrnou různá textová data vytvořená pokud možno různými žáky za různých situačních podmínek a že tyto faktory budou bedlivě monitorovat a zaznamenávat.

To se však daří naplnit jen zčásti. Je to dáno, jak jsme uvedli, značnou náročností získávání a zpracovávání jazykových dat a potřebou vytvořit soubory dostatečně velké. Nemalá část korpusů je proto v jednotlivých parametrech značně omezena a výběrový je rovněž repertoár zaznamenávaných metadat. U žákovských korpusů se např. nesetkáváme často s tím, že by zaznamenávaly vernakulární projevy nerodilých mluvčích, tedy projevy zcela neformální, v nichž je užití jazyka po formální stránce věnována minimální pozornost. V tom se žákovské korpusy do jisté míry liší od akvizičních korpusů jazyka prvního, které se na vernakulární jazyk žáků zaměřují, i když ani ony ne výlučně.

Přes tato omezení však platí, že žákovské korpusy jsou vždy vybaveny podstatně větším rozsahem informací o textu, o podmínkách jeho vzniku a sběru a o jeho autorovi, než je obvyklé u lingvistických korpusů obecných. Relativní bohatství těchto metadat nejen odlišuje žákovské (resp. akviziční) korpusy od obecných lingvistických korpusů synchronních, ale je také důležitým měřítkem jejich hodnoty. Lze říci, že čím větší počet metadat o textech, podmínkách jejich vzniku a sběru a o jejich autorovi korpus obsahuje, tím větší možnosti využití badatelům v oboru i učitelům nabízí.[28]

[27] W. Labov a na něho navazující variační sociolingvistika přinesli do studia sociálně podmíněné stylistické variace jazyka žáků zásadní vklad, relevantní také pro žákovské korpusy.

[28] Je ovšem potřeba dodat, že dosavadní výzkum nevyužívá plně všech možností, které v tomto směru žákovské korpusy badatelům nabízejí, a soustřeďuje se převážně na sledování vlivu

2.3.1. Parametry spojené s textem

2.3.1.1. Médium (psané a mluvené žákovské korpusy)

Podíl psaného a mluveného jazyka v žákovských korpusech je, podobně jako u obecných lingvistických korpusů, dán náročností získávání a zpracovávání mluvených dat. U mluvených korpusů žákovských k tomu přistupuje i skutečnost, že badatele nutně zajímají i specifika výslovnosti nerodilých mluvčích a jejího vývoje, že je tedy i přepis jejich mluvených projevů náročnější než přepis mluvených projevů v lingvistických korpusech obecných.

Zastoupení mluvených žákovských korpusů je proto výrazně menšinové. Mezi známými světovými žákovskými korpusy představují korpusy mluvené méně než čtvrtinu a relativní celkový objem mluvených dat v žákovských korpusech je ještě podstatně nižší, protože mluvené korpusy jsou až na některé výjimky malé.

K největším nekomerčním mluveným korpusům patří už uváděný japonský korpus mluvené angličtiny NICT JLE (*National Institute of Information and Communications Technology – Japanese Learner English Corpus*), který obsahuje data získaná při mluvených zkouškách o celkovém rozsahu cca 2 miliony slov; přibližně stejně velký soubor dat obsahují korpusy FLLOC (*French Learner Language Oral Corpora*), MICASE (*Michigan Corpus of Academic Spoken Language*; ten ovšem obsahuje převážně data rodilých mluvčích, projevy nerodilých mluvčích jsou v menšině) a také korpus LINDSEI. Většina mluvených korpusů však má spíše rozsah počítaný v desetitisících slov, např. LeaP (*Learning Prosody in a Foreign Language* – 73 tisíc slov), ARIDA (*Arabic Interlanguage Database* – 8 tisíc slov).

Mluvené korpusy také často obsahují, jak jsme uváděli, projevy čtené nebo čtené seznamy izolovaných slov, není tedy možné využívat je v plném rozsahu pro zkoumání jiné než zvukové stránky užívání jazyka nerodilými mluvčími.

Řidčeji se mezi žákovskými korpusy objevují korpusy založené na jiném médiu, především korpusy zachycující komunikaci zprostředkovanou počítači, např. *Padova Learner Corpus* nebo *Telecollaborative Learner Corpus of English and German Telekorp*. Jde o typ korpusů, který je nepochybně velmi užitečný, zatím ale jen málo rozšířený.

2.3.1.2. Žánr, styl a téma textu

Většina žákovských korpusů je založena, jak jsme uváděli, na textech získávaných v typicky školních situacích: jde o eseje nebo rozhovory při zkouškách, popř. eseje a rozhovory či monologické projevy získané v podobných situacích přímo pro korpus. Některé korpusy sahají ještě k dalším žánrovým, stylovým, popř. s nimi spojeným tematickým omezením. Motivaci k tomu můžeme hledat ve snaze tvůrců získat rozsáhlejší stylově homogenní materiál; výraznou nevýhodou je skutečnost, že se tím omezuje variabilita jazyka žáků – takový korpus může sloužit k posouzení pouze relativně úzkého výseku jejich komunikační kompetence.

Pokud jde o žánr a styl, mezi korpusy mluvenými a psanými jsou určité rozdíly. O psaných korpusech lze říci, že jsou dnes stylově i žánrově poměrně různorodé; zejména korpusy založené na sběrech esejů vytvořených v rámci zkoušek nebo při výuce k žádnému dalšímu žánrovému či stylovému omezení nesahají.

Část korpusů dává ovšem přednost sběru argumentativních esejů, někdy v kombinaci s eseji literárními (korpus ICLE, LOCNESS a další spojené s centrem CECL; argumentativní eseje tvoří také podstatnou část německého korpusu FALKO – *Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache*, slovinského korpusu PiKUST a dalších), ale i narativními (lundský korpus CEFLE – *Corpus Écrit de Français Langue Étrangère*; MLC – *Multilingual Learner Corpus*) nebo popisnými (*Israeli Learner Corpus of Written English*; Chy-FLE – *Cypriot Learner Corpus of French*).

Důvody pro sběr argumentativních textů se většinou neuvádějí. V jejich prospěch snad mluví známá skutečnost, že argumentativně-úvahové texty vedou zpravidla žáka k větší angažovanosti a zaujetí, takže je naděje, že svůj projev pravděpodobně méně kontroluje po stránce formální. Stejné důvody mluví i pro volbu témat emotivně laděných a osobních a v neprospěch témat neutrálních, resp. racionálních a neosobních.

Zdá se (spolehlivé statistiky v tomto ohledu není možné vytvořit, opíráme se tedy o odhad), že eseje, texty argumentativně-úvahové, narativní a popisné v žákovských korpusech převažují. Vedle nich se objevují i texty informativní, výzkumné zprávy, formální dopisy, žurnalistické texty, disertace, ročníkové práce a další žánry. Rozmanitost zastoupených žánrů je jev nesporně žádoucí; u angličtiny je jí díky velkému počtu různě zaměřených korpusů dosaženo; u korpusu jazyka méně žákovskými korpusy vybaveného je účelné dbát na to, aby v něm byly zaznamenány projevy žánrově i tematicky různorodé a aby se témata a žánr, resp. dominující styl, uváděly v lingvistické anotaci textů.

2.3.2. Parametry spojené s podmínkami vzniku a sběru textu

Způsob zadání a podmínky tvorby textů nejsou zaznamenávány důsledně u všech žákovských korpusů a ne u všech se zaznamenávají stejně podrobně, třebaže jsou důležité pro posouzení míry „autenticity" projevů, přesněji jejich volnosti, neřízenosti zřetelem k vyžadované formě vyjádření, a ovlivňují variabilitu užívání jazyka nerodilými mluvčími velmi podstatně. Jde o potenciálně velký soubor různých situ-

prvního jazyka žáků. Upozorňuje na to např. S. Grangerová (Granger, 2009, s. 17): *One must admit, however, that this facility is still seldom used and LC researchers (myself included) have had a tendency to base their analysis on the whole corpus or on subcorpora distinguished only on the basis of the learners' mother tongue. In fact, a properly coded learner corpus makes it possible for researchers to study the effect of a much wider range of variables.*

ačních parametrů, které můžeme rozdělit na ty, které se vztahují k celé situaci vzniku a sběru textu, a na faktory specifické, spjaté pouze s některou jeho fází.

(a) Pro charakteristiku situace jako celku je podstatné **určení textu,** resp. širší **kontext** jeho vzniku a sběru. Pokud je např. projev součástí jazykové zkoušky, můžeme předpokládat, že pozornost pisatele či mluvčího bude zaměřena na jeho formálně jazykovou stránku, zatímco obsah a funkce ustupují do pozadí. V takovém textu tedy můžeme počítat se strategiemi typu vyhýbání se apod.

Podobně funguje i osobnost sběrače, resp. vztah mluvčího/pisatele k němu. Tento faktor je podstatný především (ale ne výlučně) u projevů mluvených, k jeho vlivu na (ne)formálnost projevu, resp. péči věnovanou jeho jazykové stránce, srov. např. klasické práce W. Labova. Zejména u mluvených projevů je podstatný rovněž faktor připravenosti/nepřipravenosti, dialogičnost/monologičnost, počet mluvčích, prostředí (známé, neznámé, formální, neformální) a situační charakteristiky další.

(b) Do druhé skupiny lze zařadit především **zadání**. V závislosti na jeho přesnosti a striktnosti, resp. detailnosti může být žákova volnost, pokud jde o volbu jazykových prostředků, výrazně omezena, ev. některé vlastnosti jeho projevu mohou být přímo předurčeny. Je proto účelné zaznamenat, zda si žák mohl téma a/nebo žánr, popř. dominantní styl svého projevu určit sám, zda si je mohl zvolit z několika nabídnutých možností, nebo zda mu byly přímo zadány; zda mu byl stanoven minimální či maximální rozsah textu, a pokud ano, jaký, u mluveného projevu jeho délka apod.

Zadání ale působí i nepřímo – pokud je např. u písemného projevu zadán přesný čas, který má žák pro jeho tvorbu k dispozici, není tím přímo předurčena např. délka textu, ale může to jako psychologický faktor (časový stres) ovlivnit jeho kvalitu, omezit některé aktivity spojené s tvorbou textu, zvláště s jeho invenční a strukturační fází, apod.

Jinou povahu mají **přípravné aktivity** před vlastní tvorbou nebo **aktivity v jejím průběhu**. Máme na mysli aktivity spojené s působením nějakého vnějšího činitele, ať už je takovým činitelem učitel, sběrač, jiná osoba nebo třeba text, hudba apod. Žáci si např. mohou před psaním či nahrávaným ústním projevem přečíst vzorový text příslušného žánru nebo na příslušné téma, popř. si ho vyslechnout a podle jeho vzoru napsat/pronést text vlastní; mohou před vlastním psaním/ústním projevem diskutovat ve skupině o nějaké otázce a následně o ní napsat úvahu, v níž mají, mohou, ev. nesmějí použít některé obraty z předchozí diskuse; mohou postupovat při psaní a mluvení podle obrázkové osnovy apod.

Podobně mohou proces tvorby textu ovlivnit různé podmínky jeho průběhu: to, zda žáci mají příležitost radit se s jinými nerodilými mluvčími, s rodilými mluvčími, s učitelem, sběračem apod.; zda mají možnost při psaní používat pomůcky, např. slovník nebo mluvnici, zda mají k dispozici po dobu psaní/mluvení vzorový text; zda mají případně seznam výrazů a obratů, které mají, mohou, popř. nesmějí v textu použít apod. Vliv některých z těchto faktorů je jen potenciální (zpravidla nelze zaznamenávat, zda žák skutečně využil slovník a v kterých případech), jiné se naopak do podoby textu promítají velmi silně (seznamy slov, zadaná osnova textu apod.).

Repertoár dat spjatých se zadáním, přípravou a situací vzniku a sběru textu, která ovlivňují variabilitu jazyka žáků, a zasloužila by si tedy být v žákovských korpusech zaznamenána, je velmi bohatý. Jednotlivé korpusy si jich bohužel všímají pouze výběrově.

2.3.3. Parametry spojené s osobou autora

Parametry spojené s mluvčím se při výzkumech osvojování jazyka sledují tradičně. Na rozdíl od parametrů spojených s textem a situací jeho vzniku a sběru jsou zdánlivě stálejší. V korpusech se pravidelně zaznamenává věk mluvčího a jeho pohlaví; v žákovských korpusech je jedním z nejdůležitějších parametrů první jazyk, méně často se registrují rovněž další jazyky, které mluvčí ovládá, příp. (výjimečně) i úroveň jejich ovládání.

První jazyk není jen pravidelně zaznamenáván; je to také parametr, na němž se řada korpusů zakládá. Značná část korpusů je z tohoto hlediska monolingvní, tj. jde o korpusy, které zaznamenávají projevy mluvčích jednoho prvního jazyka v jednom jazyku cílovém, tedy např. anglické písemné projevy čínských žáků, japonských žáků apod.

Souvisí to s tím, že se výzkumy osvojování jazyka dosud zaměřují převážně na zjištění případné interference, zatímco vliv jiných faktorů soustavněji sledován není (viz pozn. č. 28); svou roli hrají i možnosti využití takového korpusu ve výuce žáků příslušného prvního jazyka.

V menší míře se setkáváme s korpusy s větším počtem prvních jazyků; to jsou např. oba komerční korpusy CLC a LLC, často zmiňovaný lovaňský korpus ICLE, LONGDALE, ale také řada korpusů neanglických – u nich je zaměření na mluvčí s různými prvními jazyky nepochybnou výhodou, mj. i proto, že dovolují porovnávat mezijazyky mluvčích různých prvních jazyků. Celkem ojedinělé jsou korpusy budované opačně – založené na projevech mluvčích jednoho prvního jazyka v několika jazycích cílových (korpus MLC – *Multilingual Learner Corpus*, který zahrnuje projevy brazilských mluvčích s portugalštinou jako prvním jazykem ve čtyřech různých jazycích – angličtině, němčině, italštině a španělštině).

Další skupinu parametrů spojených s osobou mluvčího mohou představovat data o době a způsobu osvojování cílového jazyka, případně o jiných okolnostech, které mohly jeho osvojování jazyka ovlivnit. Sem patří např. údaj o tom, zda žák pobýval v zemi, kde se cílového jazyka užívá, a jak dlouho; zda si cílový jazyk osvojuje či osvojoval přirozenou cestou, nebo institucionální výukou a kde tato výuka probíhala; jak dlouho už se cílovým jazykem zabývá, jak intenzivně a s využitím kterých učebnic; jaké úrovně v cílovém jazyce dosáhl; zda má případně možnost komunikovat v cílovém jazyce v rodině (pokud někdo v rodině mluví cílovým jazykem) apod.

Ne všechny tyto údaje se v žákovských korpusech objevují. Pravidelně se pracuje pouze s určením **úrovně ovládání jazyka**; zpravidla se však setkáváme pouze s ne zcela přesným určením (začátečník, mírně pokročilý, pokročilý), popř. se k určení úrovně ovládání jazyka užívá opisů typu „odpovídající třetímu a čtvrtému ročníku univerzitního studia", „univerzitní studenti", „střední škola" apod., které spíše než úroveň ovládání jazyka popisují způsob sběru (materiál se sbíral od studentů příslušných ročníků a škol). Společný evropský referenční rámec dnes nabízí možnost opřít se o poměrně přesné a dostatečně jemné stanovení úrovně ovládání jazyka, v existujících korpusech se však jako měřítko zatím příliš neuplatňuje.

Úroveň ovládání cílového jazyka je vedle prvního jazyka žáka nejvýznamnější ze sledovaných parametrů této kategorie – dovoluje studovat vývoj osvojování jazyka, a to nejen v případě longitudinálních korpusů, ale i v případě korpusů pseudolongitudinálních. Pro sledování vývoje jazyka je podstatné, aby žákovský korpus (či soubor žákovských korpusů) zachycoval projevy žáků pokud možno všech úrovní. To je však případ jen velmi výjimečný.

Ne všechny existující korpusy tento údaj uvádějí; z těch, u nichž je uveden, se naprostá většina zaměřuje na žáky jedné nebo dvou úrovní, ponejvíce na pokročilé nebo středně pokročilé studenty, na studenty cílového jazyka na univerzitách, popř. (u angličtiny) na absolventy středních škol, uchazeče o studium na univerzitách apod. Např. zmiňovaný korpus ICLE sbírá materiál od studentů angličtiny jako cizího jazyka ve třetím nebo čtvrtém ročníku univerzitního studia. Úroveň začátečník až středně pokročilý či pokročilý je v existujících (ovšem málo spolehlivých) přehledech žákovských korpusů uvedena pouze u cca devíti korpusů menšího rozsahu.

Zaměření na pokročilejší studenty souvisí se dvěma skutečnostmi. Především s tím, že projevy začátečníků, psané i mluvené, mají jen velmi malý rozsah, je tedy mimořádně obtížné vytvořit z nich korpus srovnatelné velikosti. Druhým činitelem je mimořádná chybovost v projevech začátečníků, v nichž jsou na řadě míst obtížně identifikovatelná slova a slovní tvary; ještě větší komplikace působí silná chybovost textů při lingvistické anotaci korpusů a také vyhledávání v takových korpusech je obtížné.

Souhrnně lze říci, že žákovské korpusy se, pokud jde o počet a povahu zaznamenávaných metadat spojených s textem, situací jeho vzniku a sběru a s jeho autorem, od sebe navzájem značně liší (viz i oddíl 2.3). Obecně však platí, že jsou tato data podstatně bohatší a důsledněji zaznamenávaná než u obecných korpusů lingvistických a že dávají badateli možnost sledovat působení velkého množství vnitřních i vnějších proměnných na proces osvojování a užívání jazyka a jeho variabilitu.

2.4. Zaznamenávání a další zpracování jazykových dat

2.4.1. Přepisy

Žákovské korpusy vznikají často pro potřeby konkrétního výzkumu, nezřídka je však jejich vytváření zároveň motivováno záměrem vytvořit materiálovou bázi pro širší, popř. obecné badatelské využití. Pokud se počítá s takovým opakovaným užitím korpusu pro různě zaměřené výzkumy, měla by jazyková data být zachycena pokud možno tak, aby byla co nejméně zkreslena jejich původní podoba, resp. aby bylo zachováno maximum informací o užitých jazykových formách na všech jazykových rovinách. Ideální by v tomto ohledu přirozeně byla práce s primárními záznamy – videozáznamy, popř. nahrávkami mluvených projevů a skeny písemných prací.

To však většinou není možné, jednak vzhledem k potřebě všechny projevy důsledně anonymizovat, tj. odstranit pokud možno všechny prvky, které by mohly vést k identifikaci mluvčího či školy, která data poskytla, jednak také s ohledem na potřeby automatického vyhledávání a lingvistické i chybové anotace. Volí se proto přepisy, a to jak u projevů mluvených, tak psaných.

Nejznámější systém záznamu dětských projevů byl vyvinut v rámci CHILDES. Jednou ze složek tohoto systému je CHAT (*Codes for the Human Analysis of Transcripts of child speech*), který nabízí tvůrcům akvizičních korpusů bohatou sadu nástrojů pro písemný přepis dětských projevů od fonetických detailů po komentáře zaznamenávající širší kontext.

Jednotlivé korpusy zahrnuté do databáze CHILDES však nevyužívají všech možností, které jim nástroje CHAT poskytují. Většinou je zpracování jazykových dat (podobně jako jejich sběr) nastaveno tak, aby odpovídalo primárně potřebám konkrétního výzkumu, v souvislosti s nímž korpus vzniká a jemuž má v první řadě sloužit. To platí přirozeně i o korpusech mimo CHILDES. Na těchto konkrétních potřebách závisí konkrétní volba toho, které jevy budou s využitím této sady zaznamenávány, jak přesně a jak detailně.[29] To je hlavní důvod, proč se zatím u akvizičních

[29] Sokolov a Snowová (Sokolov, Snow, 1994, s. 18) konstatují: *Few of the corpora stored in CHILDES are appropriate for a study of phonological development, for example, or for careful analysis of conversational phenomena like interruptions, latching, overlaps, back-channels, or filled pauses; the original transcription was simply not detailed enough. Many transcriptions give too little phonetic detail for one to be sure if child is saying or he, distinguishing wanna from want to, or segmenting you in could you and would you. For many analyses these imprecisions do not matter, for some they matter enormously.* Na tuto skutečnost (existenci značných rozdílů v detailnosti záznamů v korpusech CHILDES v závislosti na badatelském zaměření jejich autorů) upozorňují autoři opakovaně. Např. na s. 6 uvádějí, že badatel, *„who is interested in classifying speech at the level of communicative intent may find gestures and eye gaze crucial but*

korpusů nevytvořil obecný standard přepisů projevů mluvených ani psaných a neočekává se, že k unifikaci přepisů v této oblasti dojde v blízké budoucnosti.

V tom lze snad rovněž spatřovat určitý rozdíl akvizičních korpusů od obecných korpusů lingvistických, kde se rozmanitost pravidel přepisů např. mluvených projevů nevnímá jako výhoda a kde lze určité snahy o jednotnější podobu záznamů pozorovat.

Specifičnost záznamů žákovských jazykových dat (zejména psaných) je také v tom, že se při jejich přepisech předloha neopravuje ani v těch případech, kdy se v ní objeví zjevné nahodilé defekty či přehlédnutí, a že je žádoucí zaznamenávat i stopy procesu tvorby textu, např. u rukopisného textu škrty žáka, dodatečné vpisky, změny slovosledu a jiné zásahy, u školních prací i opravy a zásahy učitele, ve všech případech přirozeně s příslušnou značkou, která dovolí každý takový zásah žáka či učitele jednoznačně identifikovat.

2.4.2. Anotace

Závěrečnou fází zpracování jazykových dat v žákovských korpusech je jejich lingvistické a chybové značkování. Jde o proces velmi náročný, především proto, že texty nerodilých mluvčích (zejména na nižších úrovních ovládání jazyka) jsou vysoce chybové; možnosti použití nástrojů automatické anotace jsou tedy omezené, většinou je nutno spoléhat na časově i finančně velmi náročné značkování manuální.

To platí i o specifickém typu značkování žákovských korpusů, značkování chybovém. Chybové značkování navazuje na starší tradici chybových analýz (jež dosáhly svého vrcholu v 60. a 70. letech minulého století), ale odstraňuje některé jejich slabiny, které vedly k jejímu odmítání. Zejména dochází ke změně taxonomie chyb, aby byla spolehlivější a exaktnější, a mění se podstatnou měrou také způsob využití chybové analýzy s oporou o korpusová data.

Chybové značkování podstatně zvyšuje využitelnost korpusu pro následné analýzy, přináší však také řadu problémů, ať obecných, zasahujících všechny jazyky (např. skutečnost, že chybu lze opravit a v souvislosti s tím i klasifikovat, popsat a hodnotit často několika různými způsoby, ale chybová anotace tuto mnohost zachytit plně nemůže nebo jen za cenu nadměrně složitého záznamu), technických (volba mezi lineárním a vícerovinným formátem) nebo problémů spojených se specifickými rysy jednotlivých jazyků – u češtiny hraje roli zejména bohatá morfologie a variabilní slovosled. Věnujeme jim (a řešení zvolenému pro CzeSL) proto samostatnou pozornost v následujících dvou kapitolách.

have little need for phonetic detail, while researchers interested in phonological development need a detailed phonetic transcript but may include less non-verbal behaviour in their transcripts than the investigator studying speech acts".

2.5. Parametry CzeSLu

První žákovský korpus češtiny jako cílového jazyka s pracovním názvem CzeSL vzniká, jak jsme uvedli výše, jako jeden z výstupů projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk*, registrační číslo CZ.1.07/2.2.00/07.0259, v rámci programu Vzdělávání pro konkurenceschopnost s finanční podporou Strukturálních fondů EU (Evropského sociálního fondu) a státního rozpočtu České republiky, ve spolupráci TU v Liberci jako příjemce podpory, UK v Praze a AUČCJ jako jeho partnerů a s podporou a pomocí řady dalších institucí, organizací a jednotlivců, v koncepčním rámci spojeném s budováním rozsáhlejšího komplexu akvizičních korpusů češtiny AKCES.

Jeho primární funkcí je funkce pedagogická: sloužit jako nástroj přípravy učitelů (zejména učitelů češtiny) na práci s žáky s češtinou jako druhým, resp. cizím jazykem. Tato funkce ovlivňuje jak výběr jazykových dat, tak jejich zpracování. Je budován jako korpus otevřený: v r. 2012 bude zveřejněna jeho první verze, obsahující pouze projevy psané a sloužící primárně funkci pedagogické; v dalším pokračování budou práce zaměřeny na projevy mluvené a budou spojeny s navazujícími badatelskými projekty.

2.5.1. Velikost CzeSLu a volba jazykových dat

2.5.1.1. Velikost

Svým objemem patří CzeSL (resp. jeho složka obsahující projevy nerodilých mluvčích) mezi neanglickými žákovskými korpusy, jejichž velikost je známa, k největším: je v tomto ohledu přibližně srovnatelný se známými korpusy francouzštiny, němčiny či španělštiny. V době svého dokončení v první verzi v r. 2012 bude mít celkový objem 2 miliony slov; z toho zhruba polovinu (1 milion slov) budou tvořit projevy nerodilých mluvčích, kteří si češtinu osvojují jako cizí nebo druhý jazyk (bez rozlišení), druhou polovinu srovnávací korpus projevů českých žáků a korpus projevů dětí a mládeže ze sociokulturně znevýhodněných komunit ROMi (podrobněji viz kapitola 6).

2.5.1.2. Jazyková data

CzeSL je budován na základě dat klinicky elicitovaných. Velkou většinu tvoří eseje elicitované v učebním kontextu; menší část představují data získaná z kvalifikačních prací – jejich zařazení je motivováno mj. pedagogickou funkcí korpusu.

Sbírány jsou jak projevy psané, tak mluvené; první verze CzeSLu, která bude dokončena a zveřejněna v r. 2012, obsahuje pouze projevy psané, mluvená složka se bude zpracovávat a postupně doplňovat v dalších letech.

Sběr jazykových dat byl veden zřetelem k prvnímu jazyku mluvčích a k jejich úrovni ovládání češtiny, přesněji řečeno snahou získat jazyková data v takové skladbě, aby byly v korpusu zastoupeny všechny úrovně ovládání jazyka, od začátečníků po pokročilé, a v repertoáru prvních jazyků aby byly zastoupeny tři jejich kategorie: jazyky slovanské, neslovanské indoevropské jazyky a jazyky neindoevropské, typologicky a genealogicky češtině vzdálené.

Jednotlivé úrovně ovládání jazyka a skupiny prvních jazyků nejsou v CzeSLu zastoupeny rovnoměrně, korpus tedy není z těchto hledisek vyvážený; lze ho však pro studium variability užívání jazyka nerodilými mluvčími ve vazbě na tyto parametry využívat.

2.5.2. Zaznamenávaná metadata

Výběr zaznamenávaných parametrů byl veden dvěma ohledy: k pedagogické funkci korpusu a k jeho potenciálně co nejširšímu badatelskému využití. Repertoár metadat je tedy pokud možno co nejbohatší, a to jak metadat relevantních z hlediska lingvistického, psycholingvistického a sociolingvistického, tak i ryze didaktického. V tomto ohledu je CzeSL vybaven nadstandardně – žákovské korpusy takové čistě didaktické proměnné, jako je např. údaj o používané učebnici cílového jazyka nebo intenzitě jeho studia, obvykle nezaznamenávají.

Parametry zaznamenávané u všech tří složek CzeSLu, tedy u projevů češtiny nerodilých mluvčích, srovnávacího korpusu mluvčích českých a ROMi, se v některých ohledech liší. U českých rodilých mluvčích včetně mluvčích ze sociokulturně znevýhodněných komunit nejsou přirozeně zaznamenávány údaje o studiu češtiny jako druhého/cizího jazyka, o prvním jazyce či dalších jazycích, přistupují však podrobnější údaje relevantní pro hodnocení vlivu sociálního prostředí, tedy o regionu, z něhož mluvčí pochází, o typu školy, kterou navštěvuje, o velikosti a povaze sídla, kde žije (zda se např. jedná o sociálně vyloučenou lokalitu) apod. Dále viz oddíl 6.2.3 v kapitole 6.

Údaje relevantní pro možnost porovnávání projevu nerodilých mluvčích a mluvčích rodilých (vázaných k textu, k podmínkám jeho vzniku a sběru a k mluvčímu) jsou přirozeně zaznamenávány důsledně a jednotně u obou skupin. Jednotlivé zaznamenávané parametry[30] uvádíme přehledně v následujících tabulkách:

Parametry spojené s textem

Médium	mluvený × rukopisný × psaný na PC
Převažující slohový postup	informační, popisný, vyprávěcí, úvahově-argumentační
Téma	označení tématu
Typ tématu	obecné (rodina, volný čas, škola) × speciální (odborné)

[30] Při vkládání do korpusu mohou být tyto parametry i jejich repertoár modifikovány.

Parametry spojené se situací vzniku a sběru

Je zadán slohový postup?	zadaný × volný
Je zadáno téma	téma určené × výběr z několika zadaných × volné téma × jiný typ zadání
Je zadán rozsah?	ANO/NE; pokud ano, jaký
Je zadán časový limit?	ANO/NE; pokud ano, jaký
Přípravná aktivita (zvl. u psaných projevů)	žádná × obrázek × cvičení × slovní zásoba × diskuse × zadaná osnova × jiná aktivita; pokud jiná, jaká
Je projev součástí zkoušky?	ANO/NE; pokud ano, jaký typ zkoušky
Je povoleno užití slovníku nebo jiné pomůcky (u psaných projevů)?	ANO/NE
Prostředí pořízení materiálu	školní × mimoškolní × soukromé

Situační parametry zaznamenávané pouze u mluvených projevů

Forma projevu	monolog × dialog
Počet mluvčích	jeden × dva × více (kolik)
Formálnost situace	formální × neformální
Míra připravenosti	připravený × polopřipravený × nepřipravený
Sběrač – první jazyk	čeština × stejný jako žák × jiný (jaký)
Sběrač – věk	přesný údaj
Sběrač – vztah k žákovi	volné určení

Situační parametry zaznamenávané pouze u žáků ze sociokulturně znevýhodněných komunit a ze srovnávací skupiny rodilých mluvčích

Místo sběru	Praha × jiné; pokud jiné, výběr kraje (z nabídky)

Parametry spojené s osobností žáka

Pohlaví	muž × žena
Věk	přesný údaj
První jazyk	přesný údaj; modifikováno u sociokulturně znevýhodněných mluvčích, viz níže
Znalost dalších jazyků	přesný údaj; modifikováno u sociokulturně znevýhodněných mluvčích, viz níže

Parametry žáka zaznamenávané pouze u nerodilých mluvčích

Skupina prvních jazyků	slovanský × jiný indoevropský × neindoevropský
Studium češtiny – místo	ZŠ nebo SŠ v ČR × VŠ v ČR × komerční JŠ v ČR × jiný kurz v ČR × kurz mimo ČR × samostudium × individuální lekce × jiné
Studium češtiny – doba	týdny, měsíce, roky, semestry
Studium češtiny – intenzita	do 3 hod., do 15 hod., nad 15 hod. týdně
Studium češtiny – učebnice	dána nabídka, lze doplňovat jiné

Znalost češtiny podle SERR	přesný údaj
Bilingvnost	ANO/NE
Délka pobytu v ČR	kratší než 1 rok × 1–2 roky × delší než 2 roky
Umí někdo v rodině česky?	otec × matka × oba rodiče × bratr/sestra × partner/ka × nikdo

Parametry žáka zaznamenávané pouze u žáků ze sociokulturně znevýhodněných komunit a ze srovnávací skupiny rodilých mluvčích

Navštěvovaná škola	ZŠ × SŠ × SOU × SOUM × G × ZSS × ZSP
Třída/ročník	přesné určení
Mluví žák romsky?	ANO/NE; pokud ano, kdy a kde
Mluví někdo blízký romsky?	rodiče × prarodiče × jiní příbuzní × kamarádi × nikdo v okolí
První jazyk	čeština × slovenština × romština × jiný jazyk
Doma mluví	česky × slovensky × romsky × česky a romsky × jiným jazykem
Bydliště – velikost sídla	méně než 5000, nad 5000, nad 10000, nad 50000, nad 100000 obyvatel
Bydliště – region	Čechy × Morava × Slezsko
Bydliště – nářeční oblast	středočeská × jihozápadočeská × severovýchodočeská × česko-moravská × středomoravská × východomoravská × slezská × pohraničí české × pohraničí moravské
Bydliště – sociálně vyloučená lokalita?	ANO/NE
Poznámky sběrače/učitele k žákovi	volné údaje

2.5.3. Způsob zaznamenávání a dalšího zpracování jazykových dat

2.5.3.1. Přepisy

Pravidla pro přepis psaných projevů použitá v CzeSLu respektují základní požadavek zachovat v přepisu maximum informací o podkladovém rukopise, tj. jsou zaznamenávány (a kódovány) všechny zásahy, které pisatel nebo jeho kontrolor v textu provedl, a jsou zaznamenána rovněž místa nečitelná či víceznačná. Taková je základní podoba textu, jak je uložena v databance.

Ukázka přepsaného textu uloženého v databance:

> *{Na obrázku je hezký pokoj.}<dt> Na obrázku je velký a útulný pokoj. Nalevo visí ob krasný obraz a dole {v rohu}<in> stojí postel. Nahoře visí lampa. Vzadu je okno. {Vstředu|V středu} (uprostřed) stojí židle. Napravo je skříň a stůl. Na skříňí je radio a televize. Vepředu stojí taky stůl a židle. Tam taky tří človeka. To je manžel který čté knihu, mo jeho manželka která se divá o na televizi a babička.*

Pro potřeby korpusu je text od těchto komentářových složek očištěn, tj. do CzeSLu vchází čistý text, v němž nejsou zásahy učitele nebo jiné osoby kromě autora uvedeny a který odpovídá verzi z hlediska pisatele poslední (tedy se zanesením všech oprav) a je jednoznačný – nejasná místa musí zpracovatel posoudit a zjednoznačnit (tedy rozhodnout, zda určité písmeno je *o*, nebo *a*). (Detailněji o přepisech v kapitole 5.)

Ukázka čistého textu:

> *Na obrázku je velký a útulný pokoj. Nalevo visí ob krasný obraz a dole v rohu stojí postel. Nahoře visí lampa. Vzadu je okno. V středu (uprostřed) stojí židle. Napravo je skříň a stůl. Na skříňí je radio a televize. Vepředu stojí taky stůl a židle. Tam taky tří človeka. To je manžel který čté knihu, mo jeho manželka která se divá o na televizi a babička.*

Pravidla pro přepis mluvených projevů se zatím ověřují. V obecných lingvistických korpusech češtiny řady ORAL se uplatňuje přepis ortografický s mírnými odchylkami, které dovolují zachytit např. některé specifické výslovnostní jevy, zvl. obecněčeské. Dosavadní mluvené korpusy AKCES využívají stejný model, ale s výraznějšími modifikacemi (pravidla přepisu viz http://ucnk.ff.cuni.cz/schola-prepis.php)

V připravovaných mluvených korpusech nerodilých mluvčích je ale žádoucí zaznamenávat výslovnostní jevy ve větším rozsahu a důsledněji: nabízí se tedy možnost využít modelu víceúrovňového přepisu, který vyvinuli pracovníci Fonetického ústavu FF UK (J. Veroňková). Definitivní podobu však pravidla přepisů pro mluvené korpusy typu CzeSL zatím nemají.

2.5.3.2. Anotace

Korpus CzeSL bude vybaven lingvistickou a chybovou anotací, a to dvojího druhu. Jako celek projde korpus chybovou anotací s využitím modifikovaných nástrojů anotace automatické, které dovolí označkovat odchylky od spisovné jazykové normy na rovině ortografické a tvarové.

Zatím menší část korpusu CzeSL projde specifickou chybovou anotací manuální na základě modelu vícerovinné emendace textů, který byl vyvinut s ohledem na anotační problémy vyplývající jednak z typologické povahy češtiny (bohatá morfologie, specifický slovosled apod.), jednak ze zařazení výrazně chybových začátečnických textů do korpusu. Tento model dovoluje emendovat a anotovat chyby podstatně detailněji a všímat si i chyb syntaktických a slovosledných. Vyžaduje ovšem velké množství ruční práce, proto se zatím uplatní pouze u menší části CzeSLu. Podrobněji o anotacích, zvláště chybových, v následujících kapitolách.

3. Chybové taxonomie a možnosti chybové anotace v žákovských korpusech[31]

Barbora Štindlová

Počátek výstavby žákovských korpusů v devadesátých letech minulého století vedl k renesanci dvou hlavních metodologických přístupů k otázkám cizojazyčné akvizice, a zároveň také k teorii cizojazyčné výuky, které dominovaly výzkumům ve druhé polovině dvacátého století, tj. kontrastivní a chybové analýzy. Od přelomu tisíciletí jsou v souvislosti s tzv. korpusově založeným, příp. korpusově řízeným přístupem[32] v modifikované podobě znovu široce aplikovány jako kontrastivní analýza mezijazyka a počítačem podporovaná chybová analýza.[33]

Počítačem podporovaná chybová analýza se ve značné míře vyhýbá nedostatkům původní chybové analýzy.[34] Moderní žákovský korpus není zaměřen na specifický typ chyby a vlastně ani na chyby jako takové, jak tomu bylo u nedigitalizovaných korpusů shromažďovaných za účelem původní chybové analýzy. Naopak korpus jazyka nerodilých mluvčích reprezentuje žákovský jazyk komplexněji, tj. se všemi systémovými i nesystémovými prvky. Díky relativně velkému rozsahu žákovských korpusů, pokud srovnáváme s původními sbírkami žákovských projevů, a díky jejich elektronické podobě je možné aplikovat různorodé frekvenční analýzy s využitím moderních nástrojů pro statistické výzkumy (např. srovnáním relativní a absolutní

[31] Podrobněji viz Štindlová (2011).

[32] Standardně se vymezují dva přístupy k využití korpusu jako prostředku pro ověřování, exemplifikaci či budování lingvistické teorie. Pro korpusem řízený přístup slouží korpus jako empirická báze, ze které jsou extrahována data a detekovány lingvistické jevy bez předem stanovené hypotézy. Konvenuje s holistickým přístupem k jazyku. Srov. i Sinclair (1996) a Tognini-Bonelliová (Tognini-Bonelli, 2001, s. 86).
V přístupu na korpusu založeném slouží korpus jako lingvistická databanka, ze které jsou získávána relevantní data k verifikování postulované hypotézy, ke kvantifikaci jazykových jevů, jako ilustrativní příklady. Tj. korpus má funkci podpůrného výzkumného materiálu. Srov. i Tognini-Bonelliová (Tognini-Bonelli, 2001, s. 66); příp. tzv. *knowledge-based metodology* u Atkinsová, Clear a Ostler (Atkins, Clear, Ostler, 1991).

[33] Viz Štindlová (2011, s. 44n.).

[34] O chybové analýze blíže např. Corder (1967), Dulayová et al. (Dulay et al. 1982), James (1998), Ellis (1994), Ellis, Barkhuizen (2009), v českém prostředí Štindlová (2011, s. 21n.) aj.

frekvence určitého typu chyb lze vyhodnotit i problém vyhýbání). Značně rozšířené jsou i možnosti při aplikaci chybové taxonomie, která je do korpusu vnášena chybovou anotací. Některé chybové taxonomie jsou budovány hierarchicky a kombinují různé přístupy ke kategorizaci chyb, některé anotační systémy dokonce umožňují aplikaci několika odlišných taxonomií zároveň. Přepokládá se také, že elektronická podoba korpusů jazyka nerodilých mluvčích a aplikace softwarových nástrojů pro jejich značkování by do budoucna mohly usnadnit srovnání jednotlivých typů korpusů, užitých anotačních nástrojů i chybových taxonomií.

3.1. Anotace jazykových korpusů

Geoffrey Leech (1997, s. 1) poznamenal: „Anotace jazykového korpusu je obecně vnímána jako podstatný příspěvek k výhodám, které korpus přináší, protože obohacuje korpus jako pramen lingvistických informací pro budoucí výzkum." Pro efektivní využití korpusů je lingvistická anotace zcela zásadní a anotované korpusy přirozeného jazyka jsou velmi cenným výzkumným nástrojem. Umožňují verifikovat postulované hypotézy a generalizace a na jejich základě lze také formulovat hypotézy nové. Obdobně lze uvažovat o roli anotace v korpusech žákovského jazyka. Žákovský jazyk je standardně v kontextu akvizice druhého/cizího jazyka nahlížen jako jazykový systém sám o sobě, tzv. mezijazyk, a měl by být analyzován včetně nekorektních struktur.

Data nerodilých mluvčích se v žákovských korpusech mohou anotovat dvěma na sobě nezávislými způsoby. Za prvé se jedná o možnost lingvistického značkování (tím máme na mysli značkování slovních druhů, morfologickou, příp. syntaktickou anotaci, lemmatizaci, identifikaci vztahů textové koreference atd.). Pro tento typ značkování je zásadní otázkou volba vhodného schématu lingvistické anotace aplikovatelného na žákovský jazyk, jeho vztah ke konkrétnímu metodologickému rámci a koncepční ukotvenost v návaznosti na popis cílového jazyka. Ačkoli jednou z hlavních otázek výzkumů nabývání cizího jazyka je struktura jazykových pravidelností na jednotlivých úrovních procesu nabývání cílového jazyka, a to bez ohledu na to, zda jsou v kontextu národního jazyka správné, či nikoli, není problematice (automatické) lingvistické anotace prozatím věnována dostatečná pozornost (srov. Meurers, 2009). Nejčastěji se v žákovských korpusech uplatňuje slovnědruhové značkování (např. v korpusech ASU, JEFLL, ICNALE, ICLE aj.), obvykle aplikované na menší část korpusu. Pro tento typ anotace jsou využívány softwarové nástroje původně vyvinuté pro potřeby analýzy národního jazyka, jako např. TOSCA-ICLE, CLAW, Brill tagger, Oslo–Bergen tagger.[35]

[35] Podrobnější zhodnocení úspěšnosti aplikace těchto nástrojů na chybové texty viz např. van Rooy a Schäfer (2003).

Jiným typem značkování žákovských korpusů je tzv. chybová anotace.[36] Chybové anotování korpusu žákovského jazyka znamená přiřazení odpovídající značky (neboli tagu) konkrétní chybě vyskytující se v žákovském projevu, dosud, pokud je známo, jen manuální. V současné době se některá výzkumná pracoviště soustředí na vývoj automatizace chybové anotace, prozatím však jejich výsledky nebyly evaluovány, srov. Izumi et al. (2005), Reuer a Kühnberger (2005), Meurers (2009). Chybové značky jsou součástí chybové taxonomie. Vybudování validní chybové taxonomie a dostupnost srozumitelného seznamu specifických typů chyb je základem využitelnosti žákovského korpusu pro výzkumy nabývání cizího jazyka i pro specificky zaměřené otázky. Navzdory skutečnosti, že chybovou anotaci je třeba z velké části provádět manuálně, a je tudíž značně časově náročná, počet žákovských korpusů vybavených touto anotací v současné době neustále roste. Úroveň, rozsah a koncept chybové anotace se však ve značkovaných žákovských korpusech značně odlišují.

3.2. Anotace v žákovských korpusech[37]

Žákovské korpusy stejně jako jiné jazykové korpusy se vzájemně odlišují množstvím lingvistické informace, jež je přidávána k původnímu textu. Většina dostupných žákovských korpusů obsahuje řadu externích informací (tzv. metadat), které charakterizují text a autora textu (např. první jazyk, věk, dobu studia cílového jazyka, typ textu, elicitaci apod. Srov. zde kapitola 2, násl., příp. i Štindlová, 2011, s. 50násl.). Odlišná situace je při mapování implementovaných anotací jazykových. Z porovnání vyplývá, že v žákovských korpusech je uplatňována chybová anotace ve větší míře než anotace lingvistická. Jako chybově anotované se vymezuje přibližně 45 % světových žákovských korpusů. Z nich se ovšem jen 7 % pokouší o komplexně pojatou chybovou anotaci se systémovou taxonomií chyb. Zbývajících 38 % žákovských korpusů uplatňuje chybové značkování vázané na explicitně vymezenou výzkumnou hypotézu. Např. žákovský korpus ISLE značkuje pouze nedostatky výslovnostní, žákovský korpus CEDEL2 se zaměřuje zachycení problémů syntaktických, korpus CIC se zaměřuje na značkování lexikálních problémů apod.[38]

Rozsah chybového značkování podstatně ovlivňují vnější faktory, především náročnost a nákladnost manuálního značkování. Např. v žákovském korpusu ICLE je z celkového počtu tří milionů slov anotována přibližně jedna čtvrtina, v žákovském korpusu HKUST s rozsahem dvacet pět milionů slov je chybově anotováno při-

[36] Viz např. Díaz-Negrillo a Fernández-Domínguez (2006) nebo Štindlová (2011).

[37] Souhrnné statistické údaje ohledně uplatňování lingvistické a chybové anotace ve světových žákovských korpusech vychází z analýzy padesáti sedmi vybraných žákovských korpusů, uvedené v Štindlová (2011).

[38] Blíže Štindlová (2011, s. 74).

bližně dvě stě tisíc z nich. Pokud uvedená zjištění shrneme, můžeme konstatovat, že ačkoli se o chybové anotaci žákovských korpusů často hovoří jako o běžně aplikované,[39] v téměř polovině existujících žákovských korpusů ji nenalezneme. K tomu je však třeba poznamenat, že některé korpusy se chybové anotaci vyhýbají záměrně, protože ji pokládají za interpretační model, který ovlivňuje přístup k datům – srov. Fitzpatricková a Seegmiller (Fitzpatrick, Seegmiller, 2004).

3.2.1. Anotační modely

3.2.1.1. Lineární anotační model

Současné anotované žákovské korpusy používají v zásadě dva typy anotačních schémat. Majoritně se v žákovských korpusech stejně jako ve většině značkovaných referenčních korpusů[40] psaných jazykových projevů uplatňuje tzv. lineární anotační model. Velkou výhodou lineárního anotačního schématu je dostupnost mnoha kvalitních kompatibilních nástrojů pro vyhledávání. Materiál je v tomto přístupu značkován na jedné rovině chybovými kódy, které mohou být kombinovány a vzájemně i částečně rigidně hierarchicky uspořádány.[41] Zároveň jsou však možnosti anotace výrazně omezeny. Jednorovinná anotace se komplikovaně vyrovnává s označováním chyb na oddělených řetězcích, tj. nesnadné je zaznamenávání slovosledných chyb a chyb zasahujících sekvence slov. Problematická je anotace kolidujících chybových úseků, obtížně lze také zaznamenat odlišné typy chyb vyskytující se na jednom chybovém úseku, resp. klasifikovat chyby zasahující různé (obvykle lingvisticky vymezené) domény. Tento způsob chybového značkování nabízí jen limitované možnosti pro rekonstrukci a interpretaci žákovských chyb, tj. při anotaci nelze zohlednit alternativní hypotézy. Lineární model také často sjednocuje v rámci jednoho chybového tagu dva odlišné kroky chybové analýzy – deskripci a explanaci.[42]

[39] Srov. Rastelli (2009), Meurers (2005), Díaz-Negrillo a Fernández-Domínguez (2006) atd.

[40] Referenční korpus chápeme v souladu s Čermákem a Blatnou (1995, s. 52) jako korpus jádrový, „relativně stálý reprezentativní soubor pro získání běžné, základní a nespecializované informace různého druhu; jeho rozsah se pohybuje v anglickém prostředí od 20 miliónů výskytů".

[41] Srov. např. Weinbergerová (Weinberger (2002). V chybově anotovaném žákovském korpusu němčiny jde o čtyřrovinnou lineární klasifikaci, kdy jsou chybové kódy skládány do jednoho tagu (rovina 1 – lingvistické kategorie (lexikální, morfologická…), rovina 2 – slovnědruhové kategorie, rovina 3 – lingvistické subkategorie (např. ortografie, interpunkce…), rovina 4 – povrchové modifikace a další specifikace (např. redundantní výraz, výběr…).

[42] Nedostatky lineárního zachycování chyb v textech nerodilých mluvčích dále shrnují Lüdeling et al. (2005), Zeldes et al. (2009) a Fitzpatrick a Seegmiller (2003).

Nejjednodušší strukturou jednorovinné anotace je tzv. tabulární model, kdy je materiál značkován na úrovni tokenů,[43] přesněji řečeno chybová značka je spojena vždy s jednotkou psaného textu určitého rozsahu, příp. s časovými segmenty verbálního nebo vizuálního signálu.[44] V příkladu (1) je chybový tag (značka) umístěn do původního textu před chybový výraz. V daném konceptu nelze registrovat rozsah chyby. To znamená, že není odlišeno značkování chyb na jednotlivých tokenech od značkování chyby na sekvenci slov. Rekonstrukční hypotéza je v tomto konkrétním příkladě implicitní.

(1) *<LxPhCh>Es gibt eine veränderte Gesellschaft und …*[45]
Lx – lexikální doména, Ph – exponent chyby (fráze), Ch – chybný výběr (specifikace chyby)
Hypotéza*: die Gesellschaft hat sich verändert*

Typickou strukturou lineární anotace je řízené stromové uspořádání, které umožňuje zachycení rozsahu chyby a v jisté míře i anotaci disparátních částí. Tento koncept se uplatňuje u většiny anotovaných žákovských korpusů. V příkladu (2) z žákovského korpusu NICT JLE je rekonstrukční hypotéza umístěna před chybový výraz, který je zároveň oboustranně vymezen příslušným chybovým tagem. Rekonstrukční hypotéza je v tomto případě explicitní.

(2) *I belong to two baseball <n_num crr="teams">team</n_num>*[46]
n – substantivum, num – číslo, crr – korekce

Příklady (3a,b) a (4a,b) níže ilustrují problémy při aplikaci lineárního, jednorovinného modelu. Za prvé jde o zachycení konfliktu hierarchií v případě, kdy je určení chyby dvojznačné nebo nejednoznačné. Chyba v předložkové frázi (*o kamarádka*) začíná v chybném řetězci argumentové struktury (*myslela o*) a končí mimo tuto strukturu.

(3) a. *Celý den **myslela o** kamarádka.*[47]
b. *Celý den myslela **o kamarádka**.*

[43] V oblasti korpusové lingvistiky se odlišují pojmy: **token** jako výskyt slovního tvaru v korpusu, **typ** jako slovní tvar jako takový a **lemma** jako základní tvar pro skupinu tvarů (např. infinitiv pro celé slovesné paradigma).

[44] Srov. s Carletta et al. (2002, s. 3). V rámci žákovských korpusů je variací tohoto modelu značkování korpusu C-LEG (Weinberger, 2002).

[45] Weinberger (2002, s. 25), podtrženo BŠ.

[46] Izumi et al. (2005, s. 75), podtrženo BŠ.

[47] České příklady vyexcerpovala autorka z databanky textů nerodilých mluvčích, která je shromažďována pro korpus CzeSL.

Každá chybová analýza je založena na cílové hypotéze, ať již implicitně, nebo explicitně vyjádřené. Hlavním deficitem jednorovinného anotačního modelu je jeho neschopnost respektovat případnou variaci cílových hypotéz. V následujícím příkladu je možná dvojí hypotéza: buď je chybně užito sloveso (*zapamatovat* místo *vzpomenout*), nebo je chybně užita prepozice (*na*).

(4) a. *Nemůžu si **zapamatovat** na mnoho slov.*
b. *Nemůžu si zapamatovat **na** mnoho slov.*

Pro lineární anotační model, který pracuje s jedinou rovinou pro cílovou hypotézu, je nutné upřednostnit pouze jednu z alternujících hypotéz. Problém mezianotátorské shody s ohledem na cílovou hypotézu nebyl prozatím dostatečně analyzován.[48] Zachycení různých rekonstrukčních variant umožňuje víceúrovňová architektura.

3.2.1.2. Víceúrovňová distanční anotace

Odlišným anotačním schématem je tzv. víceúrovňová distanční anotace, kterou používá ve světovém kontextu žákovských korpusů pouze německý korpus FALKO (a je jím částečně inspirován i český korpus CzeSL). Tento anotační systém navazuje na modely, které byly v posledním desetiletí vyvinuty pro mluvené a multimodální korpusy.[49] Multidimezionální distanční anotace aplikovaná na žákovský korpus se metodologicky opírá o předpoklad Lüdelingové et al. (Lüdeling et al., 2005), že „při chybovém tagování není možné neinterpretovat", a z toho důvodu je třeba umožnit prezentaci několika odlišných rekonstrukčních hypotéz. Toto pojetí se vyrovnává s kritikou počítačem podporované chybové anotace, že je zásadně závislá na anotátorské interpretaci. Vlastní chybová anotace je ve víceúrovňovém distančním modelu umístěna mimo původní text a je možné ji rozšiřovat podle potřeby cílové hypotézy, tj. má pohyblivý počet anotačních rovin. To umožňuje alternativní interpretaci a rekonstrukci chybového textu. Zároveň lze v rámci tohoto schématu kódovat i překrývající se chybové řetězce, a to na různých anotačních rovinách (viz tabulka 1).

[48] Tematicky zaměřená sonda viz Lüdeling (2008).

[49] Srov. např. Bird a Liberman (1999), kteří konstruovali několikaúrovňový anotační model pro část korpusu BU Radio News. Multidimenzionální anotace mluvených i psaných korpusů je v současnosti plně etablovaným nástrojem lingvistické analýzy. Srov. např. formáty NITE (Carletta et al., 2003), EXMARaLDA (Schmidt, 2004), PAULA (Dipper, 2005), SGF (Stührenberg et al., 2006) atd.

Tabulka 1: Příklad anotace překrývajících se řetězců[50]

Word	aus	denen	sich	insgesamt	die	Bedeutung	und	den	Sinn	des	ganzen	Textes	erschließen	läßt
target					die Bedeutung und der Sinn des ganzen Textes erschließen lassen									
finiteness														x
agreement					X									
binding								x						

Tento konkrétní model víceúrovňové distanční anotace prozatím uspokojivě nevyřešil problém značkování disparátních jednotek, resp. korespondence mezi elementy na různých úrovních je zachycena pouze implicitně a může být v anotačním procesu ztracena. Možným nedostatkem tohoto schématu je i jeho anotátorská náročnost a jistým způsobem i jeho difúzní charakter. Tím máme na mysli, že pokud by měly být chyby značkovány nespojitě, bez předem definované taxonomie, pouze na základě konkrétního výzkumného záměru či individuálního přístupu uživatele, a navíc na omezeném rozsahu materiálu (na vzorku), je limitována možnost ověření výstupů a dá se reálně přepokládat, že příp. výzkumy zaměřené na analýzu stejného jevu povedou na základě odlišného přístupu ke značkování chyb k odlišným výsledkům. Tato otázka by však zasluhovala podrobnější prozkoumání.

3.2.2. Chybová taxonomie

Vedle analýz mezijazyka, kvantitativního srovnávání jazyka rodilých a nerodilých mluvčích a využití korpusových studií pro metodologii výuky cizího jazyka či tvorbu didaktických materiálů jsou jedním z hlavních témat prací vycházejících z žákovských korpusů výzkumy zabývající se chybovou analýzou. Tuto oblast zájmu lze ještě dále rozdělit na analýzy směřující k vývoji automatického značkování chyb, které je, jak jsme již zmínili, prozatím odbornou veřejností neevaluováno, a na počítačem podporovanou chybovou analýzu žákovských projevů.

Klasickou chybovou analýzu lze standardně charakterizovat v pěti krocích: sběr dat, resp. výběr vzorku pro analýzu, identifikace chyby, popis chyby, její vysvětlení a zhodnocení.[51] Explanace a evaluace chyb jsou zásadní pro výzkumy nabývání cizího jazyka, příp. pro potřeby výuky cizích jazyků, v případě budování a chybového značkování žákovských korpusů jsou však reflektovány minimálně. Klasifikace a značkování chyb v žákovských korpusech by měly být maximálně informativní a popisné, tj. srozumitelné, konzistentní, formální a dostatečně obecné, protože účelem budování jazykového, tedy i žákovského korpusu je umožnit badateli přístup

[50] Obrázek převzat z Lüdeling et al. (2005).

[51] Podrobněji o problematice chybové analýzy viz Štindlová (2011, s. 8n.).

k relevantnímu materiálu pro následný lingvistický výzkum, nikoli tento výzkum provádět.[52]

Existují dva základní přístupy k zachycování chyb v žákovských projevech. Za prvé jde o implicitní zachycení chyb, tj. rekonstrukci, kdy je v průběhu emendace[53] chyba v textu detekována a nahrazena korektní formou,[54] za druhé o explicitní chybovou klasifikaci, kdy jsou žákovské chyby identifikovány a následně tříděny a kategorizovány podle předem vymezené chybové typologie. Problémem v obou přístupech je otázka, jak docílit mezianotátorské shody v cílové hypotéze (tj. v rekonstrukci chybového textu, na jejímž základě dochází ke klasifikaci chyb).[55]

Výhodou rekonstrukčního přístupu je primárně absence klasifikačního schématu (Fitzpatrick, Seegmiller, 2004): anotátor se jej nemusí učit, tj. tento typ anotování je rychlejší, nedochází k chybnému zařazení chyby a nenastává problém s chybami (méně obvyklými, okrajovými), které nelze jednoduše do navržené typologie zařadit. Vlastní rekonstrukce textu bez kategorizace chyb však může být následně pro uživatele neprůhledná, protože nepopisuje chybu a neobjasňuje důvody pro volbu použité opravy. Zároveň také v případě, že rekonstrukční korpus není morfologicky značkován, neumožňuje přístup bez chybové typologie snadnou aplikaci kvantifikačních a statistických metod.

Chybová taxonomie, na jejímž základě dochází ke kategorizaci chyb, vždy určitým způsobem odráží teoretický koncept, v jehož rámci vznikla, a chybové kategorie, které zahrnuje, mohou reflektovat úzce zaměřený výzkumný záměr. Problémem takto postavené chybové typologie je pak její malá využitelnost pro analýzy s odlišnými badatelskými cíli. I přes dílčí nedostatky, které může klasifikační přístup k chybám vykazovat, je tento koncept při značkování žákovských korpusů významně preferován. Nabízí totiž široké možnosti statistických analýz.

3.2.2.1. Typologie chybových taxonomií

Chybová taxonomie, která vymezuje jednotlivé kategorie chyb v žákovském jazyce, je základním stavebním kamenem celého anotačního systému zaměřeného na značkování odchylek od standardu v projevech nerodilých mluvčích. Chybové kategorie jsou v systému prezentovány chybovými značkami, užívanými v anotačním procesu

[52] *„We have deliberately decided not to use distinctions such as 'errors' versus 'mistakes' or 'interlingual' versus 'intralingual' errors, which are difficult to assign and better left for a second stage in the analysis.*" (Granger, 2003a, s. 467)

[53] Termín emendace používáme ve shodě s praxí v korpusu CzeSL ve smyslu prosté opravy, resp. rekonstrukce textu dle standardů cílového jazyka.

[54] Viz Fitzpatrick, Seegmiller (2003).

[55] Standardně při anotování žákovských projevů značkují jeden text nezávisle vždy minimálně dva anotátoři. K problematice cílové hypotézy a mezianotátorské shody při značkování žákovského korpusu viz blíže Štindlová (2011, s. 121n.).

k zařazení konkrétních žákovských chyb, které se v žákovském korpusu vyskytují, do příslušných kategorií podle zvolené taxonomie.[56]

V chybově anotovaných žákovských projevech se v současnosti vedle implicitní anotace ve smyslu prosté emendace bez chybové klasifikace, tedy bez chybové taxonomie a bez značek, kterou používá pouze americký korpus MELD, uplatňují dva základní typy konstrukčních přístupů k zachycení a popisu chyb. Pro další popis je definujeme jako taxonomii parciální a taxonomii komplexní.[57]

Parciální taxonomie je zaměřená na specifické typy chyb podle deklarovaného výzkumného záměru, např. na vybrané morfémy (JEFLL), na lexikální chyby (CIC), na tzv. zásadní chyby (EARS) ap. Taková účelová a zacílená taxonomie je zároveň reakcí na neúspěšné pokusy o vybudování obecných chybových taxonomií ze sedmdesátých let. V případě parciálních taxonomií se často jedná o organicky vznikající chybovou kategorizaci, která není komplexní a systematická v lingvistickém slova smyslu, vychází primárně z anotovaného materiálu a účelu korpusu, kterým může být např. praktické využití ve výuce (např. korpus CLC).[58] Aplikace takového typu taxonomie se vyznačuje vyšší mezianotátorskou shodou, relativní jednoduchostí a rychlostí anotace.

Podrobná, komplexní taxonomie chyb je obvykle hierarchicky založená, ať už na základě lingvistických kategorií či typu povrchové realizace (např. ICLE, NICT JLE).[59] Vyžaduje podrobnou rozpracovanost anotačního manuálu a klade značné nároky na kvality anotátora. Je však významnou podporou pro analýzu nabývání i výuky cílového jazyka.[60]

Každá deskriptivní chybová taxonomie by měla reflektovat dvě svébytné oblasti: lingvistickou kategorizaci a formální popis chyby, resp. typ alternace zdrojového textu (tj. informaci o tom, že jev chybí, přebývá, je chybně umístěn, je chybně použit).[61] V anotovaných korpusech jazyka nerodilých mluvčích je formální klasifikace chyb často užívána jako komplementární k lingvisticky orientované kategorizaci. Obě hlediska jsou v tom případě strukturována buď bidimenzionálně, resp. propojeně, obvykle v podobě slovnědruhové značky a značky pro typ povrchové realizace

[56] Podrobněji k otázkám chybových taxonomií v souvislosti s tradiční chybovou analýzou viz Štindlová (2011, s. 27n.).

[57] Oba termíny zavádí do kontextu problematiky chybově značkovaných žákovských korpusů Štindlová.

[58] Srov. Nicholls, 2003.

[59] Tzv. taxonomie dle povrchové realizace klasifikuje chyby podle formálních deformací žákovských projevů. Na základě tohoto konstruktu jsou chyby standardně klasifikovány jako tzv. vynechání, nadbytečné užití, užití chybné formy a chybný slovosled.

[60] Tono (2003, s. 801): *„A generic error tagset, however, stills seems to be a very useful goal to work towards…*"

[61] Srov. James (1998, s. 104–113), Granger (2003a, s. 467), Tono (2003, s. 804), Díaz-Negrillo, Fernandéz-Domínguez (2006, s. 92).

(příklad (5) z korpusu CLC), nebo separátně, tj. tagy pro značkování chybějících/nadbytečných jevů, chyb ve slovosledu atd. nejsou spojovány s lingvistickou kategorizací (příklad (6) z korpusu ICLE).[62] Některé žákovské korpusy povrchový charakter chyby nereflektují (např. NICT JLE).

(5) *we arrived <#RT> to/at<#RT> our destination*
R – výraz je chybně použit, je třeba ho vyměnit, T – předložka

(6) *[…] big ruined walls stood(WM) 0 $rising$ towards the sky.*
WM – chybějící výraz

Lingvistické značkování chyb v korpusech nerodilých mluvčích se liší jak podrobností klasifikace, tj. od označení kategorií velmi široce pojatých (morfologie, lexikum, syntax) ke kategoriím pojatým specifickým způsobem (pomocná slovesa, pasivum, apod.), tak i v případném hierarchickém uspořádání založeném na kombinaci různých aspektů v náhledu na chybu. Značkovací systém může být pevně ukotven v analýze jazykových rovin, kdy popis chyby zahrnuje hierarchicky uspořádanou informaci (1) o doméně jako nejobecnější rovině, která určuje, zda je povaha chyby ortografická, morfologická, lexikální, syntaktická atd.; (2) o dílčí kategorii, která chybu podrobněji specifikuje (tj. např. zda jde o chybu derivační, flektivní/chybu v rodě, čísle, osobě, čase atd.); příp. obsahuje hierarchická struktura chybové značky i informaci (3) o slovním druhu. Takto je vybudována např. chybová taxonomie korpusů ICLE (tzv. lovaňský systém) a FRIDA (tzv. systém FreeText).[63] Chybové značkování korpusu NICT JLE je také lingvisticky orientované, vychází však od klasifikace slovních druhů, nikoli od jazykové domény, resp. roviny.

Chybové značkování, které aplikují současné korpusy jazyka nerodilých mluvčích a které jsem představila výše, umožňuje úspěšně zhodnotit schopnosti nerodilých mluvčích ovládat jazykový systém, tj. především jejich gramatickou kompetenci, problematičtější je však jeho využití při popisu kompetence komunikační. Zajímavé a do budoucna velmi slibné jsou proto pokusy o značkování chyb z hlediska jejich vlivu na komunikační úspěšnost, resp. analýza nabývání komunikační kompetence u nerodilých mluvčích. Prozatím je však výzkum v této oblasti na začátku.[64]

[62] Příklady převzaty z Nicholls (2003, s. 573) a Dagneaux et al. (1998, s. 166).
[63] Viz Díaz-Negrillo a Fernandéz-Domínguez (2006) a Granger (2003a).
[64] Srov. Izumi et al. (2005).

3.3. Analýza vybraných žákovských korpusů

V následující části podrobněji představíme šest vybraných žákovských korpusů a zaměříme se především na způsob zachycování chyb v textech nerodilých mluvčích, který dané korpusy používají.[65]

Lovaňský korpus ICLE v současnosti zahrnuje dvacet jedna subkorpusů, členěných podle prvního jazyka respondentů, od nichž byla data získávána. Jeho budování řídí Sylviane Granger z belgické Université catholique de Louvain. Japonský korpus NICT JLE, jenž se soustředí také na angličtinu jako cizí jazyk, je příkladem mluveného žákovského korpusu, který se cíleně zaměřuje na podrobnou klasifikaci a následnou analýzu chyb, jež se vyskytují v angličtině japonských mluvčích, a buduje proto bohatý systém chybového značkování. Korpus vzniká v National Institute of Information and Communications Technology v japonském Kjótu pod vedením Emi Izumiho. Žákovský korpus MELD, který vznikl na Montclair State University ve Spojených státech pod vedením Eileen Fitzpatrickové a Steva Seegmillera, je specifický svým přístupem k evidenci chyb, protože neaplikuje žádnou predeterminovanou chybovou taxonomii, ale zaměřuje se pouze na emendaci, tj. na postulování cílové hypotézy. Do mnohamilionového korpusu Cambridge International Corpus, který je vytvářen v rámci nakladatelství Cambridge University Press, se začleňuje i korpus CLC, který patří k největším světovým žákovským korpusům. Je částečně chybově značkován a vzhledem k tomu, že se jedná o komerční typ korpusu, je pro běžného uživatele nedostupný. Německý korpus FALKO, který je budován pod vedením Anke Lüdelingové na Humboldt-Universität v Berlíně, je unikátní svým konceptem chybové anotace, která vychází z požadavku zahrnout při anotaci variantní rekonstrukční hypotézy. Z toho důvodu využívá tento žákovský korpus systému víceúrovňové distanční anotace. Do následujícího přehledu je začleněn i tzv. pilotní korpus PiKUST, který je prozatím jediným doloženým žákovským korpusem zaměřeným na slovanský jazyk. Vznikal jako součást disertační práce Mojci Stritarové na Filozofické fakultě Univerzity v Ljubljani a v současné době je nedostupný. Na jeho základě vzniká (je v počáteční fázi budování) rozsáhlejší korpus slovinštiny jako cizího jazyka.

3.3.1. ICLE – International Corpus of Learner English

Rozsah:	3,7 mil.
Úroveň znalosti:	pokročilí

[65] Navržené chybové taxonomie jsou prezentovány pro lepší ilustraci také na příkladech, které pocházejí z databanky projevů nerodilých mluvčích, shromažďované pro žákovský korpus CzeSL. Pro autentické příklady z jednotlivých korpusů viz příslušnou literaturu.

Metadata:	respondent: 19 text: 6
Chybová anotace:	ano (částečně)
Lingvistická anotace:	POS

3.3.1.1. Korpus

Žákovský korpus ICLE je v současnosti pravděpodobně nejvlivnějším projektem v této oblasti výzkumu. Jeho současný rozsah je tři miliony slov a na jeho výstavbě se podílí osmnáct, resp. dvacet šest[66] pracovišť z celého světa, tj. korpus zahrnuje data od studentů s osmnácti různými mateřskými jazyky. Rozsah jednotlivých subkorpusů je dimenzován na dvě stě tisíc slov. Korpus ICLE je prezentován jako rozsáhlý, vyvážený soubor objektivních dat pro popis žákovského jazyka, protože jako takový je nezbytnou podmínkou jakéhokoli validního výzkumu nabývání a učení druhého, resp. cizího jazyka.[67] Budování korpusu ICLE sleduje dva hlavní cíle: za prvé možnost srovnávací analýzy interlanguage u studentů na pokročilé úrovni znalosti angličtiny, kteří mají různé mateřské jazyky;[68] za druhé možnost komparace s jazykem rodilých mluvčích.[69] Více k tomuto korpusu v kapitole 1.

3.3.1.2. Metadata

V kapitole 1 jsme již zmínili, že žákovský korpus ICLE je budován podle přísných kritérií. Jeden respondent se může podílet maximálně tisícem slov, minimální rozsah vzorku není stanoven. Striktně řízena je podoba esejí zahrnovaných do databáze, akceptovány jsou argumentační eseje, příp. eseje ze zkoušek z literatury. Popisné, narativní a odborné texty do korpusu zahrnovány nejsou. Korpus se zaměřuje na jazyk pokročilých studentů. Profil respondenta zahrnuje devatenáct proměnných (věk, pohlaví, národnost, mateřský jazyk, jazyk otce/matky, vzdělání, délka studia, jazyk výuky apod.), dalších šest parametrů se týká samotného textu (časová limitovanost, referenční pomůcky apod.). Dále viz např. i Granger (1998, s. 9n).

3.3.1.3. Chybová anotace

Data žákovského korpusu ICLE jsou automaticky slovnědruhově anotována pomocí nástroje TOSCA-ICLE, který je aplikován na bezchybné úseky textů. Chybové

[66] Osmnáct subkorpusů je v současné době hotových, osm subkorpusů je ve výstavbě.
[67] Srov. Granger (1998).
[68] A při chybové analýze rozhodnout, zda jde o chyby univerzální nebo jazykově specifické.
[69] Tj. s korpusem rodilých mluvčí angličtiny LOCNESS (Louvain Corpus of Native English Essays).

značkování je založeno na porovnání originálního textu a jeho korigované varianty, opravené rodilým mluvčím. Systém klasifikace chyb korpusu ICLE[70] je hierarchický a zahrnuje dvě roviny popisu, tzv. hlavní chybové kategorie a chybové subkategorie (celkem čtyřicet tři tagů). Hlavní chybové kategorie se vztahují primárně k lingvistickým rovinám popisu jazyka: forma, interpunkce, gramatika, lexiko-gramatika, registr[71], styl. Separátní kategorie odrážejí povrchovou klasifikaci chyb (tj. nadbytečné/chybějící/špatně umístěné slovo), tento klasifikátor nelze v anotačním systému ICLE kombinovat s lingvistickým popisem chyby. Další jazykovou kategorizaci umožňuje pouze formální klasifikátor „chybné užití".[72] Chybové subkategorie určují slovní druh výrazu, kterého se chyba týká, a typ chyby (např. pravopis, slovesný čas, stupeň, ne/počitatelnost, člen apod.). Kategorie jsou v tagu hierarchicky uspořádány (např. GVN – gramatika, slovesa, číslo; WR – slovo, redundantní apod.), srov. příklad (7a, b, c).[73] Pro anotování chyb v korpusu ICLE se využívá pro manuální anotaci specifický editační nástroj vyvinutý v Lovani, tzv. Error Editor.

(7) a. *(WR) Jsem 0 studuju na univerzitě.*
W – slovo, R – nadbytečné

b. *Na (FS) skříňi $skříni$ je rádio.*
F – forma, S – pravopis

c. *mám (GNC) sestra $sestru$*
G – gramatika, N – substantivum, C – pád

3.3.2. NICT JLE – National Institute of Information and Communications Technology Japanese Learner English Corpus

Rozsah:	2 mil.
Úroveň znalosti:	různé
Metadata:	respondent, text, diskurz: 30
Chybová anotace:	ano (částečně)
Lingvistická anotace:	POS (jen u chybových výrazů)

[70] Někdy nazývaný jako „lovaňský systém" (Díaz-Negrillo et al., 2006, s. 93).
[71] Termín *registr* používáme v souladu s teorií systémové funkční lingvistiky M. A. K. Hallidaye. V českém odborném diskurzu se termín objevuje např. u Hoffmannové (1997, s. 163).
[72] Tj. *misuse/replacement.*
[73] Viz i Díaz-Negrillo et al. (2006, s. 94).

3.3.2.1. Korpus

Mezi žákovskými korpusy se jen několik málo z nich zaměřuje na mluvený jazyk. Mezi takové patří japonský korpus NICT JLE.[74] V současné době obsahuje dva miliony slov, resp. 1281 patnáctiminutových interview nahraných při mluvních zkouškách z angličtiny.[75] Jedná se o tzv. pseudolongitudinální databázi, která díky jedinému zdroji a standardizované klasifikaci do devíti úrovní znalosti umožňuje sledovat vývojová stadia analyzovaného interlanguage. Ačkoli cílem vybudování NICT JLE korpusu bylo původně konstruování modelu interlanguage japonských studentů angličtiny, soustředí se autoři v současnosti zejména na vývoj automatické detekce žákovských chyb.

3.3.2.2. Metadata

Žákovský korpus NICT JLE uvádí třicet základních parametrů pro správní anotaci. Lze je v zásadě rozdělit do tří skupin: parametry časoprostorově charakterizující interview, parametry reprezentující profil respondenta a parametry diskurzu.

3.3.2.3. Chybová anotace

Lineární chybová anotace NICT JLE je zacílena na formální aspekty žákovského jazyka, tj. morfologické, gramatické a lexikální chyby. Obsahuje čtyřicet sedm tagů, které jsou vymezovány podle slovních druhů (substantiva, adjektiva, zájmena, slovesa, adverbia, prepozice, spojky, citoslovce), příp. jiných specifických kategorií (modální slovesa, relativa, členy, ostatní). Tyto základní kategorie jsou dále podrobněji lingvisticky definovány. Např. kategorie „adverbium" zahrnuje subkategorie „chyba v adverbiální flexi, ve stupňování adverbia, v lexému", kategorie „relativa" zahrnuje subkategorie „chyba v pádu, lexému" atp. Srov. příklad (8) a (9).

(8) *Tam je malý <n_lxc="rybník">proud</n_lxc>*
n – substantivum, lxc – lexikální chyba

(9) *<v_fml="Bál">Bojal</v_fml> jsem se ...*
v – sloveso, fml – chyba ve formě

Navržená chybová taxonomie byla aplikována na 135 tisíc slov, která byla součástí sondy analyzující akvizici morfémů u japonských mluvčích angličtiny.[76] Ačkoli

[74] Srov. Izumi et al. (2004, 2005).
[75] An English oral proficiency interview test ACTFL-ALC Standard Speaking Test (SST).
[76] Evaluace výzkumu Dulayové a Burtové (Dulay, Burt, 1974).

je NICT JLE korpusem mluveným, jeho chybová anotace se nezabývá fonetickými chybami, neřeší ani interpunkci a ortografii. V současnosti pracují autoři korpusu na dvou specifických úkolech. Jde primárně o vytvoření chybové taxonomie vhodné pro měření komunikační kompetence studentů, která by měla sloužit k odhalení chyb, jež jsou příčinou nesrozumitelnosti a nepřirozenosti jazykového projevu, a jejich odlišení od „menších" chyb, které úspěšné komunikaci nebrání. Druhým úkolem je práce na vytvoření automatické detekce chyb.

3.3.3. MELD – Montclair Electronic Language Database

Rozsah:	0,1 mil.
Úroveň znalosti:	pokročilí
Metadata:	respondent a text: 21
Chybová anotace:	ano (částečně)
Lingvistická anotace:	Ne

3.3.3.1. Korpus

Žákovský korpus MELD zahrnuje psané projevy pregraduálních studentů na pokročilé úrovni znalosti cílového jazyka, v tomto případě angličtiny. Respondenti pocházejí z heterogenního jazykového prostředí (16 mateřských jazyků). Korpus akceptuje elektronicky i ručně psané texty, jejichž tvorba nebyla časově limitována, předepsaná délka vzorků není stanovena. Databáze obsahuje přibližně 100 000 slov, polovina z nich je chybově emendována.[77] Databáze MELD je specifická v tom, že se zaměřuje pouze na shromažďování dat reprezentujících angličtinu jako druhý, nikoli cizí jazyk[78] a jejím cílem je přispět k revizi výsledků analýz v oblasti nabývání druhého jazyka. Autoři přijímají hledisko Gerharta Nickela (1989, s. 298), že nedostatečné oddělování cizího a druhého jazyka je částečně příčinou protichůdných výsledků výzkumů nabývání druhého jazyka.

3.3.3.2. Metadata

Data v korpusu jsou externě značkována jedenadvaceti proměnnými. Ke každému respondentovi jsou shromážděny standardní demografické údaje, včetně znalosti dalších jazyků a úrovně znalosti cílového jazyka, která je odvozována od délky stu-

[77] Informace ke korpusu MELD odpovídají stavu k roku 2004, kdy byl projekt ukončen. „The database currently consists of 44,477 words of tagged text and another 53,826 words of text ready to be tagged. We expect to add another 50,000 words each year; if a funding source is found, we will accelerate this pace." (Fitzpatrick, Seegmiller, 2004, s. 3)
[78] K rozdílům viz výše.

dia. Texty jsou označeny co do časové limitovanosti úkolu a pro případné navazující longitudinální analýzy jsou i datovány.

3.3.3.3. Chybová anotace

Důležitým rysem korpusu MELD, kterým se odlišuje od ostatních světových žákovských korpusů, je metoda anotování, resp. emendování chyb. Metodologicky vychází z předpokladu, že cílem studia druhého jazyka je dosáhnout performanční úrovně prvního jazyka a cílem žákovského korpusu je umožnit srovnání projevů žákovského jazyka a projevů rodilých mluvčích.[79] Autoři korpusu se vědomě odchýlili od běžného postupu predeterminované chybové klasifikace. Anotátoři MELDu nemají k dispozici žádné chybové tagy, tj. nevycházejí z předem daného manuálu chybové taxonomie. Detekované chyby jsou klasifikovány v rámci dvou široce pojímaných domén: chyby lexiko-syntaktické a stylistické. Při anotaci je chybová pasáž podle principu minimální opravy v doméně lexiko-syntaktické rekonstruována jako {chyba/rekonstrukce}, viz příklad (10a, b). Stylistické problémy se opravují jako [chyba/rekonstrukce].

(10) a. *Bojím se {pes/psa}.*
b. *Studovala {0/jsem} tam angličtinu.*

Problémem takové rozvolněné anotace chyb je častá mezianotátorská neshoda v tom, co považovat v daném kontextu za chybu a do jaké domény řešený problém spadá. Srov. např. příklad (11) a (12).

(11) a. ? *Studenti {píše/píšou} test.*
b. ? *{Studenti/student} píše test.*

(12) a. ? *Mám {hodne/hodně} kamarády.*
b. ? *Mám {hodne/hodně} {kamarády/kamarádů}.*

Tyto nesrovnalosti jsou řešeny diskusí a následnou shodou v anotátorském plénu. Daný způsob značkování, resp. rekonstrukce klade velké nároky na anotátory, na jejich odbornost. Zároveň podle našeho názoru v tomto konceptu nelze uspokojivě kvantifikovat výsledky anotace, stejně jako by z důvodu lineárního modelu anotačního formátu bylo potenciálně problematické zachycené chyby jednotně klasifikovat, a to i přesto, že autoři považují možnost různorodé interpretace stejné chyby za přínosné.

[79] Viz Fitzpatrick a Seegmiller (2004, s. 4).

Žákovský korpus MELD není lingvisticky značkován, proběhly pouze dílčí pokusy o uplatnění slovnědruhové anotace pomocí Brill taggeru (Brill, 2005).

3.3.4. CLC – Cambridge Learner Corpus

Rozsah:	35 mil.
Úroveň znalosti:	různé
Metadata:	respondent: 6 text: podle parametrů zkoušky
Chybová anotace:	ano (částečně)
Lingvistická anotace:	Ne

3.3.4.1. Korpus

Žákovský korpus CLC, který je součástí Cambridge International Corpus (CIC),[80] připravuje v Cambridge University Press ve spolupráci s Cambridge ESOL.[81] CLC obsahuje třicet pět milionů slov, z nichž patnáct milionů je chybově kódováno. Materiály, které jsou do korpusu zahrnovány, pocházejí z mezinárodních, standardizovaných ESOL zkoušek různých úrovní, od začátečnických po vysoce pokročilé.[82] Žákovský korpus je mezinárodní, zahrnuje texty od studentů se sto třiceti různými mateřskými jazyky. Je budován jako zdroj pro tvorbu slovníků, výukových a testových materiálů, a zároveň je částečně využíván pro analýzy nabývání cizího jazyka. CLC je komerčním korpusem, který není veřejně dostupný, je však možné jej po dohodě zpřístupnit pro badatelské účely. Srov. i kapitolu 2.

3.3.4.2. Metadata

Profil respondenta zahrnuje šest základních parametrů (první jazyk, národnost, věk, pohlaví, úroveň znalosti angličtiny a informace o studiu angličtiny). Informace o typu textu a realizačních faktorech jsou standardizovány dle typu zkoušky.

[80] CIC v současnosti zahrnuje více než miliardu slov a člení se do dílčích subkorpusů, např. Cambridge and Nottingham Corpus of Discourse in English (CANCODE), Cambridge Corpus of Business English, Cambridge Corpus of Financial English, Cambridge Corpus of Academic English atd.

[81] ESOL, tj. English for Speakers of Other Languages. UCLES, tj. University of Cambridge Local Examination Syndicate.

[82] Zkoušky KET, PET, FCE, CAE, CELS apod.

3.3.4.3. Chybová anotace

Chybové značkování žákovského korpusu CLC se neopírá o systematickou taxonomii žákovských chyb, jeho cílem je okamžitá praktická využitelnost. Veškerá data CLC jsou manuálně značkována pouze dvěma anotátory, aby byla zaručena co největší konzistentnost tagování. Chybová taxonomie žákovského korpusu CLC vychází z kombinace povrchové a lingvistické deskripce žákovských chyb, ačkoli tento dvojdimenzionální přístup není aplikován důsledně.[83] Taxonomie chyb, která má celkem osmdesát osm možných kódů, je primárně založena na dvouúrovňovém systému klasifikace, jež zahrnuje pětičlennou skupinu tzv. obecných typů chyb, tj. chybějící element, redundantní element, chybný slovosled, chybná forma, chybná derivace, a slovnědruhovou klasifikaci čítající osm typů.[84] Tato základní kategorizace je doplněna o kódování interpunkčních chyb, chyb souvisejících s počitatelností substantiv, chyb ve shodě a sadu nesystematizovaných, doplňkových chybových kódů, např. idiomatická chyba, kolokační chyba, chybný slovesný čas, chybný slovosled, chyba v negaci, pravopisná chyba atd. Příklady značkování chyb v korpusu CLC uvádíme v (13) a (14). Specifickým chybovým tagem je kód pro tzv. falešné přátele *(false friends),* který lze použít v případě, že se daný výraz vyskytuje v evidenčním seznamu tohoto druhu chyb.

(13) *Už jsme se <#X>někdy|nikdy</#X> neviděli.*
N – chyba v negaci

(14) *Petr <#MV> |je</#MV> doma.*
M – chybějící element, V – sloveso

3.3.5. FALKO – Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache

Rozsah:	0,264 mil.	
Úroveň znalosti:	pokročilí	
Metadata:	respondent:	20
	text:	3 + dle parametrů zadání
Chybová anotace:	ano	
Lingvistická anotace:	POS, lemma	

[83] Viz Nicholls (2003, s. 572n.).
[84] Pronoun (A), conjunction (C), determiner (D), adjective (J), noun (N), quantifier (Q), preposition (T), verb (include modals (V), adverb (Y).

3.3.5.1. Korpus

Německý korpus FALKO vzniká na Humboldt-Universität v Berlíně a v současné době obsahuje 264 432 slov. Člení se do tří částí (a celkem pěti subkorpusů) podle typu textu a mateřského jazyka respondentů. Za žákovský korpus ve vymezeném slova smyslu lze považovat pouze tři z těchto subkorpusů zahrnující projevy nerodilých mluvčích. První částí je tzv. korpus Falko Summary, který obsahuje (1) texty sumarizující jazykové a literární studie psané pokročilými studenty němčiny jako cizího jazyka (úroveň C1–C2 podle SERR[85]); (2) obdobné texty od rodilých mluvčích, které vznikaly za víceméně stejně řízených podmínek jako texty nerodilých mluvčích; (3) doplňkově jsou v rámci korpusu FALKO k dispozici i původní lingvistické a literární stati, které sloužily jako předlohy pro sumarizující texty. Druhou částí korpusu FALKO je tzv. korpus Falko Essay zahrnující (1) argumentační eseje pokročilých studentů němčiny jako cizího jazyka; (2) argumentační eseje rodilých mluvčích němčiny produkované opět za stejných podmínek jako v případě nerodilých mluvčích. Třetí částí německého korpusu je longitudinální subkorpus studentů na různé úrovni znalosti cílového jazyka shromažďovaný na Georgetown University ve Washingtonu. Rozsah žákovské části korpusu FALKO je přibližně 175 tisíc slov a zahrnuje texty od studentů s přibližně třiceti různými mateřskými jazyky.

3.3.5.2. Metadata

Žákovský korpus FALKO uvádí podrobné informace o respondentech, registruje dvacet parametrů, včetně věku, data, kdy se student začal učit cílový jazyk, a informace o délce a typu institucionalizované výuky. Proměnné týkající se materiálu jsou dány řízenými podmínkami vzniku textů (sleduje se časový limit, využití referenčních materiálů atd.).

3.3.5.3. Chybová anotace

Ve všech subkorpusech korpusu FALKO jsou automaticky značkovány slovní druhy a lemmata pomocí nástroje Treetagger (Schmid, 1994). Pro manuální anotaci chyb je využit anotační model editoru EXMARaLDA (Schmidt, 2001). FALKO jako prozatím jediný žákovský korpus na světě používá flexibilní architekturu, která umožňuje anotování na separátních, nezávislých rovinách.[86] Chybová taxonomie je založena na osmi lingvistických kategoriích (ortografie, slovosled, shoda, dominace, čas atd.), každá z těchto kategorií je pak dále specifikována v souvislosti s dílčími kroky chybové analýzy, tj. identifikací, popisem a explanací chyby. Rovina každé cí-

[85] Společný evropský referenční rámec.
[86] Srov. Lüdeling et al. (2008).

lové hypotézy se člení na tři části: jde o podrovinu identifikace, podrovinu deskripce, které využívá deskriptivní lokalizační taxonomii chyb vycházející z tradičního popisu větné struktury v němčině,[87] a podrovinu explanační (viz tab. 2).

Tabulka 2: Prezentace cílové hypotézy v žákovském korpusu FALKO[88]

Ctok	dass	nur	er	…	Konnte	durch	dieses	Tor	eingelassen	werden	
ZH					durch dieses Tor eingelassen werden konnte						
WO Identification					X						
WO Description					MF_RSK						
WO Explanation					Transfer						

ctok = původní text rozdělený na tokeny, ZH = cílová hypotéza, WO = chyba ve slovosledu, MF = „Mittelfeld", tj. střední pole, RSK = „rechte Satzklammer",[89] tj. pravá část větné konstrukce

Jednotlivé subkorpusy žákovské části FALKO korpusu jsou anotovány odlišně, ačkoli veškeré značkování vychází z uvedení jedné či více cílových hypotéz. Subkorpus Falko Summary může být anotován několikanásobně vždy na základě konkrétní cílové hypotézy a na tuto část korpusu FALKO je aplikována syntaktická anotace (Doolittle, 2009).

V návaznosti na premisu, že při definování cílové hypotézy lze buď maximálně respektovat formální podobu původní výpovědi a omezit anotátorský prostor pro rekonstrukci autorské intence, nebo reflektovat intenci autora a při rekonstrukci, resp. spíše při interpretaci, se významně odchýlit od originální podoby textu, je v subkorpusu Falko Essay možné při anotaci prezentovat dvojí typ rekonstrukce, tzv. minimální a maximální cílovou hypotézu. V příkladu (15b, c) prezentujeme ukázku dvou druhů cílových hypotéz, jejichž uvedení koncept chybové anotace subkorpusu Falko Essay umožňuje. Cílová hypotéza 1 (ZH1) je minimální rekonstrukcí původního textu, cílová hypotéza 2 (ZH2) je maximálně interpretační rekonstrukcí (na lexikální, sémantické a pragmatické rovině).[90]

(15) a. *Praha, bude líbit jí se.*

ctok	Praha	,			bude	líbit	jí	se	.
ZH1			se	jí	bude	líbit			.
ZH1Diff		DEL	MOVT	MOVT			MOVS	MOVS	

[87] Viz např. Eisenberg, P. *Grundriss der deutschen Gramatik. Band 2: Der Satz.* Stuttgart: Metzler, 1999.

[88] Převzato z Lüdeling et al., 2005, s. 7.

[89] Tj. typ větné konstrukce, při které stojí finitní slovesná část predikátu v oznamovací větě na druhé pozici, v tázací větě na první pozici (a ve vedlejší větě na konci); nefinitní část predikátu pak na konci věty.

[90] Dále viz Reznicek et al. (2010a).

DEL – odstraněno, MOVS – přesun zdroj, MOVT – přesun cíl

b. Převzato z Reznicek et al. (2010a, s. 35)[91]

ctok	über	sich	selbst	und	ihre	Erwachsenwerdenprobleme				schreiben
ZH2	über	sich	selbst	und	ihre	Probleme	mit	dem	Erwachsenwerden	schreiben
ZH2Diff						SPLIT				

SPLIT – rozděleno

c. Převzato z Reznicek et al. (2010b)

LT	Das	ist	vielleicht	das	Grund		,	aus	dem		mir		das	Wort	und	die	Ideologie	-
ZH2a	Das	ist	vielleicht	der	Grund	dafür	,			dass	ich		dieses	Wort	und	die	Ideologie	
ZH2b	Das	ist	vielleicht	der	Grund		,			weshalb	ich	für	das	Wort	und	die	Ideologie	-
ZH2c	Das	ist	vielleicht	der	Grund		,			warum	ich		das	Wort	und	die	Ideologie	(
ZH2d	Das	ist	vielleicht	der	Grund		,	aus	dem		ich		das	Wort	und	die	Ideologie	-

wenn	man		so	sagen	kann	-	,		die	sie	sich		hineinsteckt		,	nicht	echt	im	Begriff	habe	.
wenn	man		so	sagen	kann		,		die			darin	enthalten	ist	,	nicht	wirklich		begriffen	habe	.
wenn	man		so	sagen	kann	-	,	was				darin	steckt		,	kennen			Begriff	habe	.
wenn	man	es	so	nennen	kann	)	,		die			darin	steckt		,	nicht	wirklich		begriffen	habe	.
wenn	man		so	sagen	kann	-	,		die			darin	steckt		,	nicht	wirklich	im	Griff	habe	.

Jedním ze specifických rysů anotace materiálu v žákovském korpusu FALKO je značkování na makrotextové rovině, na které jsou typologicky charakterizovány úseky textů (nadpis, citace, alternativní vyjádření, cizojazyčná sekvence apod.).

3.3.6. PiKUST (Poskusni korpus usvajanja slovenščine kot tujega jezika)

Rozsah:	0,035 mil.
Úroveň znalosti:	středně pokročilí, pokročilí
Metadata:	respondent: 10 text: 2
Chybová anotace:	ano (částečně)
Lingvistická anotace:	POS (jen u chybových výrazů)

[91] Reznicek et al. (2010a, s. 35).
CTOK: Plötzlich können sie, über sich selbst und ihre **Erwachsenwerdenprobleme** schreiben und es ist interessant für die Gesellschaft.
ZH2: Plötzlich können sie über sich selbst und ihre **Probleme mit dem Erwachsenwerden** schreiben und es interessiert die Gesellschaft.

3.3.6.1. Korpus

Pilotní žákovský korpus slovinštiny jako cizího jazyka byl budován v letech 2006–2007 jako subtilní testovací korpus pro vytvoření a ověření značkovacích pravidel a principů chybové anotace, které by reflektovaly problémy slovanského, flektivního jazyka. Zpracovávání tohoto korpusu se úzce opírá o model norského žákovského korpusu ASK.[92] PiKUST obsahuje 34 873 slov od respondentů s různými mateřskými jazyky (celkem osmnáct). Největším subkorpusem jsou texty od chorvatských a srbských mluvčích, celkem jde o šedesát sedm procent celkového rozsahu korpusu. Většina textů začleněných do korpusu pochází od respondentů s pokročilou znalostí slovinštiny, texty respondentů s nižší úrovní znalosti jsou do korpusu zahrnuty jen okrajově.[93] Materiálem v tomto korpusu jsou převážně argumentační eseje, které pocházejí z větší části (92 %) z jazykových zkoušek, v menší části jde pak o texty z výukových hodin (domácí úkoly, eseje ze vstupních testů apod.).

3.3.6.2. Metadata

Ačkoli správní anotace materiálů začleněná do korpusu PiKUST je obsáhlá a zahrnuje množství metalingvistických informací týkajících se respondenta a textu (úroveň znalosti slovinštiny, rodinní příslušníci mluvící slovinsky, věk, vzdělání, profese, doba studia slovinštiny atd.), je třeba poznamenat, že tento korpus byl budován značně oportunisticky a nelze jej považovat za vyvážený.

3.3.6.3. Chybová anotace

Ve slovinském korpusu jsou částečně manuálně tagovány slovní druhy (jako jedenáctý slovní druh jsou uvedeny zkratky), pouze však v rámci chybového značkování. Slovnědruhové označení je ručně přiřazeno vždy, pokud se chyba omezuje na jednoslovný výraz. Manuální chybová anotace korpusu PiKUST je založena na lingvistických kategoriích, kombinovaných s formálním popisem. Chyby jsou klasifikovány ve dvou rovinách, podle chybové domény a podle detailnější lingvistické specifikace, příp. je uveden typ povrchové realizace. PiKUST rozlišuje čtyři domény (ortografickou, lexikální, morfologickou a strukturní), v rámci jednotlivých domén pak i jemnější lingvistickou klasifikaci, která zahrnuje jedenáct značek. Např. doména ortografických chyb se dále zjemňuje na chyby pravopisné, chybné hranice slov, chyby v použití velkých písmen a chyby interpunkční. Do domény lexikální jsou zahrnuty i chyby ve vidu, prefixaci, chyby slovnědruhové apod. Srov. např. příklad (16).[94] Strukturní doména slouží jako zastřešující pro chyby syntaktické a chyby ve slovních spojeních. Morfologická doména žádné detailnější rozpracování nemá. Fakt, že slovinština je jazykem s bohatou flexí, se odráží v množství morfologických chyb v textech nerodilých mluvčích. Stritarová (2009) chápe snahu o podrobnější kategorizování tohoto typu chyb jako příliš interpretativní a subjektivní. Při anotaci je explicitně prezentována i cílová hypotéza. Specificky jsou označeny nerekonstruovatelné chyby, příp. chyby s přílišnou interpretací. Toto značkování ale není řízeno a je intuitivní.

(16) *Dopoledne jsem (LexNonEx)lopatil $házel$ sníh*
Lex – lexém, NonEx – neexistující

Výše uvedené příklady zřetelně ukazují, že se žákovské korpusy vzájemně značně odlišují jak ve zvoleném anotačním modelu, tak ve způsobu značkování i v charakteru chybové taxonomie. Ačkoli se dlouhodobě diskutuje o standardizaci chybové anotace žákovských korpusů a o nutnosti obecného konceptu její chybové typologie, mezi odbornou veřejností nedošlo prozatím ke shodě.[95] Jednotlivé klasifikace chyb jsou vždy pevně svázány s jednotkami (morfémy, slovy apod.), které jsou kategorizovány, a s výzkumným záměrem projektu, v jehož rámci vznikají. Nezanedbatelná část žákovských korpusů se však zaměřuje na budování komplexní, metodologicky pevně ukotvené a zároveň dostatečně obecné chybové taxonomie (NICT JLE, NOCE, FALKO, FRIDA aj.). Podstatným faktem ovlivňujícím podobu klasifikace chyb v žákovských projevech je to, že koncepce chybové analýzy, resp. počítačem podporované chybové analýzy, vždy předpokládá porovnávání vyjádření nerodilého mluvčího a rekonstrukce tohoto vyjádření v cílovém jazyce. Nejběžnějším přístupem k chybové taxonomii uplatňované v žákovských korpusech s chybovou anotací je přístup založený na jazykových kategoriích,[96] protože umožňuje nejen detailnější popis konkrétních chyb, ale je také vhodným východiskem pro kvantifikační analýzy. Zároveň alespoň částečně reflektuje lingvistickou teorii, v jejímž rámci vzniká. Obvykle se jedná o schéma hlavních kategorií, např. slovních druhů jako u korpusu NICT JLE, nebo jazykových rovin jako např. u korpusu ICLE, a podrobnějších podkategorií. Klasifikace chyb v souvislosti s přiřazením k jazykové rovině, příp. ke slovnímu druhu nemusí být vždy jednoznačná, některé deviace[97] jsou kategorizová-

[92] ASK, tj. Norsk andrespråkskorpus.

[93] Korpus neobsahuje žádné texty respondentů na nízké úrovni znalosti slovinštiny s výjimkou tří textů od začátečníků Slovanů. Srov. Stritar (2009, s. 137): *„It is a general principle among learner corpora compilers to exclude real beginners due to the instability of their interlanguage and the complexity that error tagging of their texts would require."*

[94] Bohužel autentický záznam grafické anotace z korpusu PiKUST není k dispozici, rekonstruovali jsme tedy ukázku na základě korpusu ICLE.

[95] Srov. Díaz-Negrillo a Fernández-Domínguez (2006, s. 86).

[96] Podrobněji o klasifikaci chyb v textech nerodilých mluvčích viz např. James (1998), Dulay et al. (1984), v češtině Štindlová (2011).

[97] Termín *deviace* používáme ve smyslu odchylky od standardní podoby cílového jazyka.

ny a priori, přesto považujeme tento typ tagování za relativně spolehlivý a dostatečně deskriptivní.

Vedle lingvistické charakteristiky žákovských chyb je často uplatňovaným přístupem kategorizace chyb na základě jejich povrchové realizace. Tradičně se vymezují čtyři základní typy cílových modifikací (chybějící element, přebývající element, chybná forma/chybný výběr a chybný slovosled), které mají reprezentovat čistě deskriptivní taxonomii chyb, zaměřující se pouze na pozorovatelné, povrchové rysy chyb, která nepropojuje popis chyby s její explanací. Někteří badatelé chápou takovou klasifikaci žákovských chyb jako reflexi kognitivních procesů odehrávajících se při studentově rekonstruování cílového jazyka.[98] Tato myšlenka vyvolává řadu metodologických diskusí, protože předpokládá operace s povrchovými strukturami cílového jazyka spíš než budování interlanguage.

Minimálně se v chybové anotaci žákovských korpusů odráží jiné typologie chyb, např. Corderova klasifikace na chyby formální utvářenosti výrazu[99] a chyby nevhodného užití, které autor dále člení na referenční chyby, chyby v registru, sociolingvální chyby a chyby textové.

Jednou z metodologických otázek chybové anotace žákovského korpusu je, do jaké míry by měly být při klasifikaci chyb v jazyce nerodilých mluvčích odděleny jednotlivé kroky chybové analýzy, tj. především popis a explanace chyby, a je-li to vůbec možné. S výjimkou žákovského korpusu FALKO, v jehož anotačním schématu jsou jednotlivé kroky striktně separovány, chybové taxonomie světových žákovských korpusů toto téma explicitně nekomentují.[100]

Návrh konzistentního seznamu chybových značek, resp. zařazení žákovských chyb k těmto značkám, je však problematický, protože (1) široká variabilita žákovských chyb zasahuje všechny lingvistické oblasti, (2) je nesnadné vymezení hranice a rozsahu chyby, (3) klasifikace chyb je vždy určitým způsobem závislá na interpretaci anotátora. Především z těchto důvodů některé projekty chybovou anotaci zpochybňují a navrhují jiné koncepty mapování žákovského jazyka.[101] Odmítnutí počítačem podporované chybové analýzy jako metodologického konceptu výzkumů jazyka nerodilých mluvčích na základě chybově anotovaného žákovského korpusu má kořeny v původní kritice tradiční chybové analýzy zaměřené především proti nedostatkům v její metodologii, tj. obtíže při identifikaci jednotného zdroje chyby[102] a komplikované budování rámce pro deskripci žákovských chyb. Zároveň jsou výhrady zaměřené proti jejímu omezenému dosahu. Studie žákovského jazyka ze sedmdesátých let, založené na chybové analýze, se soustředily primárně na chybu, tj. pouze na jeden aspekt žákovského jazyka, neřešily otázky spojené s vývojem znalosti cílového jazyka a nereflektovaly strategii vyhýbání se v užívání cílového jazyka nerodilými mluvčími.[103]

Ačkoli vzhledem k uvedeným výhradám byla chybová analýza některými badateli odmítána jako nespolehlivá, je studium žákovských chyb kontinuálně využíváno jako validní součást performanční analýzy.[104] Chyby jsou neoddělitelnou součástí žákovského jazyka a jako takové by měly být podrobovány výzkumu stejně jako jiné aspekty mezijazyka. Chybová analýza má své uplatnění při analýze nabývání cizího jazyka a velký dosah pro pedagogickou praxi.[105] Viz i dále kapitola 8. Existence elektronických korpusů žákovského jazyka, budovaných podle striktních výstavbových kritérií, nabízí přístup k žákovskému jazyku jako celku, umožňuje jeho detailní, kvantifikační analýzy a příp. podporuje i aplikaci variabilní chybové taxonomie.

Ačkoli tvůrci žákovských korpusů uvádějí jako jeden z hlavních důvodů pro budování těchto databází žákovského jazyka přispívání k analýzám mezijazyka a k výzkumům nabývání druhého, příp. cizího jazyka, mají v současné době žákovské korpusy, resp. bádání na nich založené, větší vliv a využití v oblasti vyučování cizím jazykům. Analýzy chybově značkovaných dat zprostředkovávají povědomí o žákovské performanci; díky nim jsou odkrývány frekvenční vzorce chyb a jsou aktualizovány pedagogické potřeby studentů cílového jazyka (tj. jsou mapovány oblasti, na které je třeba se ve výukovém procesu konkrétní cílové skupiny zaměřit). Podrobněji v kapitole 8.

Někteří odborníci[106] jsou však na druhou stranu přesvědčeni, že ačkoli chybové taxonomie založené na jazykových kategoriích, příp. na typu povrchové modifikace významně ovlivňují následné pedagogické aplikace, mají jen omezený podíl na odkrývání toho, jak se žák učí cizí jazyk. Problematické je podle nich především vymezení hranice chyby, detekce zdroje chyby v mentálním lexikonu žáka a chápání povrchových modifikací jako akvizičních faktů. Zároveň zpochybňují konzistentnost a spolehlivost chybového značkování v souvislosti s jeho interpretační povahou a subjektivitou. Rastelli (2009) navrhuje při anotaci mezijazyka rezignovat na chybové značkování a navrhuje tzv. SLA-značkování.[107] Předpokládá, že příliš striktní pohled, tj. pohled řízený prizmatem cílového jazyka, na data žákovského korpusu není adekvátní, protože cílem výzkumů nabývání cizího jazyka je interlangauge,

[98] Srov. Dulay et al. (1982, s. 150n.).

[99] Tento termín používáme v širším slova smyslu, tj. nikoli pouze ve významu české formální morfologie. Srov. Corder (1974, s. 123n.).

[100] Srov. např. komentář Grangerové (Granger, 2003a, s. 467) k chybové anotaci žákovského korpusu FRIDA, která je „*descriptive rather than interpretative*".

[101] Viz Fitzparick a Seegmiller (2003), Rastelli (2009) aj.

[102] Srov. tzv. „*ambiguous goofs*", Dulay, Burt (1974b).

[103] Srov. Schachter a Celce-Murcia (1977), Long a Sato (1984), Van Els et al. (1984) aj.

[104] K termínu viz např. Larsen-Freeman a Long (1992).

[105] „*Although error analysis certainly has its limitations, it must be regarded as an important key to a better understanding of the process underlying L2-learning.*" (Ringbom, 1987, s. 69)

[106] Srov. např. Ellis (1994), Corder (1974).

[107] Srov. také Rastelli (2007), Rastelli a Frontini (2008).

a pouhý fakt, že některé formy mezijazyka se zdají být korektní a některé nekorektní, není dostatečně vypovídající ani o problémech žáka, ani o jeho mentální gramatice. SLA-značkování by mělo pomoci výzkumníkům odhalovat systematičnost (či nesystematičnost) v tom, jak studenti mapují/internalizují formy a funkce a jak budují znalost cílového jazyka. Návrh značkování žákovského korpusu, které je ukotveno v metodologii výzkumů nabývání cizího jazyka, je prozatím ve vývojové fázi.

4. Anotace chybových textů v českém žákovském korpusu[108]

Vladimír Petkevič, Alexandr Rosen, Barbora Štindlová, Tomáš Jelínek, Milena Hnátková, Petr Jäger

Tato kapitola je věnována anotaci žákovského korpusu CzeSL. Anotace je tu míněna v širším smyslu: jako celý proces zpracování vstupního ručně psaného textu až do jeho výstupní emendované a lingvisticky anotované (značkované) podoby. Mimo vlastní lingvistickou anotaci, tj. lexikální, morfologickou a syntaktickou emendaci a značkování chyb v trojrovinném anotačním systému, popíšeme tedy i věci související: přepis vstupních textů, správu a organizaci anotace i příslušné softwarové nástroje.

Smyslem tvorby emendovaného a lingvisticky anotovaného korpusu obecně je mj. umožnit učitelům češtiny jako cizího jazyka rychle a efektivně zjišťovat, jakých typů chyb a v jaké míře se dopouštějí studenti češtiny, pro něž čeština není mateřským jazykem. Korpus navíc poskytuje reprezentativní data, která umožní systematický výzkum češtiny jakožto cizího jazyka. Na základě statistik a rešerší v nashromážděných korpusových datech si tak učitelé mohou učinit objektivní představu o chybách studentů a adekvátně zaměřit svou výuku a přípravu výukových materiálů. To přispěje k nápravě stavu, kdy systematická příprava učitelů na výuku češtiny jako cizího jazyka je v počátcích a výuka probíhá často intuitivně podle individuálních zkušeností vyučujícího.

4.1. Koncepce anotace a anotační schéma

4.1.1. Anotační schéma jako kompromis

Chybově anotovaný žákovský korpus češtiny je mimo slovinský korpus PiKUST (Stritar, 2009) jediným žákovským korpusem slovanského jazyka; na rozdíl od něho jsme se však rozhodli jej anotovat se zřetelem ke specifickým vlastnostem češtiny. Ve srovnání s češtinou mají jazyky, pro něž existují anotované žákovské korpusy, jednodušší flexi a/nebo méně volný slovosled. V koncepci naší chybové taxonomie se

[108] Tento příspěvek částečně vychází z článku Škodová et al. (2011).

tedy musely odrazit specifické vlastnosti češtiny jako jazyka s bohatým flektivním podsystémem a volným slovosledem, a bylo tedy třeba řešit zcela nové problémy, aby chybová anotace žákovského korpusu CzeSL umožňovala podrobné statistické zpracování relevantních jazykových dat. Vytvoření anotačního schématu a efektivní chybové taxonomie je však z uvedených důvodů – flektivní povaha češtiny a volný slovosled – náročný úkol. Anotační schéma musí navíc 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.

Dále jsme se při tvorbě anotačního schématu museli vyrovnat s některými problémy souvisejícími se stanovením cílové hypotézy, tedy s opravami textu podle předpokládané intence autora: interferencí, interpretací, problematikou slovosledu a stylu.

4.1.1.1. Interference

Jelikož anotátoři nejsou odborníky na osvojování druhého jazyka a nelze u nich předpokládat ani znalosti všech relevantních cizích jazyků, nejsou s to zachytit případy jazykové interference z mateřštiny autora anotovaného textu nebo nějakého jiného jazyka, který autor zná. Není tedy možné od anotátorů požadovat, aby zachycovali interferenční chyby. Tak například věta *Tokio je pěkný hrad* je gramaticky správná, ale její autor, rodilý mluvčí ruštiny, zde chybně užil slovo ***hrad***, které ve vztahu ruštiny a češtiny patří mezi tzv. *falešné přátele*, neboť jeho formální ekvivalent v ruštině, *gorod*, neznamená *hrad*, nýbrž *město*. Podobně je tomu s větou *Je tam hodně sklepů*, která je sice sama o sobě gramaticky správná, ale v daném kontextu nepřípadná. Aby ji však anotátor mohl správně emendovat, musí vědět, že *sklep* znamená v ruštině *hrobka* a v polštině *obchod*.

4.1.1.2. Interpretace

U některých typů chyb spočívá problém ve stanovení interpretačních mezí. Věta *kdyby citila na tebe zlobna* je sice gramaticky chybná, ale dá se jí víceméně rozumět: znamená patrně „kdyby se na tebe zlobila". V takových případech má anotátor spíše za úkol význam věty interpretovat než větu opravovat. Větu lze tedy přeformulovat na *kdyby se na tebe cítila rozzlobená* nebo *kdyby se na tebe zlobila*, přičemž první formulace není tak přirozená jako druhá, zato je blíže původní větě. V takových případech je nesnadné poskytnout anotátorům jednoznačné směrnice, jak postupovat.

4.1.1.3. Slovosled

Jiným typem chyb specifickým pro češtinu jsou nedostatky slovosledné. Ty třeba nerespektují náležité aktuální členění, které český slovosled vyjadřuje. Často bývá obtížné stanovit – a to i v daném kontextu – zda jde o chybu. Například věta *Rádio je taky na skříni* naznačuje, že v místnosti jsou alespoň dvě rádia, z nichž jedno je umístěno na skříni, třebaže s větší pravděpodobností lze větu interpretovat tak, že mezi věcmi nacházejícími se na skříni je také rádio. Tato interpretace ovšem vyžaduje – přinejmenším v projevu psaném – odlišný slovosled, a tedy slovoslednou úpravu: *Na skříni je taky rádio*. Podobně obtížná mohou být rozhodnutí týkající se chyb lexikálních a chyb v modalitě.

4.1.1.4. Styl

Další problematickou oblast představuje pro anotátory dichotomie mezi spisovnou a obecnou češtinou, tedy diglosie: spisovná čeština se dost liší od češtiny obecné, zejména v oblasti flektivní morfologie, a pak vyvstává problém, jak anotovat obecněčeské tvary, jichž autor textu užil. Autoři textů, tj. studenti češtiny, si totiž nemusí dobře uvědomovat jejich postavení v jazykovém systému češtiny a náležitý kontext, v němž by jich měli užívat. Ačkoli takové výrazy jsou třeba gramaticky správné, v korpusu CzeSL se nahrazují svými standardními protějšky a vždy se značkují jako stylově příznakové, neboť se předpokládá, že student ve skutečnosti chtěl užít nepříznakové formy.

Výsledné schéma a typologie chyb, jež je jeho podkladem, představuje tedy jistý kompromis mezi omezeními, která z praktického hlediska klade proces anotace, a výzkumnými požadavky kladenými na žákovský korpus. Korpus se může využívat k porovnávání variet žákovské češtiny, resp. verzí mezijazyka různých nerodilých mluvčích, s ohledem na vymezený standard cílového jazyka (tj. češtiny). Mezijazyk konkrétního mluvčího lze kromě jiných kritérií charakterizovat podle jeho prvního jazyka nebo etnické příslušnosti, tedy např. jako mezijazyk ruský, vietnamský, romský apod.

4.1.2. Anotace na více rovinách

O chybové anotaci nelze předem stanovit, jaká by měla být její ideální podoba. Do značné míry záleží na cílech a možnostech projektu, a samozřejmě i na typu jazyka. Jednoúrovňové anotační schéma by stačilo pro úzce definovaný účel, např. ke zkoumání morfologických zvláštností jazyka studentů. Mohlo by zachycovat i více aspektů, pokud by se příslušné údaje daly připojit k původním formám. Pro naše účely však s sebou jednoúrovňová anotace přináší řadu problémů. Především je korpus CzeSL z hlediska budoucího využití koncipován velmi široce, takže se nelze omezit na úzký okruh jazykových jevů nebo určitou rovinu popisu. Z toho vyplývá nutnost

zaznamenávat postupné opravy a udržovat vazby mezi původní a opravenou formou i u změn ve slovosledu, změn v hranicích mezi slovy, případně i u vypuštěných a přidaných slov. Dalším důvodem je pak potřeba anotovat chyby, které se týkají více forem najednou, často v nekontaktním postavení.

V ideálním případě by anotátor měl mít k dispozici právě tolik rovin, kolik je třeba k provedení anotace, která může být i postupná. To lze zajistit buď volbou z většího počtu lingvisticky motivovaných rovin, nebo možností vytvářet roviny anotace podle aktuální potřeby oprav dané formy. Vzhledem k tomu, že anotátor by neměl být příliš zatěžován teoretickými dilematy a že výsledná anotace by měla být jednotná, zdá se velký nebo proměnlivý počet rovin pro naše účely málo vhodný. Proto jsme přijali kompromisní řešení – anotátor má pro anotaci k dispozici dvě roviny, třetí rovinou je rovina obsahující původní, nezpracovaný text (srov. i oddíl 3.2.1.2 o víceúrovňové distanční anotaci v kapitole 3). Rozhodnutí, na jaké rovině se daná forma opravuje, je dáno do značné míry formálními kritérii, ale rozdíly mezi oběma rovinami přitom mají lingvistické opodstatnění.

Rovina R0 obsahuje původní text, přepsaný z rukopisu se zachováním některých rukopisných charakteristik (varianty, nečitelné řetězce). Na rovině R1 se emendují izolované formy bez ohledu na kontext – typicky jde o překlepy a chyby v pravopisu a morfologii. Výsledkem je řetězec správných českých tvarů, i když věta z nich složená správně být nemusí. Všechny ostatní typy chyb (valence, shoda, slovosled a další) se opravují na rovině R2.

4.1.3. Formalismus

Anotované žákovské korpusy někdy využívají datových formátů a nástrojů vyvinutých původně pro anotování mluveného jazyka. Takové prostředí umožňuje arbitrární segmentaci výstupu a několikaúrovňovou anotaci segmentů (srov. Schmidt, 2009). Obvykle anotátor edituje tabulku se sloupci korespondujícími se slovy a řádky podle úrovní anotace. Buňky lze rozdělovat a spojovat tak, aby bylo možné anotovat rozdělená slova nebo posloupnosti slov jako celek, např. při opravě chyb ve shodě nebo slovosledu (Lüdeling et al., 2005).

Tabulkový formát však není příliš vhodný pro jazyky s volným slovosledem a bohatou flexí: jeden slovní tvar totiž může být chybný z různých hledisek. V krajních případech může být problematický typograficky, ortograficky, morfosyntakticky, lexikálně i slovosledně zároveň. Při slučování a rozdělování buněk tabulky však nelze zaručit, že zůstanou zachovány korespondence mezi postupně opravovanými formami. Proto jsme přistoupili k vlastnímu návrhu, kde se korespondence mezi postupně opravovanými formami vyjadřují explicitně.

Naše anotační schéma má podobu grafu složeného ze tří vzájemně propojených paralelních rovin, které představují původní text studenta (R0) a dvě úrovně anotace (R1 a R2). Každému slovu vstupního textu včetně interpunkce obvykle odpovídá nějaký uzel na každé ze tří rovin. Běžně je vztah mezi uzly na sousedních rovinách 1:1, ale slova se mohou také spojovat a rozdělovat, vypouštět i přidávat. Ve vzájemném vztahu mohou být i potenciálně nespojité posloupnosti slov, takže obecně může být počet uzlů na sousedních rovinách spjatých jedním vztahem neomezený.

Kromě tvaru mohou být u každého uzlu uvedeny další informace – lemma, morfosyntaktické kategorie, syntaktická funkce apod. Pokud byla původní forma (případně více forem) opravena na jinou, mohou být vztahy mezi uzly na sousedních rovinách opatřeny údaji o typu chyby. Na obr. 1 níže uvádíme příklad víceúrovňové anotace podle tohoto schématu.

Kromě vztahů mezi sousedními rovinami schéma také umožňuje vyjádřit jednoduché syntagmatické vztahy související s chybami určitého typu, např. u shody nebo rekce. Identifikátor chyby na spojnici mezi opravovaným a opraveným výrazem může odkazovat na jiný výraz, který určuje správnou podobu chybného slovního tvaru, např. v případě chybného tvaru finitního slovesa na podmět nebo na jiný tvar se stejnými kategoriemi shody (viz oprava *jsme* na *jsem* na obr. 1).

Častým jevem jsou tzv. sekundární chyby, jako třeba v příkladu *dívá se na americkém filmu.* Adjektivum *americkém* se náležitě shoduje s řídícím substantivem, ale po opravě pádu předmětu na akuzativ je třeba změnit i pád shodného přívlastku. V takových příkladech se používá více odkazů: od předmětu ke slovesu jako zdůvodnění opravy pádu řídícího substantiva a od adjektiva k substantivu jako zdůvodnění opravy pádu shodného přívlastku. U přívlastku jde přitom o opravu, která je vynucena jinou opravou, tzv. opravu sekundární. Při značkování chyb se tento atribut zaznamenává.

Od počátku jsme si vědomi toho, že – alespoň v netriviálních případech – lze chybu identifikovat pouze na základě stanovení hypotetické cílové podoby chybného výrazu, přičemž někdy nemusí být nasnadě podoba jediná. Práce s více cílovými hypotézami zatím existuje jako teoretická možnost.

4.2. Chybová taxonomie a její evaluace

Typický student češtiny jako cizího jazyka chybuje na všech lingvisticky motivovaných rovinách, od grafémiky až po pragmatiku. Navržené anotační schéma se z praktických důvodů omezuje na konzervativní emendaci, jejímž výsledkem je souvislý a gramaticky správný text, ale bez nároků na stylistickou vytříbenost. Anotátor by také neměl text příliš volně interpretovat. Pokud text není dostatečně srozumitelný, mohou být příslušné pasáže takto anotovány, ale mohou zůstat bez emendace.

Východiskem pro taxonomii chyb jsou lingvistické kategorie ve spojení s formálním popisem chyby (typem modifikace). Ne všechny typy chyb je nutné určovat

manuálně. Pokud je to možné, určujeme některé chyby automaticky porovnáním původní a opravené podoby tvaru a/nebo na základě výsledků automatické lemmatizace a morfologické analýzy (viz oddíl 4.3.5.3). Emendace zatím probíhá jen ručně, i když se zkoumá možnost využít automatický korektor.

4.2.1. Chyby na rovině R1

Na rovině R1, kde se opravují chyby zjistitelné bez ohledu na kontext, se kromě chyb v pravopisu a hranicích slov zachycují také chyby ve flektivní a derivační morfologii i chybné slovní základy, např. nově vytvořená nebo cizí slova. Tyto nedostatky se s výjimkou chyb pravopisných určují manuálně. Výsledkem opravy je nejpodobnější správný tvar, který může být dále na rovině R2 podle kontextu opraven na jiný – důvodem je například porušení morfosyntaktické shody nebo sémantická nekompatibilita lexému. Seznam chyb anotovaných manuálně na rovině R1 s příklady uvádí tabulka 1. Poslední tři chyby (***stylColl***, ***stylOther*** a ***problem***) se používají i na rovině R2.

Tabulka 1: Chyby na rovině R1

Typ chyby	Popis	Příklad
incorInfl	nesprávná flexe	***spám*** *málo*; *tři* ***měsícu***
incorBase	nesprávný slovní základ	*kočka se* ***jmemuje***; ***libila*** *se mi*; *musíš to* ***posvětlit***
fwFab	neemendovatelné, „vymyšlené" slovo	*je tam hodně* ***jinaků***
fwNC	cizí slovo	*jím rád* ***eggs***; *byla v* ***hangu***
flex	doplňující příznak u chyb *fwFab* a *fwNC* značící přítomnost flexe	*jdu do* ***shopa***
wbdPre	prefix oddělený mezerou a předložka bez mezery	*Petr* ***při jde***; ***dolesa***
wbdComp	neoprávněně rozdělená kompozita	***český anglický*** *slovník*
wbdOther	jiná chyba týkající se hranice slova	***mochezký***; ***atak***
stylColl	obecněčeský tvar	***dobrej*** *film*
stylOther	knižní, nářeční, slangový, hyperkorektní výraz	*holka s hnědými* ***očimi***
problem	problémová chyba (doplňkový příznak)	

Pravidlo, že na rovině R1 musí být všechny tvary správné, neplatí bez výjimky – chybu nelze opravit třeba proto, že anotátor nedokáže rozpoznat intenci autora. Na druhé straně se správný tvar nahrazuje jiným správným tvarem v případech, kdy jde evidentně o pravopisnou nebo hláskovou chybu, jejímž výsledkem bylo náhodné homonymum s existujícím tvarem.

4.2.2. Chyby na rovině R2

Opravy na rovině R2 se týkají chyb ve shodě, valenci, analytických tvarech, zájmenném odkazování, záporové shodě, v užití vidu, času, stupně, lexému a idiomu, a také ve slovosledu. U chyb ve shodě, valenci, analytických tvarech, zájmenném odkazování a záporové shodě lze při opravě chybného výrazu obvykle odkázat na jiný správně utvořený nebo již opravený výraz, který určuje morfologické kategorie nebo jiné vlastnosti výrazu opravovaného. Typy manuálně určovaných chyb na rovině R2 uvádí tabulka č. 2. (Mezi automaticky identifikované chyby patří např. chyby ve slovosledu nebo podrobnější členění chyby typu ***vbx***.)

Tabulka 2: Chyby na rovině R2

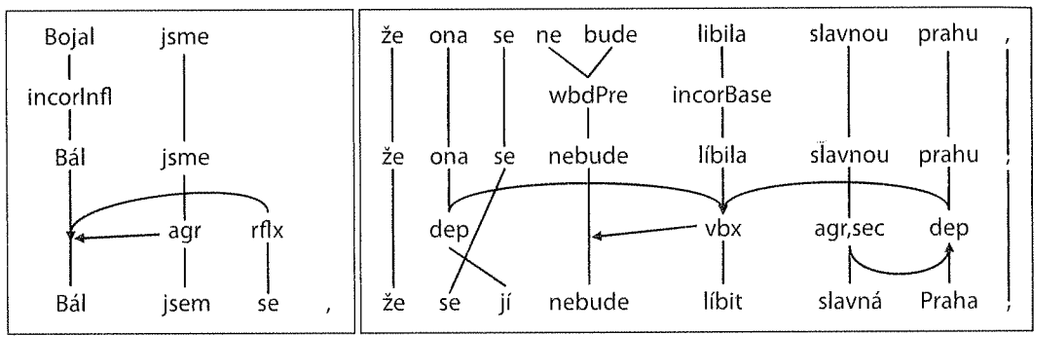
Typ chyby	Popis	Příklad
agr	narušení shody	*máme* ***hezkých*** *psa*; *Petr* ***vařím*** *oběd*
dep	chyba ve vyjádření syntaktické závislosti	*věřím* ***učitelku***; *káva bez* ***mléko***; *bojím se jí* ***zavolám***
ref	chyba v zájmenném odkazu	*paní,* ***jenž*** *jsem potkal*
vbx	chyba v analytickém slovesném tvaru a složeném přísudku	*Jana bude* ***dělá***; *guláš* ***bylo chutná*** *mi*; *začal* ***pracuje***
rflx	chyba v reflexivním výrazu	*smála* ***si***; *narodila jsem v Petrohradu*
neg	chyba v negaci	***mám*** *žádný čas*; *on* ***ne*** *velký*
lex	chyba v lexiku a frazeologii	*jsem* ***Vietnam***; *kupuju* ***housenky***
use	chyba v užití gramatické kategorie	*tričko je* ***nejvíc nejhezčí***; *celé dopoledne* ***uvařím*** *oběd*; *do polévky dáme* ***čočky***
sec	sekundární, „zavlečená" chyba (doplňkový příznak)	*dívá se na* ***americkém*** *filmu*
stylColl	obecněčeský tvar	*viděli jsme* ***hezký*** *holky*
stylOther	knižní, nářeční, slangový výraz	*rozbil se mi* ***hadr***
stylMark	výplňkové slovo jako „diskursní marker"	*no, teda, jo*
disr	rozvrácená konstrukce	*zkušební důvtip může tě řídit*
problem	problémová chyba (doplňkový příznak)	

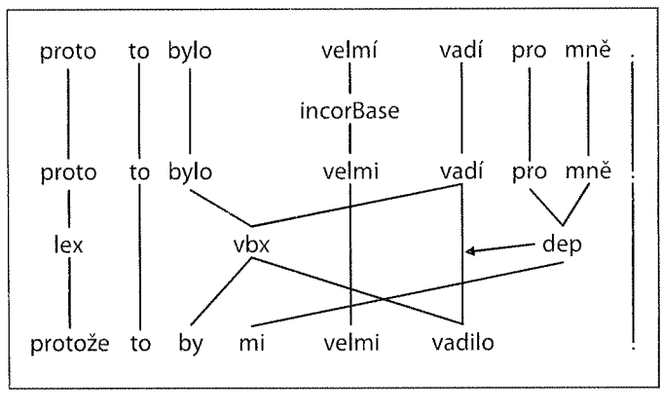
4.2.3. Příklad

Anotační schéma použité v autentickém příkladu uvádíme na obr. 1, z prostorových důvodů je příklad rozdělen na dvě části. Tři paralelní řetězce forem představují původní text a dvě roviny anotace. Jednotlivé tvary jsou spojeny hranami a většina oprav se zároveň označuje kódem typu opravy.

V první části věty se na rovině R1 tvar *bojal* opravil na *bál* s údajem, že má chybný slovní základ. Na rovině R2 se jako chyba ve shodě opravil tvar *jsme* na *jsem* s odkazem na nejbližší tvar, který je z hlediska morfologických kategorií důležitých pro shodu správně (*bojal*). Chybějící reflexivní částice se vložila s odkazem na významové sloveso. Čárka přibyla bez údaje o chybě, který se doplní automaticky.

Ve druhé části věty anotátor chybně oddělenou záporovou předponu spojil se slovesem *bude* a opravil délku v základu tvaru *libila*. Kromě toho opravil i malé začáteční písmeno u vlastního jména *Praha* (bez identifikace chyby, která se doplní automaticky). Na rovině R2 bylo nutné opravit pád zájmena *ona* s odkazem na řídící sloveso, které se z finitního tvaru *líbila* změnilo na infinitiv, neboť je součástí opisného futura – proto anotátor odkazuje na finitní tvar pomocného slovesa *nebude*. Bylo nutné opravit i pád u vlastního jména *Praha*, opět s odkazem na řídící významové sloveso. Tím je ovšem dotčen i původně správný tvar adjektiva *slavnou* – kód pro chybu ve shodě je zde doplněn údajem, že jde o „sekundární" chybu. Slovoslednou úpravu postavení příklonky *se* není třeba označovat kódem chyby – to se provede automaticky. Máme-li na výběr z více možností přesunu, které všechny vedou ke stejnému výsledku, přesouváme přednostně závislé větné členy.





Obr. 1: Příklad anotace jedné věty

Poslední věta vyžadovala na rovině R1 jen jedinou opravu (opět délka ve slovním základu). Zato bylo na rovině R2 nutné kromě spojky (lexikální oprava) změnit celý analytický slovesný tvar, což je příklad opravy typu 2:2, a s odkazem na řídící sloveso pak i předložkový pád zájmena na pád prostý (*mi*) a výsledek nakonec umístit na patřičné místo.

Oprava výrazu *pro mně* na tvar *mi* však opomíjí chybu v pádu zájmena po předložce. Aby anotátor takovou chybu mohl opravit a označit, potřeboval by další rovinu, na níž by mohl opravit *mně* na *mě* s odkazem na předložku, která pád určuje. Opravou už na rovině R1 by anotátor porušil pravidlo, že na R1 se opravují jen tvary chybné i bez kontextu. Tento problém chápeme jako kompromisní řešení, které vyvažuje jednodušší schéma.

4.2.4. Evaluace

Použitelnost anotačního schématu a taxonomie chyb byla ověřena pomocí míry shody mezi anotátory na vzorku 67 textů v průměru po 150 slovech, celkem 9373 slov (7995 slov bez interpunkce). Autory textů byli rodilí mluvčí různých jazyků. Každý text anotovali dva anotátoři, celkem bylo anotátorů čtrnáct. Jako míra shody mezi anotátory byl použit koeficient kappa (Carletta 1996), který kromě shody nebo neshody mezi dvěma anotátory při volbě dané značky bere v úvahu i pravděpodobnost náhodné shody. Blíže o evaluaci viz Štindlová (2011, s. 121n.) a Štindlová et al. (2011).

Na škále mezi dokonalou shodou (kappa=1) a shodou náhodnou (kappa=0) dosáhly hodnoty kappa velmi uspokojivých hodnot např. u značek ***incorBase*** (0,75) a ***incorInfl*** (0,61), z roviny 2 pak u značek ***agr*** (0,54) a ***dep*** (0,47). Obecně se ve srovnatelných případech považují hodnoty nad 0,4 za přijatelné. Část chybových značek jako např. ***lex*** a ***use*** však skončila pod tímto limitem (0,37 a 0,21). Zlepšení (a to i u „úspěšnějších" typů chyb) může nastat po precizaci instrukcí v anotačním manuálu, ale některé značky budou i nadále do značné míry závislé na subjektivním dojmu anotátora a vysokou míru shodu mezi anotátory u nich nelze očekávat.

4.3. Postup anotace

Jakmile jsou ručně psané texty přepsány do elektronické podoby, uloží se do databáze *AMES*.[109] Od okamžiku uložení přepsaných textů do databáze zajišťuje řízení

[109] http://ames.ff.cuni.cz.

dalšího zpracování a správu textů systém *Speed*[110] (viz podrobněji níže). Celé zpracování vstupního textu uloženého v databázi probíhá zhruba v těchto krocích:

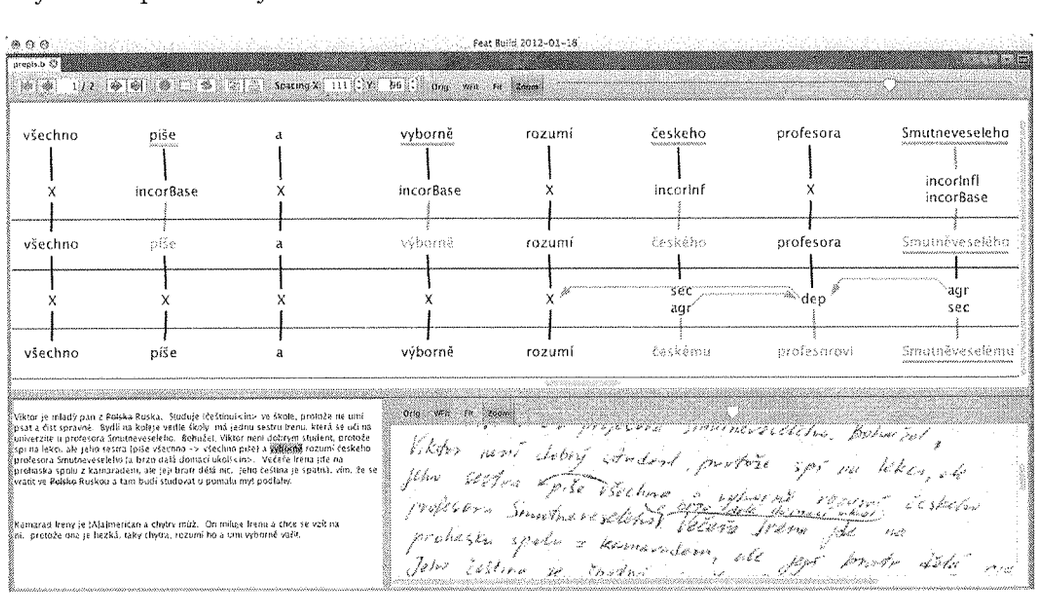
Rukou psaný text se pomocí běžného textového editoru přepíše do elektronické podoby ve formátu HTML, rozšířeném o kódy zachycující studentovy opravy, předtištěný text, text v jiných abecedách atd.

Přepsaný text v elektronické podobě se zkonvertuje do formátu pro anotaci, v němž je automaticky stanovena rovina R0 a výchozí podoba roviny R1. Obě jsou zakódovány ve formátu PML (srov. Pajas a Štěpánek (2006); je to konkretizace XML pro účely strukturní lingvistické anotace).

Anotátor manuálně opraví chyby v textovém dokumentu a určí jejich typ pomocí anotačního editoru *feat*.[111]

V posledním kroku doplní automatické nástroje klasifikaci těch chyb, které lze z ruční anotace odvodit automaticky.

Všechny podoby zpracovávaného textu se ukládají do databáze. Na obr. 2 jsou pro ilustraci uvedených kroků zachyceny různé podoby zpracovávané věty: rukopisná podoba, přepsaná elektronická podoba a emendovaná a označkovaná podoba věty v prostředí trojrovinného anotačního schématu vytvořeného anotačním editorem feat, který zároveň umožňuje všechny tyto podoby zobrazit. Jednotlivé kroky objasníme podrobněji v části 4.3.1 a dále.



Obr. 2. Příklad věty zpracovávané v anotačním editoru *feat*

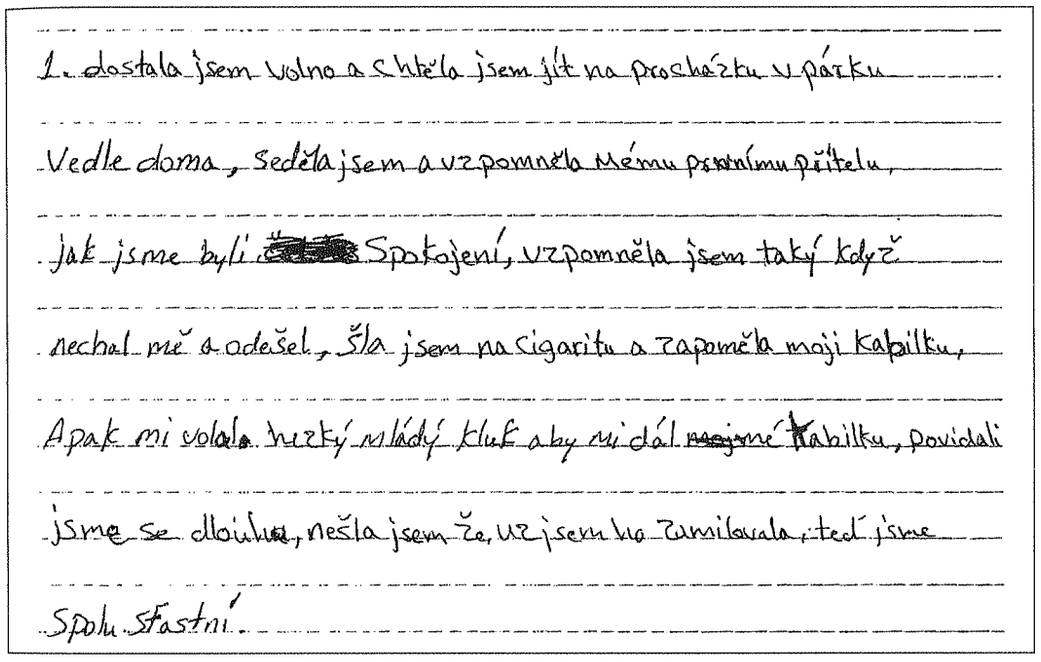
[110] http://speed.aspone.cz.

[111] Softwarový nástroj *feat* (*F*lexible *E*rror *A*nnotation *T*ool) je prostředí určené k vícerovinné anotaci žákovských korpusů, srov. Hana et al. (2010). Je volně dostupný na adrese http://purl.org/net/feat.

4.3.1. Sběr textů a jejich přepis

Na počátku celého procesu anotace jsou původní (vstupní) texty, které většinou píší studenti a žáci ve třídě při jazykových kursech nebo při zkouškách. Je tedy nutné sbírat především rukopisy (ač jsou k dispozici i texty elektronické). Pracovat s rukopisy je však rovněž vhodné proto, že zachycují podobu autentického mezijazyka mnohem věrněji než texty elektronické, jež lze snadno korigovat nebo i vytvářet automatickými nástroji, což by mohlo podobu autentického mezijazyka výrazně zkreslit. Ke sběru dat a jejich přepisu viz i kap. 2.

Na obr. 3 je uveden příklad vstupního textu psaného studentem, jehož mateřským jazykem je arabština a který absolvoval bakalářské studium bohemistiky v Egyptě a dvoutýdenní kurs češtiny v České republice.


1. dostala jsem volno a chtěla jsem jít na procházku v párku

Vedle domu, seděla jsem a vzpomněla mému pravému příteli,

jak jsme byli [redacted] Spokojení, vzpomněla jsem taky když

nechal mě a odešel, šla jsem na cigaritu a zapomněla moji kabilku,

Apak mi volala hezký mládý kluk aby mi dál ~~mojiné~~ kabilku, povídali

jsme se dlouhu, nešla jsem že, už jsem ho zamilovala, teď jsme

spolu. šťastní.

Obr. 3. Vstupní text napsaný egyptským studentem

Rukopisnou podobu vstupního textu přepisují *přepisovači* v textovém editoru Microsoft Word nebo OpenOffice.org Writer podle podrobných pokynů v Manuálu pro přepis.[112] Přepsané texty se ukládají ve formátu HTML do databázového systému k dalšímu zpracování. Na obr. 4 je uveden příklad přepisu vstupního textu zobrazeného na obr. 3.

[112] http://utkl.ff.cuni.cz/~rosen/public/transkripce.pdf, http://utkl.ff.cuni.cz/~rosen/public/transkripce_doplnek.pdf.

I když se snažíme o maximální věrnost, někdy se při přepisu rukopisných textů neobejdeme bez jisté míry interpretace. Přepisovači si musí uvědomovat specifika rukopisu dané skupiny studentů a někdy i jednotlivců (například stejný glyf je možné interpretovat v písmu různých studentů jako písmeno *l*, *e*, nebo *a*). Pokud je možné znak nebo i celý úsek textu interpretovat různě, přepisovač může uvést i více variant. Například velikost počátečních písmen nebo hranice slov jsou často nejasné. Zvlášť se označují zcela nečitelné úseky i opravy, které provádějí sami studenti (vsuvky, škrty) a které mohou být pro výzkum akvizice jazyka rovněž užitečné.

1. dostala jsem volno a chtěla jsem jít na procházku v párku vedle doma, seděla jsem a vzpomněla mému prvnímu přítelu, jak jsme byli spokojení, vzpomněla jsem taký když nechal mě a odešel, šla jsem na cigaritu a zapoměla moji kabilku, A pak mi volal hezký mládý kluk aby mi dál mé kabilku, povídali jsme se dlouh{o|u}, nešla jsem že uz jsem ho zamilovala, teď jsme spolu sťastní.

Obr. 4. Přepsaná elektronická podoba textu z obr. 3

4.3.2. Konverze a správa textů – systém *Speed*

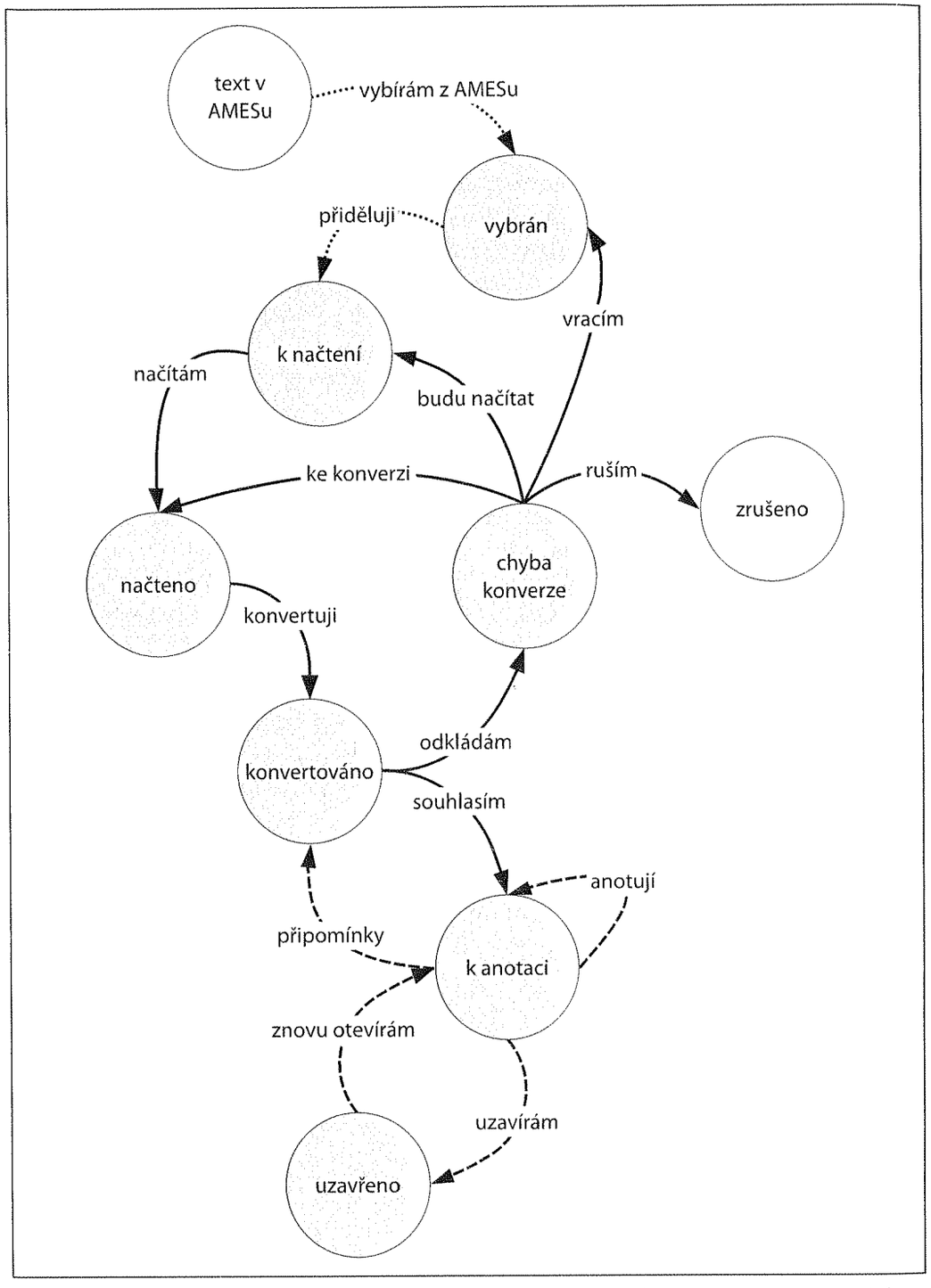
Od okamžiku uložení přepsaných textů do databáze zajišťuje řízení dalšího zpracování textů systém *Speed* pro správu přepsaných, zkonvertovaných a anotovaných textů, vyvinutý v rámci projektu Inovace vzdělávání v oboru čeština jako druhý jazyk. Nejprve se prověřuje formální správnost textů: tato činnost je svěřena *kontrolorům*. Každý přepsaný text je po kontrole zkonvertován do vstupního formátu vhodného pro vlastní lingvistickou anotaci a rovněž uložen do databáze. Konverzi provádí rovněž *kontrolor*, který navíc prověřuje, zda konverze proběhla řádně. Činnost kontrolorů řídí *koordinátoři*, kteří přidělují texty kontrolorům ke konverzi a přebírají od nich zkonvertované texty. Poté, co jsou k dispozici texty zkonvertované pro vlastní lingvistickou anotaci, ujímají se jich *supervizoři*, kteří řídí činnost anotátorů provádějících hlavní činnost: emendaci a lingvistickou anotaci chyb v textech anotačním editorem *feat*. *Supervizoři* konkrétně anotátorům přidělují texty k anotaci, přebírají od nich zkonvertované texty, kontrolují jejich anotace, upozorňují anotátory na nedostatky v jejich práci, evidují časté chyby apod. V zájmu maximální správnosti a konzistence anotace anotují každý text nezávisle na sobě dva anotátoři.

Níže jsou v přehledu uvedeny role uživatelů systému *Speed*, který řídí všechny uvedené činnosti až na vlastní anotaci, jež je svěřena anotačnímu editoru *feat*:

Role uživatelů systému *Speed*

- ***Koordinátor*** – koordinuje kontrolory. Vybírá texty z databáze a přiděluje je kontrolorům. Každý koordinátor má vymezenu skupinu textů, které spravuje. Toto vymezení je definováno na základě dat obsažených v průvodce zpracovávaných textů (například texty od vietnamských či romských mluvčích apod.) a je nastaveno tak, aby jeden text byl vždy ve správě jediného koordinátora. Role koordinátora je identická s rolí koordinátora v databázi, v němž jediný koordinátor spravuje právě jednu subdatabázi. Tak je zajištěno, že jeden a týž text má ve správě jediný koordinátor.
- ***Kontrolor*** – prověřuje formální správnost přepsaných textů uložených v databázi, konvertuje je do podoby vhodné pro lingvistickou anotaci a nakonec prověřuje, zda konverze proběhla v pořádku.
- ***Supervizor*** – řídí skupinu přidělených anotátorů a zodpovídá za správnost anotace textů jim přidělených. Prvotním vodítkem při rozdělování textů do skupin jsou údaje v průvodce textu (např. texty slovanských mluvčích apod.). Rozdělení je nastaveno tak, aby jeden text vždy spadal právě pod jednoho supervizora.
- ***Anotátor*** – provádí anotaci textů. Každý anotátor spadá právě pod jednoho supervizora, který ho řídí.

Schematicky lze vztahy mezi koordinátory, kontrolory a supervizory při zpracování textu znázornit jako na obr. 5. Operace prováděné *anotátory* je na obrázku vyjádřena slovem *anotují*, za kterým se skrývá další podrobný diagram komunikace supervizora a anotátora.



Obr. 5. Základní diagram řízení anotace v systému *Speed* pro správu přepsaných, zkonvertovaných a anotovaných textů.

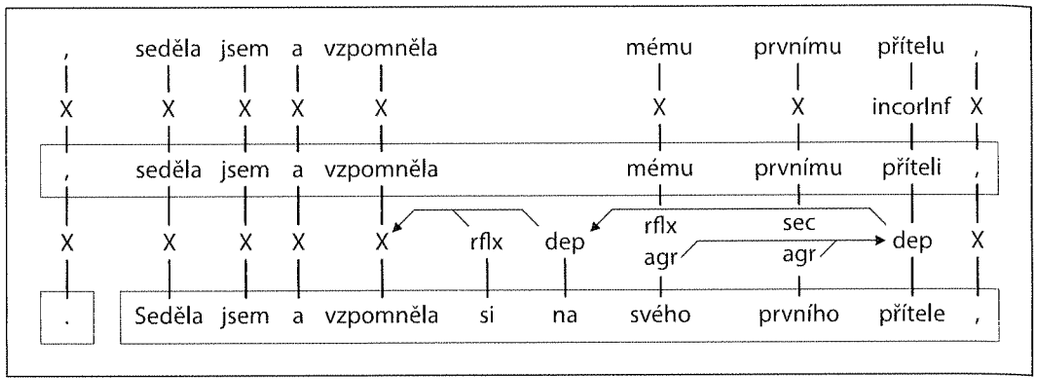
Popis diagramu:

- ***Stavy*** – jsou vyznačeny kroužky. Stav určuje, jaké operace je možné s textem provádět – je dán uloženými daty a operacemi, které jsou pro něj definovány (včetně kontrol dat apod.). Každý *text* se v jednom okamžiku nachází právě v jednom stavu.
- ***Operace*** – jsou vyznačeny šipkami. Převádějí zpracovávaný text mezi stavy, provádějí změny v okolním prostředí. V daném okamžiku lze s jedním textem provést pouze jedinou operaci; ta musí být definována na základě aktuálního stavu textu.
- ***Počáteční a koncový stav*** – jsou stavy *text v AMESu* a *zrušeno*. Jsou to stavy, kde se nachází text jednak před započetím práce, jednak na konci procesu, kdy už neexistuje žádná operace, která by s ním pracovala.
- ***Ostatní stavy*** – jsou vnitřní stavy systému, tj. všechny stavy kromě stavu *text v AMESu* a *zrušeno*.
- ***Operace koordinátora*** – jsou tyto operace označené plnou šipkou v horní části obrázku: *vybírám z AMESu* a *přiděluji*.
- ***Operace kontrolora*** – jsou tyto operace označené tečkovanou šipkou v prostřední části obrázku: *načítám*, *budu načítat*, *ke konverzi*, *vracím*, *ruším*, *konvertuji*, *odkládám* a *souhlasím*.
- ***Operace supervizora*** – jsou tyto operace označené čárkovanou šipkou v prostřední části obrázku: *připomínky*, *anotují*, *znovu otevírám* a *uzavírám*.

4.3.3. Emendace a značkování chyb

Ruční část anotace probíhá v prostředí anotačního editoru *feat*. Anotátor opraví text na příslušných rovinách, upraví vztahy mezi výrazy, které si na jednotlivých rovinách vzájemně odpovídají (implicitně jsou všechny vztahy 1:1), chyby opraví (emenduje) a u některých chyb přidá příslušnou chybovou značku. Při emendaci i značkování se anotátor řídí Manuálem pro anotaci.[113] Na obr. 6 je ukázka anotace části věty z výše uvedeného příkladu (obr. 3 a 4) v prostředí anotačního editoru. Příklad *seděla jsem a vzpomněla mému prvnímu přítelu* byl opraven na *seděla jsem a vzpomněla si na svého prvního přítele*.

[113] http://utkl.ff.cuni.cz/~rosen/public/anotace.pdf.

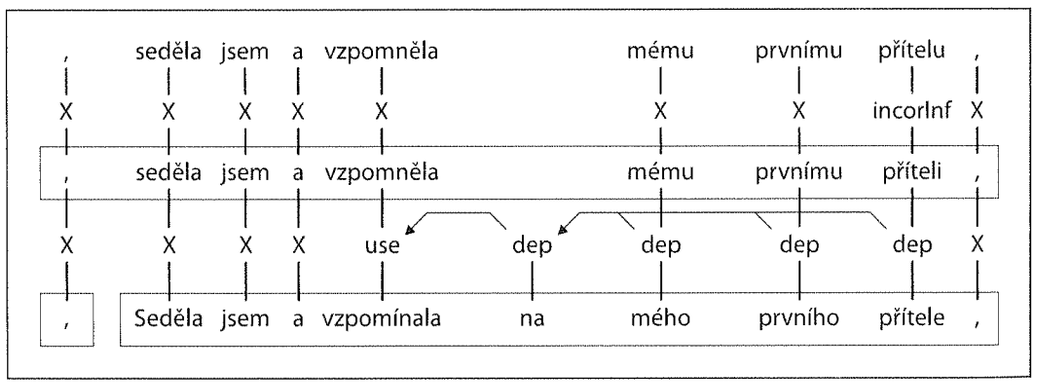


Obr. 6. Ukázka anotace části věty z příkladu na obr. 3 a 4 v prostředí anotačního editoru *feat*

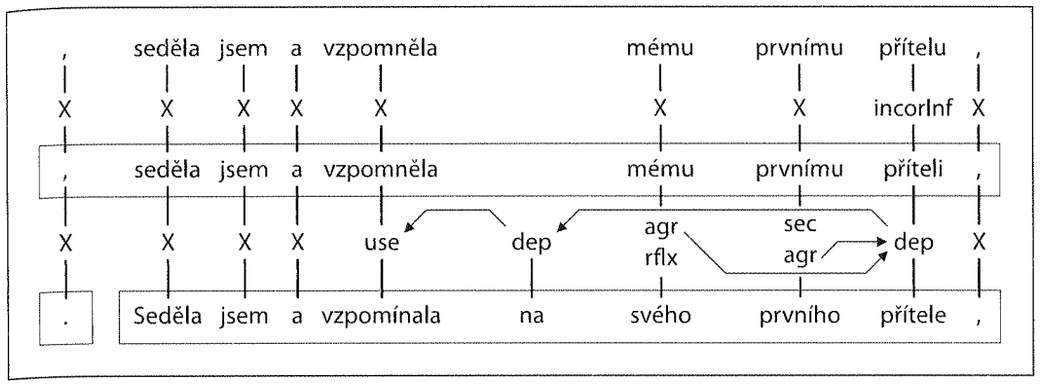
4.3.4. Kontrola a adjudikace

Jakmile anotátor dokončí anotaci svého textu, uloží text do databáze, odkud si jej vytáhne supervizor ke kontrole. Upozorní anotátora na chyby v emendaci i anotaci, ten chyby opraví, poté opravy zkontroluje supervizor a výsledkem je nakonec správně emendovaný a anotovaný text. Supervizor i anotátor se musí řídit Manuálem pro anotaci. Ke kontrole anotace slouží také programy používané k automatickému zpracování – viz níže oddíl 4.3.5.3.

Na obr. 7 je uvedena ukázka chybně anotované části věty z příkladu na obr. 6 v prostředí anotačního editoru (*dep* místo *agr*), na obr. 8 je pak uvedena správná anotace této části textu.



Obr. 7. Ukázka chybně anotované části věty



Obr. 8. Ukázka správně anotované části věty

Každý text je značkován dvěma nezávislými anotátory, a proto se jejich značkování může lišit, i když se oba řídí týmž manuálem pro anotaci. Viz porovnání interpretace části textu z obr. 6 a obr. 8:

Anotátor1: *Seděla jsem a vzpomněla si na svého prvního přítele* (obr. 6)
Anotátor2: *Seděla jsem a vzpomínala na svého prvního přítele* (obr. 8)

Porovnání odlišných anotací je v kompetenci *adjudikátora*, který obě anotace za pomoci adjudikační funkce aplikace *feat* porovná a z obou anotací vytvoří výsledný text, který musí být také v souladu s Manuálem pro anotaci. Tento text je poté zařazen do žákovského korpusu.

4.3.5. Automatické značkování textů na základě provedené manuální anotace

Poslední fází chybové anotace žákovských textů je jejich automatické zpracování pomocí sady počítačových programů. Automaticky se tak rozšiřuje a upravuje chybové značkování a doplňují se lingvistické údaje, které usnadní vyhledávání v korpusu.

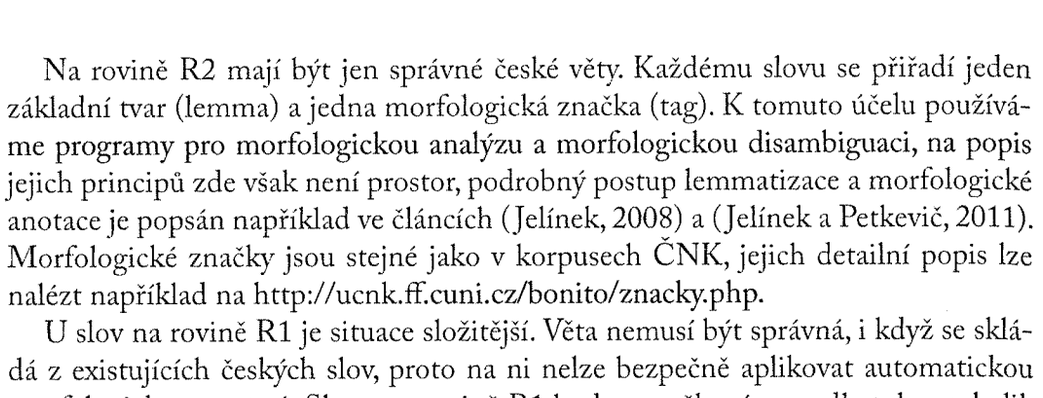
4.3.5.1. Automatické doplnění lingvistických informací

Pro vyhledávání v korpusu žákovských textů je velmi praktické, když uživatel může vyhledávat podle základních tvarů slov nebo podle slovních druhů, pádů a podobně, stejně jako například v rozsáhlých textových korpusech Českého národního korpusu (ČNK). Tyto informace se však při manuální chybové anotaci do textu nevkládají, anotace by tak kladla příliš vysoké požadavky na anotátory, a to časové i intelektuální (podrobná znalost značek, rozlišování mezi slovními druhy, pády aj.). Proto se tyto informace doplňují zcela automaticky.

Na rovině R2 mají být jen správné české věty. Každému slovu se přiřadí jeden základní tvar (lemma) a jedna morfologická značka (tag). K tomuto účelu používáme programy pro morfologickou analýzu a morfologickou disambiguaci, na popis jejich principů zde však není prostor, podrobný postup lemmatizace a morfologické anotace je popsán například ve článcích (Jelínek, 2008) a (Jelínek a Petkevič, 2011). Morfologické značky jsou stejné jako v korpusech ČNK, jejich detailní popis lze nalézt například na http://ucnk.ff.cuni.cz/bonito/znacky.php.

U slov na rovině R1 je situace složitější. Věta nemusí být správná, i když se skládá z existujících českých slov, proto na ni nelze bezpečně aplikovat automatickou morfologickou anotaci. Slovo na rovině R1 bude označkováno podle toho, nakolik se shoduje s odpovídajícím slovem na rovině R2. Pokud se spojené tvary na R1 a R2 shodují, převezme tvar na rovině R1 lemma i tag od slova na rovině R2. Pokud se tvary liší, ale slovo na R1 je jen odlišným tvarem slova na R2, přiřadí program tvaru na R1 stejné lemma jako na R2 a také všechny odpovídající značky, například kdyby tvar na R1 byl *je* a tvar na R2 *jeho*, dostal by tvar na R1 lemma *on* a značku pro zájmeno v akuzativu (střední rod i množné číslo). Pokud se tvary liší a nemohou se ani shodovat v lemmatu, dostane tvar na R1 všechny kombinace značek a lemmat, které určitý tvar může mít (např. tvar *je* by za takových okolností dostal jak lemma *být* a značku pro sloveso, tak lemma *on* či *oni* a značku pro zájmeno v akuzativu).

Na příkladu je vidět chybové značkování spolu s doplněnými lemmaty a morfologickými značkami. Tvary na rovinách R0 a R1 se shodují, na rovině R2 se liší tři slova ze čtyř:



Obr. 9. Ukázka morfologicky anotované věty

Na rovině R2 se každému slovu určilo jedno lemma a jedna značka (např. tvar *je* byl označen jako tvar slovesa *být* v přítomném čase). Na rovině R1 převzalo slovo „důležité" lemma i značku (adjektivum v nominativu singuláru neutra) od tvaru na R2, protože se s tímto tvarem shoduje. Ostatní slova se s tvarem na R2 neshodují, přiřadily se jim tedy všechny možnosti morfologické analýzy (žádné slovo ale nemá více potenciálních lemmat).

4.3.5.2. Automatické rozšíření a úprava chybového značkování

V koncepci chybové anotace se počítalo s tím, že některé typy chyb je možné spolehlivě označit automaticky a lze tak ušetřit práci anotátorům. Především se to týká formálních chyb na rovině R1, u nichž se prostým porovnáním vzájemně si odpovídajících tvarů na rovinách R0 a R1 zjistí typ formální chyby (např. chyba ve znělosti hlásky nebo v palatalizaci). Nutným předpokladem je samozřejmě správná emendace. Vzhledem ke složitosti české slovotvorby (morfematické švy jsou často zastřené, flektivní koncovky nelze zcela oddělit od slovotvorných přípon) však nelze chybu spolehlivě automaticky lokalizovat, tedy určit, zda žák chyboval ve flexi, nebo v základu slova. Toto rozlišení je nutné provést manuálně. Chyby na úrovni R2 je mnohem obtížnější spolehlivě automaticky klasifikovat, v koncepci anotace tak bylo pro automatické značkování vyhrazeno jen několik dílčích jevů.

4.3.5.2.1. Automatické doplnění formálních chyb na rovině R1

Automatické doplnění formálních chyb na rovině R1 je založeno na srovnání tvaru původního slova na rovině R0 a emendovaného slova na rovině R1. Při manuální anotaci (viz výše) se na rovině R1 značkují tyto chyby: nesprávný tvar (***incor***), chybná hranice slov (***wbd***), nově utvořené či cizí slovo (***fw***). Automatické značkování chyb s tímto tříděním nesouvisí, např. u chybně psaných českých slov může být formální chyba doplněna jak u slov, u nichž chyba typu ***incor*** nemá být vyznačena (chybný tvar na R0 se vyslovuje stejně jako emendovaný tvar na R1, např. *prozba/prosba* či *objet/oběd*), tak u slov, která mají být označena ***incorBase*** (např. *hitit/chytit*, *dedečka/dědečka*) nebo ***incorInfl*** (např. *každécho/každého*, *venkovou/venkovem*).

Typologie formálních chyb na R1 vyjadřuje, čím se chybný tvar na R0 liší od emendovaného tvaru na R1. Až na drobné výjimky (např. podkategorie chyb ve znělosti) neurčuje příčinu chyby, pouze pojmenovává jednotlivé typy často se opakujících chyb. Typů formálních chyb je cca 40, po dalším nárůstu počtu emendovaných textů (díky většímu počtu dokladů pro jednotlivé typy) bude možné třídění pozměnit: zjemnit, přidat další typy, nebo naopak zjednodušit.

4.3.5.2.1.1. Třídění formálních chyb na rovině R1

Dosud implementované formální chyby uvádíme v přehledné tabulce s příklady. Některé typy jsou čistě pravopisné (velká/malá písmena, psaní háčku ve spojeních *dě/tě/ně*, spodoba znělosti aj.), výslovnost slova se chybou nemění. Jiné typy vždy ovlivňují výslovnost (kvantita vokálů, vkladné *e*). Další typy mohou být v určitém kontextu pouze formální, v jiném kontextu ovlivňují výslovnost (psaní *i/y*, záměna *c/k*). Chyby, které nespadají do žádné přesně definované kategorie, se třídí podle počtu rozdílných znaků a podle místa, kde se rozdíl projevuje. Tyto chyby jsou uvedeny na konci tabulky.

V prvním sloupci tabulky je typ automaticky přiřazené formální chyby, v druhém sloupci popis chyby, ve třetím dva příklady, pokud se v dosud anotovaných textech vyskytly, jinak příklad jediný. Typy, které se v dosud anotovaných textech nevyskytly, v tabulce uvedeny nejsou.

Tabulka 2: Třídění formálních chyb na rovině R1

Typ chyby	Popis chyby	Příklad
formCap0	chybně použité malé písmo	*evropě/Evropě*; *štědrý/Štědrý*
formCap1	chybně použité velké písmeno	*Staré/staré*; *Rodině/rodině*
formCaron0	chyba v diakritice – chybí háček	*vecí/věcí*; *sobe/sobě*
formCaron1	chyba v diakritice – háček navíc	*břečel/brečel*; *bratřem/bratrem*
formDiaE	chyba v diakritice – *ě/é*, popř. *é/ě*	*usmévavé/usměvavé*; *poprvé/poprvé*
formDiaU	chyba v diakritice – *ú/ů*, popř. *ů/ú*	*nemůžeš*/nemůžeš; *úkoly/úkoly*
formDtn	chyba v psaní *dě/tě/ně*, *di/ti/ni*	*ňikdo/nikdo*; *ještěrka/ještěrka*
formQuant0	chyba v diakritice – chybí čárka nad vokálem	*vzpominám/vzpomínám*; *doufam/doufám*
formQuant1	chyba v diakritice – čárka nad vokálem navíc	*která/která*; *hledát/hledat*
formVoiced0	chybně neznělá/spodoba znělosti	*stratíme/ztratíme*; *nabítku/nabídku*
formVoiced1	chybně znělá/spodoba znělosti	*zbalit/sbalit*; *nigdo/nikdo*
formVoicedFin0	chybně neznělá na konci slova	*Když/Když*; *vztach/vztah*
formVoicedFin1	chybně znělá na konci slova	*přez/přes*; *pag/pak*
formVoiced	ostatní chyby ve znělosti	*pěžky/pěšky*; *hodili/chodili*
formY0	chyba *i/y* (chybně *i*)	*pražskích/pražských*; *vipije/vypije*
formY1	chyba *i/y* (chybně *y*)	*hlavným/hlavním*; *líbyl/líbil*
formYJ0	chybně zaměněné *y* a *j* (*y*)	*yaké/jaké*; *yazykem/jazykem*
formGH0	chybně zaměněné *g/h*	*gost/host*; *gorký/horký*
formCK0	chybně zaměněné *c/k*, mimo palatalizaci (*c*)	*Atlantic/Atlantik*
formPalat0	neprovedená palatalizace (*k,g,h,ch*)	*amerikě/Americe*; *matkě/matce*
formEpentE0	chyba v epentet. *e* (chybí *e*)	*najdnou/najednou*; *domček/domeček*
formEpentE1	chyba v epentet. *e* (chybně *e* navíc)	*rozeběhl/rozběhl*; *účety/účty*
formEpentJ0	chybí *j* po *i* před vokálem	*napie/napije*
formEpentJ1	chybně vložené *j* po *i* před vokálem	*dijamant/diamant*
formGemin0	chybně nezdvojené písmeno	*polostrově/poloostrově*
formGemin1	chybně zdvojené písmeno	*essej/esej*; *professor/profesor*
formJe0	chyba *je/ě* (chybně *ě*)	*ubjehlo/uběhlo*; *Nejvjetší/Největší*
formJe1	chyba *je/ě* (chybně *je*)	*vjeděl/věděl*; *vjeci/věci*
formMne0	chyba mně/mě (chybně mě)	*zapoměla/zapomněla*; *nejvýznamějších/nejvýznamnějších*
formMne1	chyba *mně/mě* (chybně *mně*, *mňe*, *mňě*)	*mněla/měla*; *rozumněli/rozuměli*
formProtJ0	chyba v protetickém *j* (chybí *j*)	*sem/jsem*; *menoval/jmenoval*
formProtJ1	chyba v protetickém *j* (chybně *j* navíc)	*jse/se*; *jmé/mé*
formProtV1	chyba v protetickém *v* (chybně *v* navíc)	*vosm/osm*; *vopravdu/opravdu*
formMeta	metateze, prohození dvou znaků	*dobrodružtsví/dobrodružství*; *provůdce/průvodce*

Jiné chyby (blíže nerozlišené), pouze jeden znak

Typ chyby	Popis chyby	Příklad
formMissChar	jiný chybějící znak (chybí jeden znak)	*protže/protože*; *oňostroj/ohňostroj*
formRedunChar	jiný přebývající znak (přebývá jeden znak)	*opratrně/opatrně*; *zrdcátko/zrcátko*
formSingCh	chyba vznikla záměnou jednoho znaku za druhý	*otevřila/otevřela*; *vezmíme/vezmeme*

Jiné chyby (blíže nerozlišené), více znaků

Typ chyby	Popis chyby	Příklad
formPre	podrobněji neurčená chyba v prefixu	*poletěla/letěla*; *potrávíme/trávíme*
formHead	podrobněji neurčená chyba na začátku slova (ne v prefixu)	*rustala/zůstala*; *žijna/října*
formTail	podrobněji neurčená chyba na konci slova	*holkamá/holkami*; *nezajína/nezajímá*
formUnspec	podrobněji neurčená chyba kdesi uprostřed slova	provudkyně/průvodkyně; kreřénu/kterému

4.3.5.2.1.2. Automatická úprava chyb wbd (chyby v hranici slova)

Chyby v hranici slov ***wbd*** jsou manuálně rozdělovány do tří kategorií: ***wbdPre*** (chybně oddělený prefix nebo chybně připojená předložka), ***wbdComp*** (chybně rozdělené složené slovo) a ***wbdOther*** (ostatní chyby v hranici slova). Automaticky se tyto chyby doplňují o informaci, zda je na rovině R0 tvar chybně rozdělený (k typu chyby se připojí ***-Split***, např. ***wbdPreSplit***, chybně oddělený prefix) nebo je tvořen chybně spojenými slovy (k typu chyby se připojí ***-Joined***, např. ***wbdOtherJoined***, chybně spojená dvě slova, ne předložka).

4.3.5.2.2. Automatická úprava a rozšíření chybové anotace na rovině R2

Většina chybové anotace na rovině R2 se provádí manuálně, variabilita chybných struktur je totiž natolik vysoká, že se spolehlivá automatická chybová anotace stává velmi obtížným úkolem. Při anotaci se tak automaticky zpracovávají jen některé dílčí úlohy. Doplňuje se označení chyby v reflexivitě (značka ***rflx***) u chyb ***dep*** (závislost), ***ref*** (zájmenný odkaz) a ***agr*** (shoda). Na tři subkategorie se rozděluje chyba ***vbx*** (složený slovesný tvar). Označují se nadbytečná a chybějící slova (***odd***, ***miss***). Značkuje se chyba ve slovosledu ***wo***. Program se může opřít o automaticky provedenou morfologickou anotaci a lemmatizaci, při rozpoznávání chyb tedy může využívat morfologických značek (např. sloveso v infinitivu) a základních tvarů slov.

4.3.5.2.2.1. Doplnění chyby v reflexivitě rflx

Pokud se v rámci chyby ***dep*** (závislost), ***ref*** (zájmenný odkaz) nebo ***agr*** (shoda) opravuje také reflexivum (*se*, *svůj*), automaticky se k této chybě připojí značka ***rflx***, jako v následujícím příkladu:

R1: *Eva stojí před její dům*
R2: *Eva stojí před svým*$_{dep,rflx}$ *domem*$_{dep}$

4.3.5.2.2.2. Rozdělení chyby ve složeném slovesném tvaru vbx

Chyba ve složeném slovesném tvaru ***vbx*** se automaticky rozděluje na tři podkategorie: chyba v analytickém slovesném tvaru ***cvf***, chyba v konstrukci s modálním nebo fázovým slovesem ***mod*** a chyba ve sponově-jmenném přísudku ***vnp***. Mezi chybami se rozlišuje automaticky na základě lemmat a morfologických značek na rovině R2, popř. i na rovině R1, a to jak u slov přímo propojených hranou, tak u odkazů.

Chyba v konstrukci s modálním nebo fázovým slovesem předpokládá na rovině R2 modální či fázové sloveso (obvykle v odkazu) a sloveso v infinitivu, popř. záměnu jednoho modálního slovesa za druhé či jednoho fázového slovesa za druhé:

R1: *nemůžu už pokračoval dál*
R2: *nemůžu už pokračovat*$_{mod}$ *dál*

Chyba v analytickém slovesném tvaru předpokládá na rovině R2 správný analytický slovesný tvar (popř. jen samotné příčestí, je-li na rovině R1 nadbytečné pomocné sloveso). Za složený slovesný tvar považujeme: préteritum, složené futurum, kondicionál (vč. kondicionálu se spojkami *aby/kdyby*) a opisné pasivum.

R1: *nestačí, aby všechno poznat*
R2: *nestačí, aby všechno poznal*$_{cvf}$

Chyba ve sponově-jmenném přísudku se na rovině R2 nebo R1 týká spojení spony (tvar slovesa *být*) a jména v nominativu nebo instrumentálu (substantiva, adjektiva, popř. i zájmena či číslovky). V žákovských textech je nejčastější chybou tohoto typu elipsa spony:

R1: *přesně nevím co pro mě nejdůležitější*
R2: *přesně nevím, co je*$_{vnp}$ *pro mě nejdůležitější*

4.3.5.2.2.3. Doplnění chyby ve slovosledu wo

Na rovině R2 se manuálně opravuje také slovosled, pokud to vyžaduje gramatická správnost věty (ne styl ani aktuální členění). Anotátor tuto chybu opraví, nemusí ji však označovat, chybovou značku ***wo*** totiž doplní program. Anotace chyby vychází ze srovnání slovosledu na rovinách R1 a R2 a z morfologických značek a lemmat na rovině R2. Chybová značka se přiřazuje závislým větným členům (např. klitikám).

4.3.5.3. Využití systému pro automatické zpracování textů ke kontrole anotace

Systém vyvinutý pro automatickou chybovou anotaci lze také použít ke kontrole anotace (před adjudikací). Do systému byly zahrnuty nástroje, které v průběhu zpracování anotovaného textu vypisují chybějící a pravděpodobně nesprávné chybové značky. Není-li například pro slovní tvar na rovině R1 možné nalézt český základní tvar, nebylo pravděpodobně chybné slovo z roviny R0 vůbec emendováno (nejde-li o vlastní jméno nebo slovo, které se záměrně neemenduje). Jestliže bylo slovo na rovině R0 emendováno na rovině R1 a nebylo přitom chybově anotováno, ověří se rozdíl mezi tvary na R0 a R1; pokud se zjištěná formální chyba projevuje také ve výslovnosti, chybí označení chyby ***incor*** (např. emendovaná, ale neoznačená chyba v kvantitě vokálu). Podobným způsobem se ověřuje i správnost a úplnost chybové anotace na rovině R2. Použitím programu pro tyto účely nelze vyhledat všechny nedostatky v manuální emendaci a chybové anotaci, je však možné rychle posoudit celkovou kvalitu provedené práce a upozornit anotátora na některé jeho chyby.

4.3.5.4. Automatická identifikace ustálených kolokací

Poslední automatickou procedurou zpracování textu na R2 je automatická identifikace a označení frazémů a ustálených spojení slov, které umožní vyhledávání chyb nerodilých mluvčích v použití frazémů, idiomů a ustálených kolokací. Označeny jsou výskyty jak nominálních (neslovesných) frazémů a slovesných frazémů, tak pří-

sloví a přirovnání. Dále uvádíme příklady výskytu frazémů ve zpracovaných textech, rozdělené podle typů.

Neslovesné frazémy:

láska na první pohled; *krok po kroku*; *jak se říká*; *na poslední chvíli*; *od rána do večera*; *s vypětím všech sil*; *psí počasí*; *havraní vlasy*; *plamenná řeč*; *stověžatá Praha*

Slovesné frazémy:

nechat to, jak to je; *dělat si srandu*; *nevěřit svým očím*; *mluvit sám za sebe*; *nemít nic v hlavě*; *udělat první krok*; *mít smysl pro humor*; *dělat si legraci*; *mít za lubem*; *spadnout z nebe*; *vydechnout naposledy*; *dát na pospas*

Přísloví:

všude dobře, doma nejlíp; *bez práce nejsou koláče*

Přirovnání:

být chytrý jako liška; *být silný jako medvěd*; *být jako med*

4.4. Perspektivy

4.4.1. Automatická anotace

Při anotaci se nabízí využít automatických postupů již při aplikaci na chybový text, např. předzpracování textu pro usnadnění úkolu anotátorů, nebo pro plně automatickou anotaci většího objemu textů, kterou z kapacitních důvodů nelze zajistit spolehlivějším manuálním způsobem. Některé pilotní studie v tomto směru už existují. Mezi kandidáty patří automatická morfologická analýza, disambiguace a lemmatizace s využitím více vzájemně odlišných metod, které u chybných tvarů vedou k různým výsledkům. Porovnání těchto výsledků by mohlo vést k automatickému stanovení hypotézy o typu chyby (Díaz-Negrillo et al., 2010). Další možností je využití automatického korektoru k emendaci. Pro chybový i opravený text pak lze uvažovat o automatické syntaktické analýze, která by mohla využívat i některé syntakticky orientované aspekty chybové anotace, jako např. odkazy u chyb ve shodě a v rekci.

Zatím jsme k automatické emendaci zkoušeli využít existující korektor pravopisu (Richter, 2010), který se snaží na základě slovníku, morfologické analýzy a stochastického modelu češtiny, natrénovaného na jazykově správných textech, nahradit chybné slovní tvary takovými, které jsou korektní, co nejméně se liší od původních a zapadají do lokálního kontextu. Tímto způsobem lze opravit i některé tvary, které jsou samy o sobě správné, ale v kontextu negramatické, např. lokálně zjistitelné chyby ve shodě. V provedeném experimentu bylo ze vzorku 9372 tokenů (7995 bez interpunkce) opraveno nezávisle dvěma anotátory stejným způsobem 13 % (celkem 1189 tokenů) na R1 a 16 % (celkem 1 519 tokenů) na R2. Na R1 dosáhl korektor úspěšnosti 72 % (je to hodnota *F-measure* při přesnosti 74 % a pokrytí 71 %, resp. 69 % a 76 % při jiném nastavení parametrů). Úspěšnost na R2 je výrazně nižší (53 %).

Z těchto předběžných výsledků je zřejmé, že plně automatická emendace (a tím i chybová anotace) je zatím myslitelná jen při výrazně snížených nárocích na míru chyb v korpusu. Proto se chceme zaměřit na kombinaci automatické a ruční anotace, kde by automatické metody nabízely anotátorovi nejpravděpodobnější variantu opravy a chybové klasifikace, ale poslední slovo by zůstávalo na anotátorovi.

4.4.2. Korpusový manažer

Pro využívání žákovského korpusu CzeSL, jako je vyhledávání v jeho datech, pořizování statistik a další funkce, je nezbytný *korpusový manažer*. Koncepce korpusového manažeru vychází z anotačního schématu o třech rovinách (R0, R1, R2), na nichž jsou korpusová data v různých podobách uložena. Manažer by měl splňovat tyto základní požadavky:

- umožňovat zpracování dat v třírovinném formátu použitého anotačního editoru *feat*,
- pokrýt plánovaný rozsah korpusu s dostatečnou rezervou pro budoucí rozšíření (cca 5 milionů slov),
- reagovat na dotazy s rozumně dlouhou odezvou při předpokládané zátěži více uživatelů současně (předpokládá se přístup nejvýše dvaceti uživatelů současně).

Manažer by měl uživateli poskytovat dostatečně bohatý repertoár vyhledávacích funkcí vyvolávaných na základě uživatelských dotazů do korpusu. Rovněž by měl umět vytvářet souhrnné informace v podobě filtrů, statistik, a to zvláště v podobě frekvenčních seznamů a kolokačních měr, a tříděných seznamů.

4.4.2.1. Dotazy do korpusu

Korpusový manažer by měl umět v korpusu vyhledávat podle dotazu na formu/lemma/značku v libovolné kombinaci, případně s dotazem na typ chyby a odkaz na libovolné rovině. Vyhledat příslušná data na základě většího množství forem, lemmat či značek mají umožnit regulární výrazy, včetně možnosti dotazovat se na pozici relativně k jiné pozici i na označené frazémy. Dotazy na jednotlivé morfologické kategorie mají zahrnovat i specifikaci identity hodnot u více pozic (například k vyjádření shody). Dotaz může zahrnovat i požadavek zobrazit odpovídající pozice na ostat-

ních rovinách (včetně roviny s přepsaným textem) spolu s volitelným kontextem (viz podrobněji dále). Manažer by měl umožňovat i hledání v neanotovaných datech. U všech dotazů by zároveň měl být k dispozici metadatový filtr. Zadané dotazy by se měly uchovávat v zásobníku dotazů. Manažer by měl také umožňovat export výsledků dotazu ve vhodném formátu a také generování, ukládání a správu subkorpusů.

Dotazy na pozice by měly dále zahrnovat mj. tyto typy:

- dotaz na všechny pozice, které na určité rovině mají/nemají svůj protějšek,
- dotaz na pozice, které korespondují se sousedními rovinami vztahem 1:1:1/m:n:o
  (kde *m, n, o* jsou nezáporná celá čísla),
- dotaz na identitu formy/lemmatu/značky na rovině R0 se značkou na rovině R1 a R2,
- různé statistiky forem, lemmat, značek, typů chyb.

Manažer by měl umět zpracovat i dotazy týkající se změny slovosledu, zejména vyhledat všechny přesunuté pozice s danými vlastnostmi, dále věty, kde k něčemu takovému došlo, a dotčené pozice (uzly na hranách, které přesunutý uzel překřížil svou hranou). Dotazy na typy chyb by měly zahrnovat mimo jiné tyto možnosti: určit počet dotčených pozic na sousedních rovinách; zjistit pozici, na kterou se odkazuje; zjistit výchozí pozici odkazu.

Obecně by korpusový manažer měl být schopen na výstupu zobrazovat korpusové pozice, typy chyb, odkazy, metadata a jejich kombinace. Hlavním typem výstupu by měly být konkordanční řádky obsahující příslušný úsek textu. Na přání uživatele by měl manažer zobrazit roviny podobně jako v anotačním nástroji *feat* s tím, že na konkordančním řádku je implicitně rovina R0 s možností parametrizace a zobrazování jiné roviny. Manažer by měl umět zobrazovat též odlišné formy na jiných rovinách i v lineárním zobrazení na konkordančním řádku a stejně tak chyby, značky, lemmata, a to např. po kliknutí jen na pozici nebo po přejetí myši. Výstupní data by se měla zobrazovat s ohledem na rozlišení běžných a privilegovaných uživatelů korpusu.

4.5. Závěr

Chybová anotace je velmi náročný úkol, ale plody takového úsilí mohou být velmi užitečné. Uživatel korpusu s chybovou anotací má přístup ke statistickým údajům o typech chyb, které nelze získat jiným způsobem a které podávají věrný obraz mezijazyka studentů. To umožňuje modifikovat pedagogické metody a materiály používané při výuce tak, aby řešily nejčastější slabiny v jazykových dovednostech studentů s ohledem na úroveň jejich znalostí a mateřštinu.

Anotace přináší řadu podnětů, které se promítají do anotačního manuálu a školicích setkání. Důležitým nástrojem pro zdokonalování popisu chybové taxonomie i vlastního anotačního schématu je také internetové fórum, které slouží k řešení aktuálních problémů anotátorů. Reakce anotátorů již umožnily alespoň částečně zpřesnit pokyny k rozhodování v některých obtížnějších případech, např. při nejistotě o intenci autora, inferenčních chybách, o optimální míře intervence do původního textu a o způsobu anotace nestandardních variet jazyka. Ve všech těchto případech je třeba skloubit požadavky potenciálních uživatelů korpusu s imperativem konzistentní anotace.

5. Jazyková chyba a práce s ní v jazykovém vyučování

Milan Hrdlička

Práce s jazykovou chybou je podstatnou součástí jazykového vyučování, jazyková chyba byla tedy předmětem zájmu lingvistů a jazykových lingvodidaktiků odedávna. V úvodních pasážích této kapitoly pojednáváme o jazykové chybě a jazykové správnosti z perspektivy teorie spisovného jazyka. Jelikož je daná problematika značně složitá, diskusní a v mnoha aspektech problémová, je jí vymezeno relativně více prostoru. Druhá část kapitoly je věnována problematice chyby z hlediska lingvodidaktického. V závěru kapitoly je prezentována jedna z možností elementární klasifikace morfologických tvarů, která se úzce vztahuje k psaným i mluveným projevům (nejen) jinojazyčných mluvčích komunikujících na základě různé kvality komunikační kompetence naší mateřštinou.

5.1. Jazyková chyba a jazyková správnost

Jen málokterý druh lidské činnosti je tak úzce, nerozlučně a nevyhnutelně spjat s chybováním, jako učení se cizímu (ale i mateřskému) jazyku.

„Chyba“,[114] zde zatím pracovně vymezená především jako určitý nežádoucí odklon od jazykové normy (v jistých případech rovněž od řečové normy), bývá obvykle jedním z prvků, které s vysokou mírou spolehlivosti odlišují (pochopitelně, že při užívání jednoho a téhož jazykového kódu) mluvčího rodilého od mluvčího jinoja-

[114] Vzhledem k lingvodidaktickému zaměření příspěvku operujeme přednostně s pojmenováním „chyba“, byť jsme si vědomi jeho určité problémové podstaty. Uvědomíme-li si totiž existenci pestrého spektra různých jazykových a řečových prohřešků i různých příčin pochybení účastníků komunikace, můžeme oprávněně uvažovat i o jiném pojmoslovném aparátu: odchylka, omyl, přeslechnutí se, přeřeknutí se, nedopatření, nedorozumění, neporozumění aj.
Pokud jde o přírodní vědy, v souvislosti s chybou se někdy hovoří o jakémsi „zklamaném očekávání“, neboť se za chybu pokládá odchylka naměřené hodnoty od hodnoty předpokládané (viz Korčáková, 2004).

zyčného.[115] Hausenblas (1979) v této souvislosti poukazuje na zajímavý poznatek z psychologie, a sice že u člověka vyvolávají negativní jevy zpravidla větší pozornost než jevy pozitivní. Tento fenomén je ostatně průkazně empiricky ověřený i ve sféře řečové komunikace prostřednictvím přirozeného jazyka.[116]

Na základě zkušeností a poznatků z řečové i vyučovací praxe se někdy připouští, že různé národy projevují k chybám ze strany jinojazyčných mluvčích nestejnou míru tolerance a pochopení. Zdá se (přesná zjištění v tomto směru nemáme k dispozici), že Češi (alespoň starší generace) jsou v tomto směru, např. na rozdíl od mnohých obvykle tolerantních anglických mluvčích, poměrně dosti nároční a přísní.

Hovoříme-li o problematice chybování ve sféře řečové komunikace,[117] považujeme za nutné zmínit již na tomto místě jednu podstatnou, avšak nikoliv samozřejmou a vždy náležitě doceňovanou skutečnost: jeden z klíčových momentů představuje detekce chyby, tedy odhalení jazykového prostředku (řečové formy), který není v souladu s určitými konvencemi, ustálenými zvyklostmi (čili s územ, stručně řečeno s běžným výskytem určitého jazykového jevu – ať „správným" či „nesprávným" – v komunikaci)[118] či s explicitně deklarovanými (kodifikovanými), závaznými[119] a vyžadovanými pravidly, normami.[120] K tomu je ovšem mimo jiné nezbytná náležitá

[115] Upozorňujeme na známou skutečnost, že se různých jazykových (řečových) přehmatů poměrně běžně dopouštějí ve své mateřštině rovněž rodilí mluvčí, a to nikoliv pouze v dětském věku (srov. potíže s velmi komplikovaným náležitým pravopisem, s ortoepií, s morfologickými formami prestižní variety českého národního jazyka aj.).

[116] Hausenblas (Hausenblas, Kuchař et al., 1979, s. 169) komentuje situaci slovy: *„Veřejnost, běžní uživatelé jazyka mívají vyhraněnější představy o jazykových chybách než o jazykové správnosti, existují také jednoznačnější názory na to, co je chyba, než na to, co je správné."*

[117] Stranou ponecháváme zajímavou problematiku funkčního využití záměrných pravopisných nebo gramatických chyb v uměleckém textu, a to jak v původním, tak v přeloženém. Srov. blíže Hrdlička, 1995.

[118] To, co je ve spisovné normě, nemusí být v úzu (např. přechodníky, kondicionál minulý, genitiv záporový aj.) a naopak. Encyklopedický slovník češtiny (2002, s. 516) vymezuje úzus slovy: *„Soubor jaz. prostředků, které jsou ve vžité podobě užívány jaz. společenstvím (bez ohledu na to, zda jsou vhodné, n. nevhodné, správné, n. nesprávné). Ú. se někdy rozumí i běžné užívání jaz. prostředků. V teorii spis. jazyka je ú. vysoce hodnocen; i ty prostředky, které jsou utvořeny nesystémově, lingvisté akceptují, jestliže jsou přijaty jaz. společenstvím. Např. slovo šlehačka bylo původně odmítáno jako neústrojně (nesystémově) utvořené: příponou -čka se od sloves tvoří pojmenování přístrojů (řezat – řezačka), produkty činnosti se tvoří příponou -nka (řezat – řezanka). Protože se však slovo šlehačka ve významu ‚produkt šlehání' vžilo, bylo akceptováno jako noremní."*

[119] Jejich nedodržení, porušení bývá často různým způsobem sankcionováno.

[120] Pojetí normy není jednotné (srov. např. Hrbáček, 1994, Nebeská, 1996 aj.) a její definice je značně komplikovaná. Starší, preskriptivní pojetí spatřovalo v normě předpis, návod k náležitému užívání jazyka. Novější pojetí (takovým způsobem ostatně přistupovali k normě i protagonisté Pražské školy) chápalo normu jako (vnitřní) systémovou zákonitost, jako objektivní vlastnost jazyka, kterou je možné pouze zjišťovat (srov. výstižný

znalost jazykového systému (a potažmo i zásad řečového chování),[121] v našem případě kodifikace normy spisovného jazyka (upozorňujeme však, že ne každá odchylka musí být nutně pokládána za chybu,[122] srov. i dále) i řečové etikety příslušného jazykového společenství.

Rozsáhlá a složitá problematika jazykové chyby je logicky těsně spjata s komplexní a ne vždy snadno uchopitelnou otázkou jazykové správnosti. Názory na ni se vyznačují – a to nejen v české jazykové situaci – rovněž velmi zajímavou historií a genezí. Jazyková správnost byla v minulosti kupř. vztahována k čistotě jazyka, k jeho historické přímočarosti a pravidelnosti, k respektování kodifikace spisovné normy apod.; Mathesius (1932) nadřazuje pojmu jazyková správnost jazykovou vytříbenost, tj. míru, do níž je spisovný jazyk propracován jako prostředek literární, vědecké aj. komunikace (zdůrazňuje rovněž skutečnost, že podmínkou kultivovanosti není vždy nutně spisovnost).[123]

5.1.1. Jazyková chyba v užším smyslu

Pokud jde o tuzemskou lingvistickobohemistickou tradici, bylo by možné konstatovat, že se v pojetí jazykové chyby vymezují v zásadě dva základní přístupy. Uvažuje se totiž o chybě v užším a o chybě v širším pojetí (Jelínek, 2001). To má své podstatné důsledky, a to zdaleka ne jen ve sféře lingvodidaktické.

V našem dalším výkladu se nejprve zastavíme u prvního ze zmíněných přístupů, na který posléze navážeme pojednáním o diskusi spjaté s široce chápanými otázkami jazykové kultury v našem prostředí.

Máme-li na mysli chybu v užším pojetí, dostáváme se na obecné rovině k široce pojímané a bohatě strukturované problematice odchylky od standardu, a to na všech

Chloupkův výrok *„Norma je současný stav jazykové struktury"*, uvedeno podle Hrbáček, 1994, s. 76). V současnosti se někdy situuje norma mezi systém jazyka (langue) a mezi jeho užití (parole), poukazuje se na společensky závaznou (konvencionalizovanou) realizaci jazykového systému – na normu se pohlíží jako na systém závazných realizací jazyka, které jsou přijaty příslušným jazykovým společenstvím.

[121] Ve sféře výuky češtiny pro cizince zaznamenáváme první pokusy o zmapování této významné oblasti, viz blíže např. Bischofová, Hrdlička (2007).

[122] Za chybu nejsou považovány různé aktualizace (srov. již Mukařovský, 1932), tedy nevšední umělecké použití jazykových prostředků. Hausenblas (in Hausenblas, Kuchař et al., 1979 aj.) např. v tomto smyslu rozebírá Holanovo poetické vyjádření *Mlčím vás, jabloně!* v básni *Smrt umírajícího v sadě* ze sbírky *Vanutí*.

[123] V této souvislosti podotýkáme, že užívání kupř. obecné češtiny není možné a priori hodnotit jako nekultivované a „pokleslé": obecná čeština (a potažmo běžná mluva) má v neformální komunikaci své nezastupitelné místo, funkčně a záměrně se jí však využívá rovněž v oblasti kultury, umění (viz kupř. literární texty Haška, Wericha, Horníčka, Suchého, Hrabala a dalších autorů).

jazykových rovinách a plánech. V této souvislosti si můžeme položit otázku, kde hledat a v čem spatřovat hlavní příčiny oněch odchylek od systému (normy) prestižní variety českého národního jazyka.

Mezi nejčastější příčiny chyb[124] v užším pojetí podle Jelínka (2001, s. 79n.) patří následující jevy:

a) Pronikání obecněčeských a nářečních prvků do spisovného jazyka. Tento důvod, podle našeho mínění jednoznačně nejčastější a nejvýraznější, je obecně přijímaný, není proto (v jednom i druhém případě) o četné a různorodé příklady nouze.

Ponecháme-li stranou ostatní jazykové roviny a plány, jako výmluvný doklad stačí připomenout běžné a frekventované deklinační různotvary, např. obecněčeské pádové koncovky jména (*s těma třema velkejma klukama*), nebo oblast formálního tvarosloví českých sloves (*oni dělaj* apod.).

b) Dalším evidentním důvodem je nevhodná volba (záměna) lexikálních prvků slangových (zde a částečně i na dalších místech použijeme Jelínkův (2001, s. 79) příklad) – *volačka do Brna* místo *předčíslí*.

c) Nelze zapomínat ani na neznalost náležitého užití knižního výraziva: i v této oblasti jsme svědky častého užití tvarů, které se neslučují se synchronní kodifikací spisovné normy (viz např. mnohdy chybné užití tvarů zájmen *týž, jenž*), zařadili bychom sem rovněž případy tzv. hyperkorektnosti, čili přehnanou snahu o užití prestižní variety českého jazyka bez náležité znalosti standardu (*před dvěmi lety* aj.).

d) Dalším důvodem je nedodržování spisovného systému slovotvorby při doplňování slovní zásoby spisovné češtiny. Jelínek (2001, s. 79) v této souvislosti uvádí následující příklady neústrojně utvořených slov, která však – což není ojedinělý případ – přesto zakotvila v úzu[125] a stala se součástí standardu: *velín; majitel; ovlivnit*.

e) K chybování mohou přispívat i různé kontaminační procesy, ať už českých prepozic (např. *Sejdeme se před, nebo po obědě?;* chybné užívání prepozice *mimo* s genitivem, nikoliv s akuzativem pod vlivem analogické prepozice *kromě + G* apod.), nebo spojek *(Pokud se týká zítřejšího počasí; Když bych* ev. *Jestli bych/bysem přišel, přinesl bych/bysem to*), dále nerozlišování slovesných vazeb typu *pamatovat si něco* × *pamatovat se na něco*) apod.

[124] Jelínek (2001, s. 80) však přináší zajímavý protichůdný příklad, a to pospisovnění původní chyby: „*Např. až do roku 1993 se za nespisovný považoval tvar infinitivu péct, jehož koncové -t bylo k základu slovesa přidáno podle infinitivu jiných slovesných typů. Pravidla českého pravopisu z r. 1993 přiznala této formě status prostředku spisovného, i když se v komentářích k Pravidlům upozorňovalo na stylový příznak hovorový. Nebo uzuální přestavba větného vzorce je slyšet hudbu v podobu je slyšet hudba si poměrně nedávno vynutila vřazení nové syntaktické varianty do kodifikované spisovné normy.*"

[125] Podobných případů lze najít celou řadu. Jedním z nich je i nevhodná počeštěná přejímka z angličtiny ***sexual harassement***, užívaná u nás jako ***sexuální harašení*** (vedle vhodnějšího *sexuálního obtěžování*). Lexém ***harašit, harašení*** se v češtině tradičně pojil s výrazem „zbraněmi".

Doplnit bychom mohli ještě pestrou škálu nejrůznějších dalších stylizačních a formulačních nedostatků a neobratností ze strany českých rodilých mluvčích.

Otázkou např. zůstává, kdy se stanou spisovnými stále častější případy doposud převážně kritizovaného a odmítaného souvýskytu dvou prepozic, jevu z oblasti jazykové ekonomie, k němuž (zejména v mluvené komunikaci) dochází v zájmu dosažení větší kondenzace sdělení a který patří mezi nejvýraznější příklady vývojových tendencí současné češtiny (*Trenér spoléhá na v zámoří hrající hráče; Koupili to za pro ně nevýhodných podmínek*, viz blíže Machová, 2000).

f) Jelínek (2001) zmiňuje rovněž rozmanitý vliv cizích jazyků.[126] V této souvislosti můžeme poukázat na jeden z nápadných trendů současné češtiny, a sice na předsouvání (antepozici) vlastního jména před jméno obecné: *fotbalová Gambrinus liga; Tipsport extraliga ledního hokeje* atd.

Je zajímavé, že o tomto fenoménu hovoří Kučera (1990) jako o jevu příznačném pro americkou češtinu (viz spojení typu *v Rio Grande údolí*), v současnosti je však již běžný i v dnešní češtině. V určitých komunikačních kontextech (reklama, prodej zahraničních výrobků aj.) dochází pod vlivem cizojazyčných prvků k procesu tzv. deflektivizace, tedy k potlačování pádových koncovek a k užívání jmen cizího původu v nominativu (*akční nabídka Samsung; Allianz rodinný běh* aj.).

5.1.2. Jazyková chyba v širším smyslu

Za jazykovou chybu v širším slova smyslu se pokládá odchylka od systému národního jazyka vůbec (Jelínek, 2001, s. 80), nikoliv tedy pouze odklon od normy spisovné češtiny. Těchto různě závažných jazykových, resp. řečových provinění se dopouštějí cizinci, kteří se učí česky a kteří z důvodu nedostatečné znalosti češtiny porušují v psaném i v mluveném projevu její normu na všech úrovních jazykového systému (o příčinách vzniku tohoto druhu chyb viz oddíl 5.4.2).[127]

5.2. Odchylka od normy: puristé versus funkcionalisté[128]

Otázka jazykové chyby se úzce váže k dvěma důležitým, odlišným (nebo spíše protikladným) myšlenkovým proudům v lingvistické bohemistice, a sice purismu a vy-

[126] K dané problematice viz příspěvky např. Adama, Mareše aj.

[127] Odhlížíme zde od anomálních případů jako agramatismus aj.

[128] K uznávanému funkčnímu strukturalismu, resp. spíše k neuspokojivým výsledkům a důsledkům jeho aplikace, se staví kriticky Starý (1995, s. 18, resp. s. 30), když poukazuje na stávající – a od doby Pražské školy stále poměrně neutěšenou – jazykovou situaci u nás:

braným názorům předních představitelů pražského funkčního strukturalismu. Obě tyto antagonisticky laděné koncepce pokládáme za významné pro pochopení podstaty řešené problematiky (jeví se jako klíč k postižení polemik o povaze a roli variet českého národního jazyka v komunikaci), neboť stály – každá pochopitelně jiným způsobem – u zrodu teorie spisovného jazyka (a potažmo teorie jazykové kultury), která dlouhodobě sehrávala a stále ještě sehrává v otázce jazykové chyby a jazykové správnosti klíčovou roli.

5.2.1. Český purismus – jeho podstata a druhy

Puristické tendence se v našich podmínkách projevovaly po několik staletí velmi významnou měrou, byť byly zaměřeny poměrně velmi úzce a monotematicky: soustřeďovaly se převážně na oblast slovní zásoby.

Purismus, resp. brusičství,[129] prošel v českých zemích dlouhým vývojem. Geneze snah o různě výrazné a intenzivní „očištění" českého národního jazyka a o jeho zbavení domněle či skutečně cizojazyčných prvků a vlivů, především německých, se datuje už od doby Husovy. Vznik puristických tendencí v zemích Koruny české se vysvětluje naší geopolitickou polohou a značnými německými vlivy na český jazyk (připomínáme soupeření češtiny s němčinou ve středověku,[130] pohnuté osudy našeho národního jazyka v době pobělohorské,[131] emancipační úsilí buditelů v době národního obrození, snahy vlastenců v boji o samostatný český národ a stát ve druhé polovině 19. století aj.). V tomto smyslu se pak někdy hovoří o purismu obranném, kdy jde o snahu vymanit se z germánského vlivu[132] (především pokud jde o germanismy ve slovní zásobě češtiny), a o purismu historickém – ten spočívá v nezřídka dogmatické snaze o zachování přímočaré a pravidelné historické linie vývoje jazyka.

„Mezi živou jazykovou praxí a standardní češtinou totiž po bezmála půlstoletí funkční kultivace panuje citelné napětí."/…/„V tomto smyslu funkcionalismus nesplnil očekávání v něj kladená a jako teoretický princip řešení otázek jazykové praxe selhal. V tomto smyslu selhala i teorie jazykové kultury, jejímž jádrem funkční princip je." Na Starého skepsi a zpochybňování teorie jazykové kultury s odstupem navazuje Cvrček (2008) s tzv. Konceptem minimální intervence.

[129] Název patrně vzniká podle názvu jazykové příručky *Lima linguae Bohemicae* (Brus jazyka českého, autor Constantius) z roku 1667.

[130] Srov. např. silně vlastenecky (a protiněmecky) laděnou Dalimilovu kroniku z počátku 14. století.

[131] Proti pojímání období baroka jako „doby temna" se však staví např. Stich, 1969, 1979, Starý, 1995 (viz jeho „syndrom národního údělu") aj.

[132] Situace je ovšem složitější: M. Jelínek (2007, s. 542) upozorňuje, že u nás na podzim roku 1938 po mnichovském diktátu a zradě Francie proběhla silná kampaň proti galicismům. Jindy, a to i v závislosti na dané politické situaci, určitá část bohemistické obce brojí proti anglicismům, rusismům apod.

5.3. Meziválečná a poválečná diskuse o jazykové kultuře

V souvislosti s postupným utvářením teorie spisovného jazyka pražskými funkčními strukturalisty a v důsledku vleklých a nezřídka bezvýchodných a stereotypních polemik týkajících se základních otázek jazykové kultury, se vytvářelo několik vzájemně provázaných klíčových tematických okruhů, o nichž se – i kvůli zásadové neústupnosti účastníků diskuse a neslučitelnosti jejich argumentů – bez hmatatelnějšího názorového posunu a bez vyhlídky na alespoň částečné dosažení konsensu diskutuje dodnes. Mezi ona bázová, úzce související témata[133] podle našeho soudu patří a) role úzu při kodifikaci standardu, resp. oprávněná míra zasahování do kodifikace spisovné češtiny (preskripce versus deskripce) b) role obecné češtiny v komunikaci, c) otázka pojetí, resp. (ne)existence hovorové češtiny. U zmíněných problémových bodů se nyní stručně zastavíme.

5.3.1. Role úzu při kodifikaci standardu

Daná problematika je jedním ze stěžejních, nejčastěji diskutovaných a současně nejkonfliktnějších bodů. Můžeme se zde setkat s různě početně a argumentačně zastoupenou pestrou paletou názorů, které se nacházejí na názorové ose mezi dvěma bipolárními variantami: (velmi) omezená (a kontrolovaná) role úzu versus výlučná (nekriticky přijímaná) a (prakticky) neregulovaná role úzu.

Není bez zajímavosti, že na tuto otázku nepanovala shoda ani v rámci Pražské školy. Zatímco „anglistická složka", zastoupená především Mathesiem a Trnkou, se přikláněla k názoru, že rozhodující vliv na spisovnou normu má mít úzus (viz Starý, 1995, Cvrček, 2006), Havránek, představitel „slovanské" části, hájil stanovisko opačné – řadil se ke stoupencům kvalifikovaných zásahů, regulace, vědecky podloženého pěstění (kultivování) spisovného jazyka. Ve sborníku Úkoly spisovného jazyka a jeho kultura se k tomu vyjadřuje slovy: *„usus sám nevytváří normu spisovného jazyka: vytváří se, totiž vzniká a dále se vyvíjí, z různých tendencí za teoretických zásahů a tím se liší od normy lidového jazyka. Tedy i teorie jazykovědná zasahovala a může zasahovati do vývoje spisovného jazyka."* (Havránek, 1932, s. 32)

[133] Z prostorových důvodů ponecháváme stranou otázku komunikačních funkcí spisovné a obecné češtiny, problematiku vymezení a komunikačního uplatnění běžné mluvy, otázku existence diglosie v české jazykové situaci a střídání, resp. míšení jazykových kódů aj.

Tato Havránkova[134] koncepce nakonec v PLK převážila a v následujících obdobích na ni navázala řada význačných bohemistů (Jedlička, Daneš[135] aj.). Dlužno ovšem podotknout, že je tato otázka i nadále otevřená a v našem prostředí uspokojivě nedořešená. Lze konstatovat, že zejména zásluhou korpusové lingvistiky nabývá v posledním desetiletí role úzu na významu (viz práce Čermákovy, Cvrčkovy aj.).

5.3.2. Obecná čeština – její územní rozšíření a role v komunikaci

Je známou skutečností, že PLK zastával v otázce spisovného jazyka výrazně elitářské postoje. Na spisovný jazyk bylo pohlíženo jako na nejpropracovanější, funkčně a stylově diferencovaný a vědomě kultivovaný prestižní útvar národního jazyka, jako zdroj poznání normy spisovné češtiny mělo sloužit jazykové povědomí intelektuálních vrstev i literární praxe posledních 50 let (srov. i Ertlovu koncepci „dobrého autora").

V padesátých letech se v našem prostředí prosazuje teze o demokratizaci jazyka (užívání prestižní variety přestává být záležitostí úzké elity národa a cílem se stává naopak její celonárodní dostupnost) a ke slovu se dostávají i koncepce třídního charakteru jazyka. V podobném klimatu se mohly dostávat do popředí i různé extrémní názorové a populistické proudy. Cvrček (2006, s. 20, 21) k tomu uvádí: „*V souvislosti s třídním pojetím a s demokratizačními tendencemi se tedy objevil názor, že spisovný jazyk je třídním jazykem buržoazie (tedy nevhodným pro novou společenskou situaci). To se pravděpodobně stalo v rámci oficiální diskuse o jazykovědě, kterou uspořádal tehdejší ministr školství Z. Nejedlý 24. 11. 1949, kde tehdy mladý lingvista J. Krejčí položil otázku o zrovnoprávnění obecné češtiny se spisovnou.*"

Je třeba zdůraznit, že tyto dobové snahy byly odmítnuty jako pokus „zrušit spisovnou češtinu" (Havránek, Dokulil), podobné tendence se však nicméně objevily

[134] Starý (1995, s. 114) situaci vystihuje slovy: „*Pražská škola a puristé tedy sdíleli základní modalitu své činnosti kolem češtiny – jejich přístup k jazykové praxi byl bytostně intervenční. Z tohoto hlediska vystoupení teoretiků jazykové kultury proti puristům na přelomu dvacátých let v léta třicátá tedy nebylo vystoupením proti intervencím puristů, ale proti purismu těchto intervencí.*"

[135] Např. Daneš (1979) stanovuje tři objektivně zdůvodněná kritéria: a) kritérium noremnosti (noremní se tu jeví ten jazykový prostředek, který je daným jazykovým společenstvím přijat, je tedy konvencionalizován); b) hledisko adekvátnosti (zjišťuje se, zda, resp. do jaké míry je jazyková forma schopna splňovat funkční potřeby daného jazykového společenství, jde tedy o vhodnost, přiměřenost); c) hledisko systémovosti (jde o to, zda, event. jak je příslušný jazykový prvek v souhlase s již existujícími vztahy, pravidly jazykového systému, zda, event. jakým způsobem přispívá k vnitřní soudržnosti, pravidelnosti, dynamické rovnováze systému). Podotýkáme, že tato kritéria mají předchůdce v Trávníčkově ústrojnosti (rozumí se jí systémovost), úkonnosti (jde o funkčnost) a vžitosti (lze ji chápat jako noremnost, resp. uzuálnost).

i při jiných příležitostech, např. v roce 1950 na FF UK při diskusi o reformě pravopisu (viz blíže Cvrček, 2006).

Určitá menšinová část bohemistů (a to i zahraničních) přistupovala k roli obecné češtiny v komunikaci nekonformním způsobem. Někteří se pokusili prosadit ji do pozice alternativní variety k češtině spisovné (nesporným pozitivem těchto snah bylo obrácení pozornosti bohemistické obce i k jiné než spisovné varietě národního jazyka, srov. Sgall (1960) i následující diskusi na stránkách *Slova a slovesnosti*, *Naší řeči*, *Českého jazyka a literatury*, řady sborníků apod.

K stoupencům teze o celonárodní platnosti obecné češtiny patřili Sgall, Hronek aj. (viz Cvrček, 2006 aj.); jiní lingvisté ji pokládali a pokládají dosud za útvar regionálně omezený.[136] Sgall (1996) později vystoupil s návrhem zavedení úrovní spisovnosti,[137] čímž by se i obecná čeština dostala do pozice standardu, tedy prestižního útvaru českého národního jazyka.

Na regionální omezenost užívání obecné češtiny poukazuje např. Uličný (1996, s. 61), který v této souvislosti konstatuje: „*Především se tzv. obecná čeština nešíří na východ Moravy; nemá tu dostatečné psychosociální podmínky, protože vždy byla vnímána jako cizí a neprestižní jev. Tato situace se dnes díky televizi poněkud změnila, avšak ve východomoravských interdialektech se to projevilo jen velmi málo.*"

V souvislosti s důvody akceptování určité variety za celonárodní prestižní útvar se zdůrazňuje právě požadavek její „regionální přijatelnosti" (Uličný, 1994). Jelínek (1963) nadto dovozuje, že obecná čeština nemůže být přímým zdrojem změn v normě spisovného jazyka, ale že se tyto změny mohou uskutečňovat jen prostřednictvím hovorového jazyka jako celonárodní mluvené varianty spisovné češtiny.

V odborných kruzích se v souvislosti s nadhodnocováním role obecné češtiny (a s jejím „zrovnoprávněním" s češtinou spisovnou) rovněž opakovaně upozorňuje na nebezpečí stylové nivelizace spisovných a nespisovných prostředků, které by mohlo vést až k závažnému narušení funkční stylové variability naší mateřštiny. Můžeme konstatovat, že v posledním období dochází v pohledu na roli češtiny spisovné a obecné k určitému posunu. Začíná se klást větší důraz na potřebu terénních materiálových výzkumů nezatížených předpojatostí.

[136] Dokládáme jedním citátem za všechny (Cvrček, 2006, s. 47): „*Odlišnosti můžeme vysledovat i v náhledu na obecnou češtinu. Zatímco Sgall v ní vidí celonárodní útvar, který je v běžném hovoru prestižní (a jediné, co brání uznání spisovnosti prvků, je jejich nepřítomnost v kodifikaci), Bělič ji naopak chápe jako útvar nespisovný, regionálně omezený, který se navíc projevuje v hybridním kombinování jevů obecné češtiny a spisovných.*" (Poznamenáváme: v tomto pojetí by se dnes již uvažovalo o běžné mluvě; pokládáme obecnou češtinu za útvar strukturní, nikoliv však celonárodní).

[137] Mohlo by s trochou nadsázky dojít k situaci, kdy by bylo v zásadě vše spisovné, pouze s tím rozdílem, že v různé míře (čili něco „více", něco „méně").

5.3.3. Čeština spisovná a obecná ve výuce jinojazyčných mluvčích

Je možné konstatovat, že v oblasti výuky češtiny jako cizího jazyka převládá názor, že je – v obecné rovině – vhodné prezentovat jinojazyčnému mluvčímu přednostně češtinu spisovnou.[138] Argumentuje se zpravidla skutečností, že má prestižní varieta naší mateřštiny celonárodní platnost (je regionálně nepříznaková) a že má její užívání v určitých oblastech (oficiální styk, hromadné sdělovací prostředky, odborná sféra) své pevné místo.

Poukazuje se také na poznatek, že se od nerodilého mluvčího obecně očekává, resp. toleruje se jako přiměřené jeho vyjadřování spisovné (včetně náležité ortoepie) než kolokviální. Z hlediska úspěšnosti komunikace se totiž považuje za nepatřičné, užívá-li jinojazyčný mluvčí substandardní kód, aniž češtinu ovládá na náležité úrovni – jeho jazykový projev se tak může setkat s negativní odezvou.

V souvislosti s lingvodidaktickou prezentací prestižní variety našeho národního jazyka je možné v průběhu posledních desetiletí sledovat postupný přechod od češtiny spisovné neutrální (až knižní) k češtině hovorové; můžeme však zaznamenat i pokusy o přednostní zařazení češtiny obecné (viz blíže Hrdlička, 2009, 2010). Postoje zahraničních bohemistů k roli a zastoupení zmiňovaných útvarů českého národního jazyka jsou různé, s určitým zjednodušením lze konstatovat, že s rostoucí úrovní komunikační kompetence mluvčího v češtině stoupá jeho zájem i o některé útvary nespisovné.

V zájmu náležité prezentace češtiny jako cílového jazyka pokládáme za žádoucí, aby byl cizinec poučen již v počátcích studia češtiny (zejména pohybuje-li se v česky mluvícím prostředí) alespoň v hrubých rysech o české jazykové situaci, o roli jednotlivých variet v komunikaci a o jejich teritoriálním zastoupení. Pokud se v učebním materiálu češtiny pro cizince prezentuje obecná čeština jako nejrozšířenější nespisovný útvar, mělo by se tak dít kvalifikovaně a uváženě. To znamená, že by měla být obecná čeština popsána komplexně (všechny jazykové roviny a plány), nikoliv pouze dílčím způsobem (pozornost bývá věnována pouze stránce morfologické) a odděleně od češtiny spisovné, resp. signalizovaně.

Živelný průnik češtiny spisovné a obecné, jak k němu v některých učebnicích češtiny pro cizince dochází, komplikuje cizinci interpretaci tvarů spisovných a nespisovných, narušuje jeho schopnost vědomě a účelně přepínat kódy a může vést k nediferencovanému hybridnímu vyjadřování, což je z komunikačního hlediska nežádoucí.

[138] Záleží pochopitelně na celé řadě faktorů, mimo jiné na cizincových komunikačních potřebách a prioritách, na úrovni jeho komunikační kompetence v češtině aj. Podotýkáme rovněž, že se v popisech češtiny jako cizího jazyka podle SERR až do úrovně B1 včetně nepočítá se zařazováním nespisovného výraziva.

Domníváme se, a to nejen z perspektivy lingvodidaktické prezentace českého jazyka jako jazyka cizího, že stávající komplikované a v zásadě patové situaci nepřispívá chápání dané problematiky jako konfrontační, tedy jako soupeření češtiny spisovné a obecné. Vhodnější a prospěšnější by byl podle našeho soudu přístup, který by na souvýskyt oněch útvarů národního jazyka pohlížel jako na jejich funkční a účelnou koexistenci (viz blíže Hrdlička, 2009).

5.3.4. Hovorová čeština, resp. otázka její (ne)existence

Nezřídka diametrální odlišnost postojů a rozpolcenost diskusí se názorně projevuje v otázce hovorové češtiny, a to nejen co se týká obsahu zmíněného pojmu, ale i oprávněnosti samotné jeho existence.

Zatímco nemálo tuzemských i zahraničních bohemistů existenci této vrstvy standardu popírá (Mathesius, 1932, Čermák, 1996 aj., Sgall a Hronek, 1992, Starý, 1995 aj.) a kvůli nekompaktnímu („mezerovitému") charakteru se o ní jako o chiméře vyslovuje i Daneš (1988), jiní lingvisté s tímto pojmem i nadále operují a existenci hovorové češtiny uznávají.

O konstituování pojmu hovorová čeština se přičinil především Havránek (1932), když uvažoval o hovorovém funkčním jazyku. Jako jeden z prvních se o vymezení tohoto pojmu pokusil Kopečný (1949), který v hovorové češtině spatřuje mluvenou formu spisovné češtiny[139] a současně se domnívá, že jako podklad (pozadí) pro její formování slouží čeština obecná. Tento názor je ostatně v souladu s rozšířenou představou, podle níž tvoří hovorová čeština „nárazníkové pásmo" s oblastí nespisovnou, vystupuje jako určité síto, jehož prostřednictvím se do prestižní variety postupně dostávají vybrané frekventované prvky nespisovné.

Této myšlence oponuje Bělič. Pro něho je doba pobělohorská (období barokní češtiny) obdobím úpadku,[140] kdy se spisovná norma *„rozkládá nesystematickým pronikáním četných prvků dialektických."* (Bělič, 1950, s. 10) Hovorová čeština má podle Běličova názoru prostor pro svůj rozvoj až ve druhé polovině 19. století, kdy český jazyk postupně proniká do veřejného života. Bělič (1955, 1958, 1959) charakterizuje hovorovou češtinu jako mluvnou formu češtiny spisovné, ovšem bez knižních prostředků na straně jedné a nářečních jevů na straně druhé, V tomto smyslu je hovorová čeština: *„nevtíravě správná řeč, zbavená na jedné straně výlučných znaků knižního jazyka, zachovávaných někdy v oficiálních promluvách veřejných, na druhé straně však*

[139] Hrbáček (1995, s. 53) správně uvádí, že *„Mluvený projev nemusí být hovorový a naopak hovorový projev může být i psaný."*

[140] U Běliče se opakovaně (viz Bělič, Havránek, Jedlička, Trávníček, 1961; Bělič, Havránek, Jedlička, 1962) setkáváme s názorem, že hovorová čeština existovala již před třicetiletou válkou, ale v době úpadku, zvláště vyháněním nekatolíků a germanizací šlechty postupně mizely ty vrstvy obyvatelstva, které jí užívaly v běžném denním styku.

neobsahující ani jevy nářeční, které jsou i v této volnější „neoficiální" podobě spisovného jazyka pociťovány jako nespisovné." (Bělič, 1959, s. 437)

Bělič usiluje o posilování mluvnosti spisovné češtiny, kritizuje jak zaostávání normativních příruček za skutečným vývojem jazyka, tak neblahý vliv vzdělávací soustavy, kde se učí: „*spisovné češtině více či méně archaické, strojené, neživotné; a to jen podporuje a dále udržuje situaci /.../ že pro značnou část příslušníků národa je spisovný jazyk běžným dorozumívacím prostředkem zpravidla jen v projevech písemných /.../, kdežto v ústních projevech nanejvýš jenom oficiálních.*" (Bělič, 1959, s. 436)

Sgall (1960) vystupuje proti chápání hovorové češtiny jako (strukturního) útvaru, neboť usuzuje, že např. její gramatická struktura nemá ve srovnání s češtinou spisovnou a obecnou žádné viditelné rozlišovací rysy. S tímto postojem lze ovšem polemizovat. Kupř. už Hausenblas (1962) poukazuje na skutečnost, že se do oné „požadované" míry mezi sebou neliší ani dialekty a spisovný jazyk. Nadto je dnes možné uvést oněch distinktivních mluvnických rysů celou řadu, srov. např. morfologické spisovné (neutrální) a hovorové tvary jako *mohu – můžu, děkuji – děkuju, oblékl – obléknul, oni rozumějí – oni rozumí, co jsi dělal? – cos dělal?, bez tří, čtyř – bez třech, čtyřech, vidí mou knihu – vidí moji knihu, zná se s mou sestrou – zná se s mojí sestrou, je u tvé kolegyně – je u tvojí kolegyně, čeká na něho – čeká na něj, Američané – Američani, šachisté – šachisti, diplomaté – diplomati, o střediscích – o střediskách, méně – míň* atd.

Na tomto místě je však třeba zmínit nezřídka nejednotný, nesystémový a problémový přístup k posuzování charakteru určitých morfologických forem. Zatímco část frekventovaných výrazů již do hovorové vrstvy spisovné češtiny patří (byla do ní dříve či později akceptována), jiné tvary – s neméně četným výskytem – zatím nikoliv, viz dichotomii spisovných a obecněčeských tvarů jako *déle – dýl; abychom – abysme*[141]*; kdybychom – kdybysme; lidé – lidi; Jitce jsi náhodou nevolal? – Jitces náhodou nevolal?* apod., viz i dále.

Situace se často komplikuje i vinou obecně rozšířené dezinterpretace onoho pojmu, jímž mnozí chápou nikoliv vrstvu spisovné češtiny, nýbrž češtinu běžně mluvenou, kolokviální. S běžnou mluvou v dnešním pojetí ji v podstatě ztotožňuje i Mathesius (1932): ten totiž hovorovou češtinou rozumí dosti volně vymezenou množinu spisovných a částečně též nespisovných prostředků užívaných v běžném hovoru mluvčími, kteří jsou zvyklí aktivně užívat spisovný jazyk.

Výstižně se o hovorové češtině vyjádřil Hrbáček (1995, s. 54–55): „*Jestliže však „hovorovou češtinu" chápeme jako mluvnou formu spisovného jazyka, pak by nespisovné, tj. nekodifikované jevy do ní patřit neměly, ať mají povahu obecně českou nebo i nespisovnou celonárodní. Že se do ní v různé míře zahrnují, vyplývá z toho, že pod pojmem hovorový jazyk se myslí dvě různé věci: 1. hovorová vrstva prostředků spisovného jazyka, 2. hovorová forma jazykových projevů.*"

[141] O zařazení této formy do hovorové vrstvy spisovné češtiny se osvíceně zasazoval již Hausenblas v roce 1962.

Celkově lze konstatovat, že naznačená nejednota názorů a značná (i potenciální) rozkolísanost spisovné normy má značně negativní důsledky pro komunikační praxi, a tím i pro oblast lingvodidaktickou (tedy potažmo i pro námi sledovanou problematiku jazykové chyby a jazykové správnosti ve sféře češtiny jako cizího jazyka).

5.4. Jazyková chyba a výuka cizího jazyka

K podstatným změnám docházelo a dochází v námi sledované oblasti jazykové chyby rovněž v lingvodidaktice, přesněji řečeno v cizojazyčné výuce. V minulosti byla po celá staletí chyba oddělována od vlastního procesu učení se jinojazyčnému kódu, byla pokládána za defektní, a tudíž nežádoucí prvek (viz dále), bylo na ni pohlíženo především jako na neúspěch a na selhání mluvčího (s různě negativními důsledky). V rámci institucionalizované výuky cizího jazyka byla chyba mnohdy významným (ne-li rozhodujícím) hodnotícím kritériem řečové (ať psané, tak ústní) produkce mluvčího.

V současnosti převládá v moderní lingvodidaktice trend zásadně odlišný, ba přímo protichůdný. Učení se jinojazyčnému kódu (i společenská komunikace prostřednictvím přirozeného jazyka) se totiž chápe jako specifický druh lidské činnosti. Výzkum se proto zaměřuje na analýzu průběhu dané aktivity a odborníci mimo jiné usilují o nalezení optimální a efektivní cesty k jejímu ovlivňování a řízení. Adekvátní rozbor chyby (její typologie a specifická výpovědní hodnota)[142] se tak stává pro vyučujícího i pro jinojazyčného mluvčího cennou zpětnou vazbou a nenahraditelným zdrojem relevantních informací. K chybě se již nepřistupuje negativně, hodnotí se naopak jako přirozený jev, jako nevyhnutelná a integrální součást složitého procesu nabývání znalosti jinojazyčného kódu, a to jak cestou učení se cizímu jazyku (*learning*), tak formou osvojování si cizího jazyka (*acquisition*). (Srov. Hendrich et al., 1988 aj.) Filozofie tohoto přístupu by se dala stručně shrnout větou *Chybami se člověk učí.*

5.4.1. Druhy chyby z hlediska lingvodidaktického

Podobně jako je tomu i v jiných oblastech lingvodidaktiky, setkáváme se v souvislosti s vymezením druhů (typů) chyby s poměrně širokým spektrem názorů a koncep-

[142] Rozumí se jí kupř. informace o úrovni komunikační kompetence mluvčího v cizím jazyce, o úspěchu jím zvolené strategie učení se jinojazyčnému kódu, o jeho schopnosti analýzy a syntézy, o jeho dovednosti aplikovat lingvodidaktickou poučku na užívání určitého mluvnického jevu/mluvnické kategorie atd. Z tohoto důvodu se v odborné literatuře setkáváme s pojmy jako „smysluplná chyba", „smysluprostá chyba", „chytrá chyba" atd.

cí. V případě chyby však rozdíly mezi nimi nenabývají zásadního charakteru, proto stručně přiblížíme pouze několik vybraných přístupů. Stranou ponecháme otázku řečové etikety i problematiku jazykových rovin a plánů.

S konstruktivním přístupem k chybě přichází Corder (1967, 1973)[143]. Vyčleňuje několik druhů jazykových odchylek, a sice ***error*** a ***mistake, lapse***.

V případě *error* se jedná o chybu kompetence, kterou mluvčí sám nedokáže opravit (jedná se v jistém smyslu o „chybování z nevědomosti"), neboť se s příslušným jevem doposud buď neseznámil (je pro něho nový, neznámý, ještě se mu neučil) nebo jeho řečové uplatnění nesprávně pochopil a není schopen ho v komunikaci adekvátním způsobem aplikovat. V této souvislosti pokládáme za důležité zdůraznit nutnost náležitého (tedy kognitivně pojatého, dostatečně častého a intenzivního) praktického procvičování jazykového učiva ve výuce, protože pouze touto cestou, a nikoliv pouhým nebo převážně teoretickým výkladem si mluvčí mluvnickou (i jinou) látku náležitě osvojí.[144] Zmiňovaný typ chyby (bez ohledu na jeho různé možné příčiny) je systematický, častý, lze ho předvídat i na základě zkušeností z vyučovací praxe.

Jako jeden z typických frekventovaných příkladů z oblasti výuky češtiny pro cizince můžeme uvést nepatřičné užívání spojení *„dělat rád"* začátečníky, popř. mírně pokročilými cizinci, např. anglicky a francouzsky hovořícími. Větu *Studuju rád češtinu* pod vlivem interference zpravidla chybně vyjádří následovně: *Mám rád studovat češtinu* (= *I like studying Czech; J'aime étudier le tchèque*).

Pokud jde o *mistake, lapse* (chybná morfologická forma, záměna syntaktické struktury, přeřeknutí se, stylistické nedopatření aj.), jde podle Cordera o chyby performance. Tyto jazykové prohřešky mohou být detekovány a opraveny žákem, který zná správné řešení a který se dopustil chyby např. v důsledku nedostatečného zautomatizování užití příslušného jevu.

Jinou klasifikaci chyb přináší Norrish (1987). Rozlišuje ***error, mistake a lapse.***

Termínem *error* označuje systematickou chybu, resp. odchylku od náležité formy cílového jazyka, které se cizinec soustavně dopouští z toho důvodu, že je určitý jev pro něho nový nebo nedostatečně pochopený a osvojený (řekli bychom nedostatečně internalizovaný).

Mistake proti tomu představuje odchylku nestálou, kolísavou, nesystematickou (občasnou). Tento druh chyby svědčí o nedůsledném učení se, o neadekvátním pochopení příslušného jevu (po stránce formální i řečové).

Prohřešku nazvaného *lapse* se dopouštějí jak mluvčí rodilí, tak nerodilí. Je způsobován nedostatečnou koncentrací mluvčího, jeho únavou, stresem apod.

[143] Uvedeno podle Korčáková, 2004. Corderův příspěvek „Die Rolle der Interpretation bei der Untersuchung von Schülerfehlern" vyšel ve sborníku *Fehlerkunde. Beitrage zur Fehleranalyse, Fehlerbewertung und Fehlerterapie.* G. Nicker (ed.), Berlin: Verlag für Lehrmedien KG, 1973, s. 38–50.

[144] Srov. diametrální rozdíl mezi „učením se jazyku" a „učením se o jazyce" (srov. např. Hrdlička, 2009 aj.).

Mnozí vyučující ještě uvažují o dalším druhu chyby, a sice o tzv. *careless slip,* přeřeknutí, tedy o chybě zapříčiněné žákovou nedbalostí (nesoustředěnost, nedostatečný zájem apod.).[145]

5.4.2. Příčiny vzniku jazykové chyby

Rovněž výše uvedené téma je v odborné literatuře poměrně často diskutováno. Z perspektivy zaměření této kapitoly lze z jednotlivých přístupů stručně připomenout následující důvody.

Mezi nejčastějšími, a to na všech jazykových rovinách a plánech, bývá uváděn (mezijazykový)[146] ***negativní transfer (interference)***, který je objektivní povahy (viz dále). Role mateřského jazyka v procesu nabývání znalosti jinojazyčného kódu je značná a prokazatelná, přestože se některé přístupy, např. přímá metoda, snaží úlohu výchozího jazyka potlačit.

Velmi negativní roli sehrává také ***nevhodná aplikace lingvodidaktických pouček*** v procesu učení se cizímu jazyku, a to jak v učebních materiálech, tak ze strany vyučujícího. Není výjimkou, že optimální poučení a) chybí (máme-li na mysli oblast češtiny jako cizího jazyka, pak se jedná např. o absenci návodů, jak užívat primární prepozice), b) je zavádějící, zjednodušené (odkazujeme na prezentaci slovesného vidu, tvoření rozkazovacího způsobu atd., viz blíže Hrdlička, 2009).

V odborné literatuře je rovněž zmiňována ***objektivní obtížnost***[147] některých ***mluvnických kategorií a jevů*** (opomenout nelze ani roli mnohdy podstatné typologické odlišnosti mateřského a cizího jazyka) a snaha mluvčího realizovat svůj komunikační záměr navzdory nedostatečné znalosti jazykového systému cizího jazyka.

[145] V lingvodidaktické odborné literatuře se však setkáváme ještě s celou řadou dalších přístupů. Tak kupř. Edge (1994) člení chyby z pozice vyučujícího na *slips* (žák je schopen chybu opravit), *errors* (žák chybu schopen opravit není, je však nicméně zřejmé, který tvar chtěl, resp. měl použít; spolužáci správné řešení znají) a *attemps* (chybování v neznámém a doposud neprobraném jevu, event. nejasnosti s komunikačním záměrem mluvčího – není jasné, který tvar chtěl žák použít a co chtěl vyjádřit). Bartram a Walton (1991) klasifikují chyby na *slips* (jde o odchylky způsobené únavou, nepozorností, stresem aj.), *mistakes* (chyby typické pro jinojazyčné, nikoliv pro rodilé mluvčí) a *covert mistakes* („skryté chyby", žák se vyjádří správně pouze náhodou, nikoliv na základě znalosti systému cizího jazyka) atd.

[146] Někteří autoři uvažují rovněž o tzv. vnitrojazykové interferenci, srov. např. Hendrich et al. (1988, s. 367): *„... vnitrojazyková interference s počtem osvojovaných jevů stoupá. Zvláště silný je vliv právě osvojené struktury na struktury osvojené dříve (žáci mají např. sklon užívat nově osvojeného slovesného času příliš často, na úkor ostatních časů)."*

[147] V této situaci je ovšem nutno poznamenat, že jde zpravidla o hodnocení subjektivní a že ve velké míře závisí na úrovni lingvodidaktické prezentace příslušného jazykového (gramatického) jevu – na způsobu jeho výkladu, procvičení apod.

Důvodem může být také ***chybná aplikace lingvodidaktické poučky*** nebo pravidla ***mluvčím*** (např. v důsledku jeho nedostatečné schopnosti uplatnění příslušné instrukce nebo následkem její přílišné generalizace;[148] projevit se může i nedostatečné procvičení daného jevu apod.). Připomíná se i podstatná role ontogeneze[149] mluvčího, jeho věk, intelekt, jazykové vlohy, paměť, motivace apod.

Svůj podíl mají pochopitelně i další psychosomatické faktory, v nichž zmíníme ***únavu, nesoustředěnost, stres, spěch***, a v neposlední řadě i ***osobnost vyučujícího***, resp. jeho možný negativní vliv na mluvčího (nedostatečná znalost vyučovaného jazyka, nesprávná výslovnost atd.). Netřeba jistě zdůrazňovat, že ve vyučovací praxi běžně dochází k různé kombinaci uvedených příčin a že by je měla v náležité míře reflektovat i práce s chybou.

5.4.3. Jazyková chyba a vyučovací metody cizích jazyků

Většina vyučovacích metod cizích jazyků přistupuje k chybě pouze nebo převážně jako k jevu negativnímu, stojícímu mimo odbornou pozornost badatelů. Odchylka mluvčího/pisatele od jazykové normy je v zásadě pokládána za nežádoucí (některé z lingvodidaktických přístupů, např. metoda sugestopedická, chybu do určité míry tolerují), odlišnosti mezi jednotlivými metodami spočívají především v hodnocení závažnosti příslušného řečového pochybení. Pro komunikační metodu jsou z pochopitelných důvodů nejzávažnější chyby zabraňující porozumění psanému, resp. mluvenému projevu.[150]

[148] V podobných případech se hovoří o vytváření mylných analogií. Jde kupř. o to, že se určité pravidlo vztahující se na pravidelné jazykové jevy mechanicky aplikuje i na jevy nepravidelné, srov. nevhodné tvoření nom. pl. Ma *člověk – člověci, člověkové* či nevhodné stupňování adjektiv typu *dobrý – dobřejší* apod.

[149] Korčáková (2004, s. 57, 58) s odvoláním na názory německých lingvodidaktiků (Koutiva aj.) uvádí: „*Žák si vytváří určité teorie o fungování jazyka, jejichž prostřednictvím přijímá a zpracovává nové informace. Při osvojování jazyka tedy není pasivní, ale vystupuje aktivně: analyzuje, dekomponuje, zjednodušuje, tvoří a modifikuje hypotézy, syntetizuje, tvoří analogie atd. Na učení se jazyku proto nelze pohlížet jako na tzv. norimberský trychtýř, kterým do hlavy žáka nalijeme vědomosti, díky nimž se pak budou ze žáka ven „sypat" jen správné věty. Nová látka musí být vysvětlena, zpracována a integrována do již naučeného systému. Znalost gramatického pravidla a schopnost aplikovat ho v průběhu rozhovoru jsou totiž dvě odlišné věci.*"

[150] Co se „gramatické správnosti" týká, Společný evropský referenční rámec pro jazyky (2002, s. 116) vyčleňuje v souladu s jednotlivými úrovněmi popisů z pozice jinojazyčného mluvčího šest stupňů znalosti mluvnice cílového jazyka:

A1 – *Ovládá jen v omezené míře několik základních gramatických struktur a typů vět, které jsou součástí osvojeného repertoáru /…/.* A2 – *Používá správně některé jednoduché struktury, ale přitom se systematicky dopouští elementárních chyb – např. má sklon k zaměňování časů, zapomíná na gramatickou shodu. Nicméně je stále jasné, co se pokouší vyjádřit /…/.* B1 – *Komunikuje přiměřeně správně ve známých kontextech; všeobecně ovládá gramatiku dobře, ačkoliv vliv mateřského jazyka je postřehnutelný. Chyby se objevují, ale je jasné, co chce vyjádřit. /…/ Přiměřeně správně používá repertoár běžných gramatických prostředků a vzorců v rámci snadno předvídatelných situací. /…/* B2 – *Ovládá dobře gramatiku: jen občas se dopouští malých nebo nesystematických chyb a mohou se objevit menší nedostatky ve větné stavbě, ale nejsou časté a mohou být zpětně opraveny /…/ Nedopouští se chyb, které by mohly způsobit nedorozumění. /…/* C1 – *Dodržuje důsledně vysoký stupeň gramatické správnosti; zřídka se dopouští chyby a chyby jsou sotva postřehnutelné /…/* C2 – *Dokáže důsledně ovládat gramatiku jazyka v její komplexnosti, i když věnuje pozornost něčemu jinému (např. promýšlení dalšího sdělení, sledování reakce jiných) /…/.*

Soustavná pozornost věnovaná jazykové chybě se datuje od padesátých let dvacátého století, a to hlavně díky behaviorismu a jeho lingvodidaktickému vyústění, tedy audioorální metodě.

Podle behavioristů jde v cizojazyčné výuce především o vytváření návyků (prostřednictvím drilových cvičení) a zautomatizovaných reakcí mluvčího na jinojazyčné podněty na základě principu stimul – reakce (srov. např. Šebesta, 1999, Hrdlička, 2009 aj.). Vliv interference z mateřského jazyka má být potlačen intenzivním nácvikem (memorováním modelových struktur),[151] který by měl minimalizovat možnost odchylky, resp. nesprávné (jiné než modelové) formulace ze strany jinojazyčného mluvčího.

Chyba je pokládána za defektní jev, hodnotí se jako výraz nedostatečně kvalitní výuky (projev neadekvátní motivace a koncentrace mluvčího apod.).

Zcela opačný přístup k jazykové chybě vzniká v druhé polovině šedesátých let dvacátého století, a sice teorie interlanguage[152] (Corder, 1967, Selinker, 1972 aj.). Odchylka je považována za přirozený fenomén a za zákonitou, nevyhnutelnou průvodní okolnost procesu nabývání přirozeného jazyka (ať mateřského, tak cizího). Mnoho „systémových" (předvídatelných) chyb vzniká proto, že si jinojazyčný mluvčí z různých důvodů (viz výše) ve fázi učení se jazyku (osvojování si jazyka) vytváří (částečně) nesprávnou představu o jeho struktuře a o jeho fungování.

Mluvčí si průběžně (jde o sérii kontinuálních přechodů) vytváří o cílovém jazyce stále přesnější hypotézu. Onen „mezijazyk", přechodný a přechodový „mezikód" (někdy se hovoří o tzv. třetím systému) tak postupně obsahuje stále méně rysů jazyka mateřského a naopak stále více rysů jazyka cílového (srov. např. Hrdlička, 2009).

[151] K tomu měl výrazně přispět tzv. kontrastivní přístup (Fries, Lado aj.), resp. detailně propracovaný popis jazyka výchozího (L1) a cílového (L2). Jak už bylo výše naznačeno, zůstalo pouze u očekávání. Ukázalo se totiž, že jak odlišnost jazykových systému mateřského a cizího jazyka, tak i oblast lingvodidaktiky jsou mnohem složitější povahy.

[152] Onen pojem se vykládá dvojím způsobem: jednak označuje polohu onoho autonomního jazykového útvaru mezi dvěma jazykovými systémy (tedy mezi mateřským a cizím jazykem), jednak dané pojmenování odráží potenciální kvalitu znalosti cílového jazyka pohybující se v intervalu 0–100 %.

V moderní lingvodidaktice se klade důraz mimo jiné na kognitivní aspekt, na odlišování komunikační závažnosti chyby i na náležitou práci s ní (chyba má být nejen opravena, ale mluvčímu má být zřejmé, v čem a proč chyboval).[153]

Na základě dosavadních poznatků lze konstatovat, že v současnosti převládá ve společenské komunikaci prostřednictvím přirozeného jazyka požadavek přijatelnosti jazykového, resp. řečového projevu jinojazyčného mluvčího. Míní se jí nejen jeho srozumitelnost, ale i další atributy, mezi něž patří zejména přiměřenost dané komunikační situaci.

5.5. Morfologické tvary systémové, nesystémové a defektní

Dostáváme se nyní k problematice zaujetí lingvodidaktického postoje k různým druhům jazykových (morfologických) tvarů a odchylek (ať mluvčích rodilých, či nerodilých), resp. k vymezení základních druhů morfologických forem z hlediska jejich přináležitosti k určitým varietám českého národního jazyka i z hlediska jazykové chyby a jazykové správnosti. Podle našeho mínění (podotýkáme, že se z perspektivy lingvodidaktického využití jedná o určité nutné zjednodušení) jsou jimi morfologické tvary systémové, nesystémové a defektní. Na základě předchozího výkladu je následující naznačená klasifikace tvaroslovných forem vcelku zřejmá a neproblémová.

Bylo by možné konstatovat, že za ***systémové*** by byly označeny takové ***morfologické tvary***, které jsou buď v souladu se (současnou) kodifikací normy spisovné češtiny, nebo v souladu s normou toho kterého strukturního útvaru českého jazyka (obecná čeština, určitý dialekt). Je nasnadě, že jeden a týž morfologický tvar (kupř. *Jde/De* ***do školy***) může být součástí několika útvarů, konkrétně v tomto případě češtiny spisovné a obecné, nikoliv však již např. nářečí z oblasti Slovácka (správný tvar by zněl *Ide* ***do škole***).

Domníváme se, že z hlediska výuky češtiny jako cizího jazyka by měly být přednostně prezentovány k aktivnímu osvojení morfologické tvary spisovné neutrální a hovorové. Celá záležitost je však podstatně složitější, srov. blíže Hrdlička, 2009, 2010 aj.

Jako ***nesystémové morfologické tvary*** by bylo možné interpretovat „cizorodé", resp. nesourodé systémové tvarotvorné formy z jiného strukturního útvaru českého jazyka (např. z perspektivy standardu by jimi byly formy nespisovné, kolokviální, např. *před* ***dvouma*** *roky*; z pozice obecné češtiny by jimi byly např. spisovné formy: *před* ***dvěma*** *rokama* apod.).[154]

Především z pohledu prestižní variety českého národního jazyka je pak zajímavé sledovat přechod (ovšem nedůsledný a mnohdy dosti kontroverzní a bolestný) „nesystémových" obecněčeských morfologických forem k „systémovým" (přesněji řečeno ke spisovným hovorovým).

Na „čekací listině" by se tak mohly, resp. měly ocitnout velmi frekventované kolokviální morfologické formy, které doposud nebyly uznány za spisovné hovorové. Máme nyní na mysli především:

a) již zmíněný tvar komparativu, event. superlativu adverbia *dlouho*, tedy *dýl*, resp. *nejdýl* (zde bezesporu působí rušivě ona hláskoslovná změna *-é-* (*déle*) v *-ý*, neboť v jiných případech, konkrétně *málo*, resp. *méně – míň*, již k pospisovnění došlo);

b) účelovou a podmínkovou spojku *abysme* („*jsme*"), *kdybysme* – srov. analogické *abyste* („jste"), *kdybyste*);

c) nepravidelný nominativ plurálu substantiva *člověk*, tedy *lidi* (srov. např. obdobné případy jako *hosté, hosti; Židé, Židi; diplomaté, diplomati; akrobaté, akrobati; občané, občani* apod.);

d) kontrahované tvary préterita, konkrétně pospisovnění běžně užívaných substantivních tvarů s koncovým *-s* ve druhé osobě singuláru: *Chlebas náhodou nekoupil? Aleněs to dal taky?* (Srov. analogické spisovné případy typu *Modrous koupil? Komus to řekl? Kams šel potom? Cos tam napsal? Kdys přijel zpátky? Pročs to neřekl dřív? Jsem rád, žes mi pomohl. Zůstals doma, nebos šel ven?*).

Je však třeba poznamenat, že jde o problematiku velmi citlivou, konfliktní, na niž panují (diametrálně) odlišné názory, jak pokud jde o regionální původ bohemistů (jde zejména o dichotomii Čechy versus Morava a Slezsko), tak o jejich generační a jiné relevantní charakteristiky.

Podotýkáme, že jsou podstatné rozdíly v názorech na zařazení toho kterého morfologického prvku na ose spisovnost – nespisovnost rovněž z pohledu komunikátu psaného (v takovém případě jsme svědky přístupů spíše konzervativnějších), či mluveného (zde se můžeme setkat i s postoji poněkud liberálnějšími).

Ve výuce češtiny jako cizího jazyka by podle našeho soudu mělo jít mimo jiné rovněž o to, aby (alespoň pokročilejší) jinojazyčný mluvčí dokázal zmiňované morfologické „systémové" a „nesystémové" tvary správně identifikovat a adekvátním

[153] V odborné literatuře se zpravidla hovoří o otázce detekce (identifikace), interpretace a korekce chyby.

[154] V podobných případech, v nichž nedochází ke střídání jazykových kódů (*code switching*), nýbrž k jejich míšení (*code mixing*), se u nás hovoří o běžné mluvě jako o nestrukturním útvaru. Daneš (1997, s. 15) charakterizuje běžnou mluvu slovy: „*Termínem běžně mluvená čeština lze označovat repertoár všech různých, různorodých jazykových prostředků (nespisovných, ale zčásti i spisovných, resp. společných), kterých se užívá v situacích, v nichž se nepředpokládá závazné užívání spisovného jazyka. Jde tedy o rozsáhlou, nejednotnou, nehomogenní řečovou oblast vykazující vysoký stupeň variantnosti, kterou probíhá řada parametrů (lokální, regionální, generační, sociální aj.).*"

způsobem (tedy vědomě, soustavně, diferencovaně) jich v mluvené i psané komunikaci používat.

Za ***defektní morfologické tvary***, kterými se žákovský korpus zabývá především a kterých existuje nepřeberné množství i druhů, pak lze pokládat nesystémové odchylky od normy strukturních útvarů českého národního jazyka, kterých by se, stručně a zjednodušeně řečeno, (dospělý) rodilý mluvčí patrně nedopustil. Jako poněkud kuriózní příklad z pedagogické praxe uvedeme situaci, v níž došlo ze strany jedné neslovanské mluvčí ke konjugaci substantiva. Větu *Já hledat divadlo* z cvičení zaměřeného na časování sloves typu *dělat* dotyčná mluvčí upravila na *Já hledat divadlám*.

Díky reprezentativnímu množství shromážděného jazykového, resp. řečového materiálu, který obsahuje žákovský korpus, lze tyto defektní morfologické tvary, resp. jazykové chyby sofistikovaným způsobem třídit, analyzovat a diferencovaným způsobem (podle jejich povahy, typologie, s ohledem na jejich příčinu i na komunikační závažnost, podle jazykové provenience mluvčího, který se jich dopustil, a dalších podstatných faktorů) s nimi efektivně pracovat.

6. Budování specializovaného korpusu mluvčích ohrožených sociálním vyloučením a předpoklady jeho chybové analýzy – databanka ROMi

Zuzana Bedřichová, Kateřina Šormová

ROMi představuje v rámci korpusu CzeSL specifický subkorpus. Shromažďuje jazykové projevy – mluvené a psané – českých romských dětí a mládeže ve věku 6–28 let. Součástí tohoto subkorpusu je také část psaných jazykových projevů českých neromských dětí, která slouží jako srovnávací materiál.[155] Na rozdíl od korpusu jazykových projevů jinojazyčných mluvčích je tento zaměřený na jazyk dětí a mládeže a zároveň (opět na rozdíl od subkorpusu češtiny jinojazyčných mluvčích) obsahuje velké množství nahrávek mluveného projevu. V současnosti sestává z 1557 nahrávek mluveného projevu a 4865 textů. Svým rozsahem a zaměřením představuje zcela unikátní soubor jazykových pramenů, neboť se jedná o první takto rozsáhlý soubor jazykového projevu romských dětí a mládeže v češtině vůbec, který je navíc vybaven chybovou anotací.[156]

Od ostatního jazykového materiálu obsaženého v CzeSLu se však liší především tím, že se nejedná o projevy češtiny jako druhého jazyka, ale o projevy jazykově specifické skupiny romských mluvčích, pro které (nebo naprostou většinu z nich) je čeština jazykem prvním. Tato čeština má specifickou podobu, odlišnou od češtiny většinové českojazyčné společnosti, a to na úrovni mluveného a sekundárně také psaného jazyka. Jedná se o tzv. romský etnolekt češtiny, tedy o varietu češtiny užívanou romskými komunitami především na území České republiky. Jsou v ní patrné vlivy romštiny a slovenštiny.[157] Většina z těchto mluvčích navíc žije v sociokulturních

[155] Většina těchto textů, které slouží jako srovnávací materiál, vznikla současně s texty romských mluvčích. Jsou to převážně školní písemné práce, které vznikly ve třídách, jež navštěvují také romští žáci. Je tak možné zkoumat texty romských i neromských dětí vzniklé za stejných okolností, se stejným zadáním a ve stejné třídě.

[156] K chybové anotaci korpusu CzeSL viz více kapitola 4.

[157] Více k romskému etnolektu češtiny viz Bořkovcová (2006) či Šotolová (2008).

podmínkách, které hraničí se sociálním vyloučením, anebo je lze přímo označit jako sociální vyloučení. Je tedy otázkou, do jaké míry ovlivňují podobu zejména písemných textů oba tyto faktory – romský etnolekt i vlivy sociokulturní, a do jaké míry se prolínají a navzájem podmiňují. Hledání odpovědí na tyto otázky je jednou z možností bádání, jemuž může databanka ROMi sloužit.

V našem příspěvku se budeme věnovat především specifickým nárokům kladeným na proces sestavování takto specializovaného korpusu, tedy na problémy spojené se sběrem a zpracováním jazykového materiálu v prostředí ohroženém sociálním vyloučením, i na to, jakým způsobem mohou tyto faktory ovlivnit předpoklady pro chybovou analýzu.

6.1. Základní předpoklady sestavování specializovaného korpusu

Sestavování specializovaného korpusu přináší řadu specifických otázek, s nimiž bývají jeho autoři konfrontováni ještě před jeho zahájením. Protože se v případě ROMi jedná o korpus zaměřený na jazykový projev určité etnicky vymezené skupiny mluvčích češtiny (limitované navíc věkovou hranicí 6–28 let), týkala se řada těchto otázek jednak značně problematického vymezení skupiny mluvčích, jednak navázání kontaktu s mluvčími, samotného sběru dat a jeho dalšího zpracování.

6.1.1. Vymezení skupiny respondentů

Metodika sběru jazykových dat vychází ze skutečnosti, že je velmi obtížné oslovit větší skupiny romských žáků, jejichž jazykové projevy by bylo možné zařadit do korpusu. Protože jsme při oslovování potenciálních sběračů[158] jazykového materiálu naráželi na obavy z nařčení z diskriminace, zejména u institucí, jako jsou základní školy, bylo nutné stanovit takové kritérium, které by bylo přijatelné pro všechny situace sběru jazykového materiálu.

U sběru materiálu mimo školní prostředí, např. v rámci volnočasových sdružení, nízkoprahových klubů, individuálního sběru apod., tento problém nenastával, protože respondenti[159] neměli problém se identifikovat se svým romstvím.[160] Jak již bylo řečeno, zcela opačná situace nastávala při navazování spolupráce se základními školami, kdy nebylo možné se žáků dotazovat na jejich romství ani je jiným způsobem veřejně označovat za Romy. Oficiální statistiky o počtech romských žáků na jednotlivých školách neexistují, navíc v některých případech nebyli učitelé nakloněni myšlence podílet se na vzniku jazykové databanky češtiny romských žáků.[161] Zároveň bylo nutno vyřešit velmi citlivou otázku, koho označit za Roma, a tudíž zařadit jeho jazykový materiál do korpusu. Vycházeli jsme z obecně užívané metodologie dotazníkových šetření, kdy je za Roma považován ten, kdo se sám považuje za Roma a zároveň ho za Roma považuje také jeho okolí.[162] Jako určující faktor jsme pak použili kritérium jazykové,[163] kdy byly za respondenty považovány ty osoby ve věku 6–28 let, které umějí (alespoň pár slov) romsky.[164] V případě sběru materiálů na školách nebylo vhodné klást dotaz na vnímání romství přímo žákům, proto rozhodnutí, zda daného žáka zařadit do výzkumu, bylo na učiteli, který své žáky zná a měl by mít přehled i o jejich jazykovém zázemí. Tento přístup má samozřejmě svá úskalí a neřeší např. asimilované Romy, nicméně považujeme ho za nejpřijatelnější jak pro naše účely, tak pro učitele.

6.1.2. Etické aspekty sběru a zpracování dat

Značná část respondentů, a v případě nahrávek i sběračů, pochází ze sociálně vyloučeného prostředí anebo je sociálním vyloučením ohrožena. Proto je třeba si v neposlední řadě klást etické otázky, které ze spolupráce s takovou skupinou mluvčích plynou. Kromě samozřejmého zajištění anonymity respondentů (zajištěné přepisem veškerého, jak psaného, tak mluveného materiálu a záměnou všech údajů, které by mohly respondenty identifikovat), je potřeba se zamyslet, jaká je vlastní motivace vytvoření korpusu, a nevyhýbat se přitom zodpovědnosti, která z práce na korpusu plyne. Je účelem nasbírat unikátní jazykový materiál, který umožní autorům korpusu dosáhnout určitých vědeckých cílů, nebo získat podporu pro sociálně atraktivní projekt? Může projekt nějakým způsobem sociálně ohroženým osobám prospět? Na všechny tyto otázky je třeba hledat pravdivé odpovědi a mít je při celém procesu vzniku a dalšího zpracování korpusu na zřeteli.

[158] Sběračem zde rozumíme osobu, která v rámci projektu získávala jazykový materiál pro účely databanky. V případě mluvených projevů to byla osoba nahrávající rozhovory, v případě písemných textů osoba, která pro potřeby projektu texty zprostředkovávala.

[159] Respondentem zde rozumíme osobu, která je autorem jazykového projevu.

[160] Nejjednodušší situace nastala při pořizování nahrávek, kdy se nám podařilo navázat spolupráci s jedním romským sdružením. Jeho členové sami nahrávali uvnitř své vlastní komunity položené rozhovory a mohli tak sami určovat, kdo je a kdo není Rom.

[161] Při kontaktování škol jsme vycházeli zejména z Mapy sociálně vyloučených a sociálním vyloučením ohrožených romských lokalit v České republice, mapa je dostupná z: http://www.esfcr.cz/mapa/int_CR.html.

[162] Viz Zpráva o stavu romských komunit v České republice, dostupná z: www.vlada.cz.

[163] Viz Šatava, 2001.

[164] K jazykové situaci romských dětí a jejich kompetencích v češtině a romštině viz více Červenka, Kubanik, Sadílková (2010).

Databanka ROMi byla od počátku koncipována jako jazykový materiál, který bude sloužit pedagogickým účelům, mj. zkoumání funkční gramotnosti romských žáků.[165] Jako takový byl také prezentován sběračům a respondentům, přičemž byl důraz kladen na to, že databanka v budoucnu poslouží jako podklad pro tvorbu nových učebních pomůcek, které pomohou romským žákům překonat potíže s češtinou, jež bývají uváděny jako nejčastější důvod odchodu romských dětí ze základních škol na školy praktické.[166]

Přesto to neznamená, že databanka bude skutečně mít pozitivní dopad na romské žáky; rozhodující bude její reálné využití, zejména to, zda se podaří prosadit její využití ke konkrétním pedagogickým účelům, jež se promítnou do metodiky výuky romských žáků.

Etické nároky klade také obsahová stránka jazykového materiálu; texty i nahrávky jsou často velice zajímavé, respondenti vyprávějí o svých životních situacích, které jsou většinové společnosti často zcela vzdálené nebo pro ni nepředstavitelné a zároveň vzbuzující senzaci. Je proto potřeba vyhnout se při dalším zpracování (např. publikaci ukázek) určitému voyerství a senzacechtivosti. Nevyhnutelný eklekticismus, ke kterému při vybírání ukázek vždy dochází, může představovat nebezpečí konstruování určitého obrazu dané skupiny mluvčích na veřejnosti (k tomuto konstruování dochází pochopitelně vždy; je ale nutné, aby si autoři jakýchkoli publikací, které používají ukázky z korpusu, byli vědomi, jaký obraz svými ukázkami tvoří, a zpytovali vlastní předsudky, motivace a pohnutky, které je mohou při volbě ukázek ovlivňovat).

Na druhé straně každá publikace ukázek představuje výzvu; jde o to, jak jí využít k seznámení širší veřejnost s vlastními výpověďmi respondentů o sobě samých, o prostředí, ve kterém žijí, sociálních a rodinných vztazích apod. V tomto ohledu nelze nikdy uvažovat o objektivní neutralitě.[167]

Určitou etickou výzvu, jak již bylo řečeno, představuje také finanční zázemí projektu. Při sestavování ROMi jsme se snažili, abychom mohli do projektu začlenit na placené pozice (tj. v našem případě pozice sběračů jazykového materiálu) co nejvíce Romů. Podařilo se nám tak navázat spolupráci s několika romskými asistenty na školách, ale především s řadou individuálních sběračů a s pracovníky některých romských občanských sdružení, kterým vděčíme za jazykově i obsahově velice zajímavý materiál.

[165] K funkční gramotnosti viz více Palenčárová, Šebesta, 2006.

[166] Více viz zpráva GAC 2007.

[167] Např. Samarin (1967) a s ním Mosel považuje obsahovou zajímavost lingvistického korpusu za jedno z 6 kritérií dobrého jazykového korpusu.

6.1.3. Práce v terénu

Jak již bylo řečeno, při interakci s prostředím ohroženým sociální exkluzí je velice důležité vycházet především z potřeb a možností členů této skupiny a také z potřeb osob, které se v tomto prostředí pohybují či v něm pracují. V průběhu vzniku korpusu jsme se jednoznačně přesvědčili, že je třeba především respektovat odlišné nároky na materiální kvalitu textů a nahrávek a nezatěžovat respondenty ani sběrače omezujícími požadavky či nesplnitelnými nároky.

Při sestavování zásad sběru jazykového materiálu jsme se proto cíleně rozhodli nezadávat žádná omezující kritéria, která by zbytečně omezila jazykový materiál (např. v případě zadávaných témat bychom mohli oslabit obsahovou výpovědní hodnotu projevů) a zároveň by mohla odradit některé sběrače či respondenty. Chtěli jsme také oběma zúčastněným stranám poskytnout možnost vést rozhovor či iniciovat a psát text podle spontánních možností dané situace a do průběhu vzniku jazykového materiálu nijak nezasahovat. V případě písemných textů jsme např. uváděli možná témata pouze jako určité vodítko pro sběrače, případně jako usnadnění vzniku textů.

Součástí korpusu jsou tak i nahrávky, které nejsou technicky příliš kvalitní, jež vznikaly např. v domácnosti za účasti dalších osob, hrající televize apod. Nahrávka by však v jiných podmínkách nikdy nemohla vzniknout (je přirozené, že nahrávky rozhovorů s mladšími dětmi vznikaly většinou přímo v domácnosti za přítomnosti rodičů), anebo by vznikla, ale byla by ovlivněna nepřirozeností situace, což by ovlivnilo i jazykový materiál. Kromě nahrávek mladších dětí se to týká zejména rozhovorů sběračů se svými přáteli vrstevníky např. v prostorách volnočasového klubu v plném provozu.

Podobně jsme se snažili co nejvíce zjednodušit metajazykové informace o pořizovaných jazykových projevech. V případě písemných textů sice bylo nezbytné zachytit do samostatného formuláře alespoň základní údaje o respondentovi (podrobněji dále), ale u nahrávek jsme se rozhodli nahrát informace o respondentovi (tj. věk, pohlaví, třídu a eventuelně typ školy, který navštěvuje) přímo jako součást nahrávky. Díky tomuto zjednodušení se nám podařilo zjednodušit nahrávací situaci na pouhou přítomnost zapnutého diktafonu.

6.1.4. Cílová skupina uživatelů korpusu

Podobně jako ostatní subkorpusy CzeSL je i databanka ROMi zaměřena na cílovou skupinu pedagogů, budoucích pedagogů a odborníků v oblasti jazykovědy a vzdělávání. Pedagogické využití usnadňuje chybová anotace, a to jak anotace ruční, tak automatická.

6.2. Charakteristika ROMi – První žákovský korpus pro češtinu

Korpus ROMi zahrnuje písemné práce romských žáků v češtině a nahrávky řízených rozhovorů těchto žáků s učiteli, pracovníky volnočasových organizací a studenty. Minimální rozsah korpusu byl plánován na 500 tisíc slov; již v současné době, kdy nejsou zpracovány všechny materiály, je však jasné, že rozsah bude větší.

6.2.1. Metodika sběru jazykových dat

V rámci projektu jsou sbírána jazyková data psaná i mluvená. V průběhu projektu jsme dosud spolupracovali s přibližně stovkou sběračů – zhruba polovinu z nich tvoří učitelé nejrůznějších typů škol, druhou polovinu tvoří pracovníci neziskových organizací a studenti středních a vysokých škol. Sběr materiálů probíhá ve dvou fázích. V první fázi byly sbírány jazykové projevy romských mluvčích, jak mluvené, tak psané. V současné době je nasbíráno přibližně 1,6 milionu slov prostřednictvím nahrávek a přibližně 450 tisíc slov prostřednictvím psaných textů.

Sběr materiálů od romských mluvčích probíhá na dvou rovinách – prostřednictvím spolupráce s pražskými i mimopražskými školami různých typů (jde o běžné základní školy, základní školy praktické a speciální) a prostřednictvím individuálních sběračů, jimiž mohou být například romští pedagogičtí asistenti, pracovníci různých nevládních neziskových organizací, kteří spolupracují s Romy, případně studenti středních a vysokých škol, kteří sbírají materiály přímo v terénu.

Na školách probíhá ve spolupráci s učiteli češtiny zejména sběr písemných materiálů, tedy školních slohových prací, jejichž téma vybírá učitel. Ten také ovlivňuje způsob zadání tématu (vybírá vstupní aktivitu před samostatnou prací; rozhoduje, zda půjde o výběr z více témat, či zda je zadáno pouze jedno téma; diskutuje s žáky o základních pojmech v souvislosti s vybraným tématem apod.). Učitelé mají také k dispozici modelová témata, která mohou podle vlastního uvážení modifikovat, případně použít témata vlastní (*Jaké jídlo mám nejraději; Jaká je moje oblíbená hračka/zábava; Jakou poslouchám hudbu a proč; Můj nejlepší kamarád; Můj nejhezčí den/moje nejhezčí narozeniny; Nejlepší vynález lidstva, který mi usnadňuje život; Co bych dělal, kdybych vyhrál 10 milionů; Co mě baví dělat, když skončí škola*).

Celkem jsme dosud spolupracovali s 43 školami[168], přičemž 55 % nasbíraných materiálů připadá na základní školy, 34 % na praktické školy, 9 % na speciální základní školy. Mimo školní prostředí spolupracujeme přibližně s 35 individuálními sběrači. Působí nejčastěji v různých občanských sdruženích zaměřených na vzdělávání romských žáků, někteří z nich jsou sami Romové, případně se jedná o romské pedagogické asistenty. Významnou část sběračů tvoří studenti romistiky nebo bohemistiky s profesním zájmem orientovaným na češtinu romských žáků.

Nahrávky jsou zatím pořizovány pouze od romských mluvčích, autory písemných prací jsou jak romští žáci, tak žáci neromští (procentuální zastoupení je přibližně 64 % Romů a 36 % žáků neromských). Zastoupení chlapců a dívek je vyrovnané, 54 % respondentů tvoří chlapci, 46 % dívky. Důvodem sběru nahrávek pouze od romských respondentů je zejména časová a technická náročnost sběru, stejně jako finanční náročnost zpracování.

Z celkového počtu respondentů tvoří prozatím 52 % žáci prvního stupně základních škol (nebo odpovídající věková skupina žáků škol praktických a speciálních), 46 % žáci druhého stupně, 1 % žáci středních škol a 1 % žáci SOU nebo nestudující. To odpovídá našemu přednostnímu zaměření na věkové skupiny odpovídající devíti ročníkům základní školy; v tomto věkovém rozmezí je naším cílem přibližně vyrovnané zastoupení všech věkových skupin.

Většina písemných prací je poměrně malého rozsahu (přibližně 60 slov na text), délka textů je velmi proměnlivá a bývá ovlivněna zejména věkem respondentů. Dolní věková hranice je omezena schopností produkovat psaný text (tzn. minimálně druhé pololetí prvního ročníku), horní věková hranice je stanovena na 28 let.

6.2.2. Zpracování materiálů

Materiály odevzdané sběrači jsou nejprve naskenovány a nahrány do jazykové databanky AMES[169] (o databance blíže v kap. 4, odd. 4.4), přes kterou probíhá další distribuce materiálů k přepisům, kontrolám a následně i anotacím, zároveň databanka obsahuje doplňující informace ke vzniku textu, sběrači, respondentovi a lokalitě vzniku materiálu. Každý jazykový materiál obsahuje evidenční číslo, které podává informaci o sběrači, respondentovi, místě sběru i pořadí nasbíraného materiálu. Materiály jsou sbírány anonymně, tzn. obsahují pouze takové identifikační údaje, aby nedošlo k záměně jednotlivých respondentů (od některých respondentů máme materiálů několik) a sběračů. Pokud materiály obsahují jakékoliv osobní údaje, které by mohly vést k identifikaci respondenta (mnoho zejména písemných prací slouží zároveň jako klasifikované školní práce, navíc žáci jsou ze školy zvyklí podepisovat se celým jménem apod.), jsou při dalším zpracování odstraněny a nahrazeny kódy.

Ke každému jazykovému materiálu náleží průvodka, která podává informace o okolnostech vzniku jazykového materiálu, a anamnestický dotazník (viz dále).

[168] Všechny uváděné číselné údaje se mohou ještě změnit, sběr materiálů není ukončen a bude pokračovat v roce 2012.

[169] www.ames.ff.cuni.cz, přístup na webovou stránku je chráněn uživatelským jménem a heslem, přístup do databanky také.

Údaje z dotazníků i průvodek jsou postupně zanášeny do databanky AMES, po zpracování všech materiálů bude na jejich základě možné vytvořit reprezentativní statistiky. Databanka slouží zejména pro distribuci jazykových materiálů přepisovačům a kontrolorům, kteří mají přístup do databanky chráněný heslem a umožněný pouze k těm materiálům, které sami zpracovávají. Přepisovači a kontroloři jsou v naprosté většině případů studenti bohemistických oborů, kteří se zajímají o problematiku jazykově handicapovaných komunit.

6.2.3. Sledovaná metadata

Ke každému nasbíranému materiálu sběrač vyhotovuje anamnestický dotazník a průvodku. Společně s údaji o sběrači a místě jeho působení jsou tyto údaje převedeny do databanky, která následně obsahuje několik typů průvodních informací. (K metadatům viz i kap. 2, odd. 2.3.1.) Jde o údaje vztahující se:

a) k jazykovému materiálu a místu jeho vzniku (datum sběru; počet slov; počet znaků; místo sběru; velikost sídla a nářeční oblast; region; prostředí (školní, mimoškolní, soukromé); sociálně vyloučená lokalita;
b) k jazykovému materiálu a okolnostem jeho sběru, které vyjadřují míru řízenosti jejich tvorby učitelem (téma zadáno/nezadáno; slohový postup a rozsah zadán/nezadán; povaha a rozsah přípravných aktivit; okolnosti sběru – psáno nebo namluveno jako součást zkoušky; psáno nebo namluveno přímo pro korpus apod.);[170]
c) k respondentovi (věk žáka; třída/ročník; pohlaví; typ školy; subjektivní znalost romštiny; první jazyk – který jazyk žák pokládá za svůj první a komunikační prostředí v rodině – který jazyk/jazyky se užívají v rodině ke komunikaci, příp. zda někdo v rodině mluví romsky);
d) k místu sběru materiálu – v případě škol jde o charakteristiku podle typu školy (základní, praktická, speciální, SOU, SŠ), podle zřizovatele (státní, církevní, soukromá) a sociologicky (sídlištní škola, vesnická škola, tradiční městská zástavba), v případě místa individuálního sběru označení organizace, zájmové skupiny apod.;
e) ke sběrači (zkratka jména sběrače a místo jeho působení, v některých případech věk a skutečnost, zda jde o Roma/Neroma).

[170] Materiál pro žákovské korpusy obecně jen zřídka představují autentické, přirozené jazykové projevy, vzniklé z autentické komunikační potřeby v autentických situacích reálného života. Jde zpravidla o texty klinicky elicitované. K tomuto termínu blíže v odd. 2.2.2, kap. 2.

Organizace sběru jazykových materiálů a údajů z průvodek a dotazníků je administrativně velice náročná, mnozí sběrači působící v sociálně vyloučených lokalitách nemají možnost v místě sběru využívat počítač, proto je převážná většina průvodních materiálů dodávána v papírové podobě, což klade velké nároky na čas strávený jejich zpracováním. V některých případech jsou anamnestické dotazníky a průvodky nekompletní, důvodem může být to, že se jedná o starší písemné práce, ke kterým sběrači nemohli všechna sledovaná metadata zjistit, nebo v některých případech sběrači odmítli žádané informace zjišťovat.

6.2.4. Charakteristika textů – rozsah, nejčastější žánr a téma

Rozsah textů je ovlivněn věkem respondentů – 98% všech nasbíraných materiálů spadá do věkové kategorie žáků základní školy, tzn. do 15 let. Více než polovinu z nich (52%) tvoří žáci prvního stupně, průměrný rozsah textu je kolem šedesáti slov. Co se žánru týká, převládá vypravování, u starších respondentů úvaha. V materiálech se objevují témata velice různorodá, zastoupena jsou jak vzorová témata, která dostávají učitelé společně s pokyny ke sběru materiálů, tak témata vlastní. Ve věkové kategorii 6–10 let jsou to nejčastěji vyprávění z prázdnin, o rodině, úvahy typu co bych dělal, kdybych vyhrál velkou sumu peněz nebo čím bych chtěl být. Ve věkové kategorii 11–16 let se často opakují témata spojená s volbou budoucího povolání a výběru školy, překvapivě často také úvahy na téma rasismus, diskriminace, romství, co pro mne znamená být Rom apod. Tyto úvahy se nesou jak v obecné rovině, tak v souvislosti s aktuálním děním (popálení malé Natálky zápalnou lahví, nepokoje na severu Čech apod.).

Vzhledem k tomu, že nemalé množství textů vzniká mimo školu, vyskytují se také stylizace textů do formy dopisu nebo básně pro učitele (resp. sběrače), drobné dedikace a obrázky, vzkazy psané na okraj.

6.2.5. Přepis materiálů a jejich primární zpracování

6.2.5.1. Přepisy textů

Úkolem přepisu je zachovat autentickou podobu původního textu, při přepisech se neopravují chyby ani neuvádí správná verze.[171] Pravidla pro přepis textů jsou v základu shodná s pravidly pro přepis textů jinojazyčným mluvčích, jsou pouze doplněná o některé specifické situace vyskytující se v romských textech. Mnoho nestandardních situací při přepisech souvisí s faktem, že jde o rukopisné texty psané nedospělý-

[171] Problematika přepisu rukopisů pro potřeby žákovského korpusu je podrobněji zmiňována již v odd. 2.4.1, kap. 2.

mi respondenty, často na hranici čitelnosti. Sníženou čitelnost ještě podporuje fakt, že přepisovači pracují s texty naskenovanými, v krajních případech je nutné vyhledat text v archívu (je-li k dispozici v papírové podobě) a přepsat ho přímo z originálu. Čitelnost textů také v některých případech ovlivňuje fakt, že vznikají v neformálním prostředí, některé texty jsou pomačkané, plné nejrůznějších ilustrací a emotikonů, jednotlivé části textu mohou být v některých případech psány přes sebe.

Nejčastější sporné situace vznikají v těchto případech:

a) nesystematické používání a míchání velkých a malých písmen v rámci jednoho textu,
b) varianty jednotlivých písmen, např. a/o,
c) varianty velké/malé písmeno,
d) varianty jedno/dvě slova,
e) autorovy vlastní rektifikace a přesuny,
f) rektifikace způsobené zásahem učitele,
g) chybějící diakritická znaménka,
h) chybějící interpunkce,
i) nečitelná písmena/slova.

Texty vznikající v neformálním prostředí, ale i některé texty z prostředí školního, obsahují velké množství osobních a citlivých informací. V případě osobních údajů (tzn. jména, adresy, místa bydliště, telefonní čísla, emailové adresy apod.) jsou tyto informace nahrazovány zástupnými jmény (v případě osobních a místních jmen), v případě dalších osobních údajů kódy pro osobní údaje. Přístup do korpusu bude v budoucnu víceúrovňový, přístup na rovinu skenů obsahující tyto údaje nekódované nebude standardně přidělován.

6.2.6. Charakteristika nahrávek – rozsah, nejčastější žánr a téma

Nasbírané nahrávky pocházejí jak ze školního prostředí (zejména rozhovory romských pedagogických asistentů s jednotlivými žáky), tak z prostředí volnočasového (zájmové kroužky, mimoškolní doučování). Ve většině případů jde o rozhovor nahrávajícího s jednotlivcem, případně dvojicí respondentů, objevilo se ale i několik nahrávek vyučovacích hodin s vyšším počtem mluvčích. Rozsah nahrávek je velice proměnlivý, od několika minut až po šedesátiminutové rozhovory. Kvalita nahrávek je ovlivněna možnostmi nám dostupné techniky a snahou o co největší autenticitu. Nahrávky byly pořízeny se souhlasem žáků, nicméně diktafony byly umístěny tak, aby žáky příliš nerozptylovaly a neovlivňovaly autenticitu jejich projevu. Z toho důvodu se na některých nahrávkách projevuje zvýšený šum a snížená kvalita.

6.2.6.1. Přepisy nahrávek

Sběr nahrávek a jejich přepisy byly v polovině roku 2011 pozastaveny. Do té doby se podařilo přepsat přibližně 15 % všech nahrávek. Pravidla pro přepis nahrávek[172] vycházejí zatím z pravidel pro přepis nahrávek korpusu SCHOLA. Přepis se provádí folkloristickou transkripcí, tj. co nejblíže záznamu psanému, speciálně upravenou pro účely počítačového zpracování podle úzu zavedeného v Českém národním korpusu.

Nejčastější sporné body se při přepisech objevovaly v těchto situacích:

a) větná interpunkce – užívá se tak, jak je to obvyklé v textech spisovných, tj. nezachycuje se přerušování věty pauzami;
b) v pravidelných jevech se ponechává spisovná forma (např. znělé a neznělé souhlásky: *dub, sbírat, jablko* – a tedy i *jabko* atd.);
c) zachycují se příznakové rysy běžné mluvy (např. *dóle, vzádu*);
d) důsledně se zaznamenává jakákoliv jiná než standardní výslovnost, a to jak v případě spodoby znělosti – *Miz Róma (Miss Roma), šestnáz let*, tak v případě předložkového spojení: *vokně*;
e) v případech, kdy se svými ustálenými podobami běžná mluva od výslovnosti spisovné odlišuje, spisovný zápis nerespektujeme a tuto odlišnou výslovnost zaznamenáváme (např. *ste, von, pudu*);
f) změkčenou výslovnost zaznamenáváme (např. *uňiverzita, Jan Benďík*);
g) zaznamenáváme odlišnou kvantitu (fonologickou – *včéra, rohlík,* i emfatickou – *bóže*);
h) prodloužení způsobené váháním se nezaznamenává;
i) výrazný přízvuk (důraz) se označuje podtržením dané slabiky.

6.3. Chybová anotace

Podobně jako u ostatního jazykového materiálu zpracovaného v rámci databanky CzeSL probíhá i na písemných projevech romských žáků chybová anotace, a to jak manuální, tak automatická. K této chybové anotaci viz více kapitola 4. Pro jakoukoli chybovou analýzu textů obsažených v ROMi, ať již bude provedena na základě chybové anotace, nebo jinými metodami, je potřeba mít na paměti specifické aspekty vzniku, sběru a zpracování jazykového materiálu, které tuto chybovou analýzu mohou ovlivnit.

[172] Jak bylo uvedeno v kapitole 2, pravidla pro přepis nahrávek se budou ve spolupráci s pracovníky Fonetického ústavu FF UK ještě upravovat.

6.4. Předpoklady chybové analýzy v ROMi

6.4.1. Vliv sběru a zpracování jazykových dat na jejich chybovost

Jednou z velkých předností databanky ROMi je její chybová anotace, jež umožní řadu lingvistických výzkumů. Dříve než bude možné vypracovat detailní chybovou analýzu využívající např. statistické metody pracující s chybovou anotací, je potřeba se zamyslet nad tím, jak tuto chybovost ovlivňuje metoda sběru jazykového materiálu a jeho další zpracování. Věříme, že následující poznatky mohou být podnětné při budování jakékoli další podobně specializované databanky. Nejdříve je ovšem nutné vymezit, co pod pojmem chybovost rozumíme.

6.4.2. Pojetí chybovosti u psaných textů

Protože je naše databanka primárně zaměřena na pedagogické využití a jejím nejvlastnějším cílem je přispět k poznání funkční gramotnosti romských uživatelů češtiny, chápeme chybu jako nefunkční odchylku od normy (k tomu srov. i kap. 5), tak jak ji chápe školní pedagogická praxe (opírající se o pravidla pravopisu pro školu a veřejnost a další odborné texty, zejména učebnice schválené Ministerstvem školství).[173] Tato norma je proměnlivá vzhledem k ročníku dítěte (např. u žáka první třídy se nevyžaduje znalost vyjmenovaných slov, ale stranově správná orientace písmen apod.).

Takový přístup lze relativně bez obtíží, alespoň těch teoretických, uplatnit u textů, které vznikly jako školní práce v rámci výuky. Problém ovšem představují texty, které vznikly v neoficiálním či polooficiálním kontextu. U nich lze předpokládat, že jejich autoři vnímali normu méně závazně. Jedná se především o projevy obecné češtiny včetně lexikálních a gramatických jevů, chybějící diakritika – jev běžný v elektronické komunikaci, méně závazné užívání malých a velkých písmen běžné v neformální elektronické komunikaci (sms, skype, různé formy chatování apod.) včetně zcela nahodilého střídání malých a velkých písmen či psaní celého textu velkými písmeny.[174]

[173] Viz např. Skalková (1999).

[174] Zde je třeba upřesnit, že texty, na nichž byla provedena chybová anotace, byly anotovány na základě jednotných kritérií, nehledě na typ textu a okolnosti jeho vzniku.

6.4.3. Vliv metody sběru a přepisu na chybovost písemných projevů v ROMi a její hodnocení

Při sestavování a analýze podobně specializovaného korpusu je proto třeba dbát na jasné odlišení jednotlivých typů textů nejen podle zadání (slohová práce v hodině, domácí úkol apod.), ale také, a to především, podle okolností vzniku textu. Při budování naší databanky jsme záměrně zvolili co největší různorodost typů textů i situací jejich vzniku; texty vznikaly v situacích s odlišným stupněm formálnosti. Jejich autoři mají navíc široké věkové rozpětí, jsou jimi žáci od 1. třídy základní školy po koncové ročníky školy střední či vyšší odborné, a patří mezi ně i mladé osoby, které školu nenavštěvují, ale spadají do věkové kategorie, kdy by ji navštěvovat ještě mohly (v případě vysoké či vyšší školy cca 26–28 let). Mezi respondenty jsou zastoupeni žáci velmi různých typů škol; setkáváme se navíc s jevem, že vlivem přechodu z jedné školy na druhou či opakování ročníku variuje i vztah mezi věkem a navštěvovaným ročníkem.

Prvním a zásadním problémem je čitelnost rukopisu, mnohdy značně ztížená, a jeho následná interpretace při přepisu. Příčinou nečitelného rukopisu může být samozřejmě jak charakter rukopisu jako takový, tak možné specifické poruchy učení[175]. Do značné míry může čitelnost ovlivnit i fakt, že přepisovači pracují nikoli s originálním textem, ale s jeho skenem, případně se skenem fotokopie. Při přepisu byla volena zásada, aby se přepisovač vždy snažil pochopit individuální rysy rukopisu tak, jak by je hodnotil z pohledu učitele (např. nevýrazné odlišování *b*/*l*, které se objevuje v celém textu, nepovažujeme za chybu).

Přes uplatnění těchto zásad nebylo vždy jednoznačně možné rozhodnout, o jaké písmeno/písmena či slovo se jedná. Problematické byly jednak skupiny písmen, která jsou si podobná (*a*/*o*; *l*/*k*/*h*/*b*/*t*; *e*/*i*), jednak špatně čitelná či přepisovaná místa, příp. nejednotnost psaní malých a velkých písmen (např. jejich libovolné střídání). Navíc se v textech objevily i chybné podoby písmen (stranově převrácená písmena; *m* psané s jedním obloučkem navíc apod.). V případě nejasností tedy byla zvolena metoda zápisu obou variant slova, která však může místy přepis zbytečně zatěžovat a přidávat chybovost tam, kde původně není.

6.5. Závěr

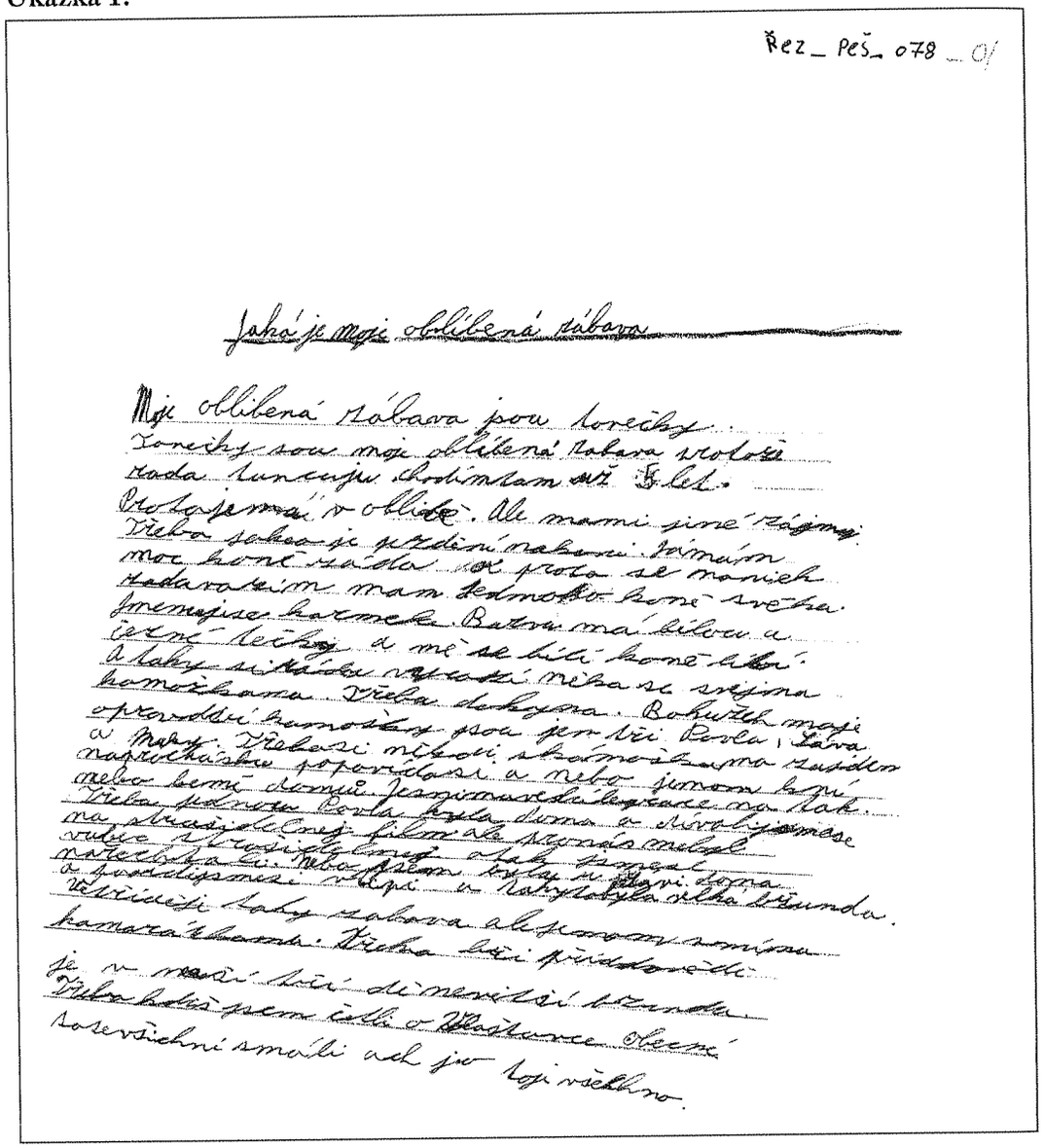
V této kapitole jsme se pokusili představit specifické nároky na výstavbu specializovaného korpusu mluvčích ohrožených sociálním vyloučením, jakým je chybově anotovaná databanka projevu romských žáků ROMi. Kromě detailního představení

[175] Např. dysgrafie.

databanky jsme se pokusili nastínit i ty aspekty, které mohou ovlivnit možnou chybovou analýzu, ať již prováděnou na základě chybové anotace, jíž je část textů ROMi vybavena, nebo pomocí jiných vědeckých metod.

6.6. Ukázky

Ukázka 1:



Ukázka 2:

Moje oblibená zábava jsou tonečky. Tonečky|Tanečky sou moje oblíbená zabava protože rada tuncuju. chodímtam už 5 let. Protojemá v oblibě{předposlední písmeno vícekrát přepisované}<co>. Ale mami jiné zájmy|zájmi. Třeba jakoo je jezdění nakoni. Jámám moc koně ráda a proto se nanich radavozím mam Jedmoho koně svéha. Jmemujese karmela. Barvu má bílou a černé tečky a mě se bílí koně líbí. A taky siráda vyrazí měka se svejma kamožkama. Třeba dokyna. Bohužeh maje|moje opravdiví kamošky jsou jem tři Eva1, Eva2 a Lily. Třebasi někdi skámoškama zajden naprochásku popovídasi a nebo jemom kní nebo|mebo kemě domiů. Jesnimavelálegrace na tak. Třeba jednou Eva1 byla doma a dívalijsmese na strašidelnej film ale pronás mebyl vubec strašidelmej atak jsmese nařechtali. Nebo jsem byla u Evi2{špatně čitelné}<co>. doma a povídijsmesi vtipi a takytobyla velká bžunda. Vetříděje taky zabava alejemom smíma kamarátkama. Třeba bři příddovědě je v neaší{druhé a třetí písmeno přepisované}<co> tří dě{ }nevětší bzunda. Třeba kdiš jsem četli o vlaštovce|Vlaštovce obecné tosevšichni smáli ach jo to je všchno.

Ukázka 3:

Slohová prace: Vzpomínky na prazdníny
Moje Prazdníny žačaly brygadou delal sem s otcem a s bratrem. delal sem u Firmy polak zemní prace. Poměsíci jsmesi v zali vslna a jeli sme na Slovensko k rodině kterou sem nevíděl dva roky. Sbratranci jsne chodili na kopaliště teta si novy rodiny domek tak sme stravili pordní jomoponím sXXX bratrem sme jim položily novou podlahu take vymalovalí do bytu novou vodu. Povečerech jsne chodily do boru pobavitse těch desetní ubjehlo rychle přijelí jsme domu a zase jsmese pdí do prace tak takle probjehly moje Prazdníny:
<img krumpáč, lopata, hromada hlíny, za ní chlapcova hlava>

Ukázka 4:

Horymíruv skok
Kníže Václav chtěl odsoudit Horymíra k popravě a Horymír řekl že se chce ještě jednou projet na svém koni šemíkovi kníže Václav na to když máš svého koně tak rád můžeš se na něm naposled projet. Horymír šel do stáje a potichoučku mu něco po{š|ž}eptal pak v[y']jely ven. Horymír pak dakrát hvízdnul a Šemík začal skákat Horymír pak hvízdnul třikrát a Šemík se rozeběhnul a s vipjetím všech sil viskočil a jako vítr přeletěl hradbi a už byl na druhém břehu Vltavy. Všichni se divili a začali tak dlouho přemlouvat Knížet Václava za ten Horymírův žazrak že nakonec svolil. Horymír pak přijel na Václavův hrad a řek že Šemík nechce ani pít ani jíst a že už kolabuje.

7. Nástin využití žákovských korpusů pro jazykové vyučování

Svatava Škodová

V posledním desetiletí žákovské korpusy mění náš pohled na jazyk a na jeho užívání. Jednou z otázek, kterou existence žákovských korpusů vyvolává, je, jak je využít přímo pro potřeby jazykového vyučování. V roce 1999 Yukio Tono[176] představil možnosti využití žákovských korpusů, které shrnul do pěti bodů: 1. popis vývojových úrovní žákovského mezijazyka, 2. studium vlivu mateřského jazyka a jazykový transfer, 3. vymezení nadužívání a podužívání jazykových prvků v žákovském jazyce, 4. rozlišení mezi univerzálními chybami a chybami vzniklými na základě vlivu žákova mateřského jazyka a 5. rozlišení prvků komunikace rodilých a nerodilých mluvčí, které způsobují dojem cizosti. Po více než deseti letech výzkumu v této oblasti je stále třeba souhlasit s tímto tematickým rozložením a s důležitostí vymezených okruhů. Avšak je také možné říci, že více vědecké práce se doposud soustředilo na samotný sběr a budování korpusů, výzkumné studie a závěry nejsou tak rozsáhlé, jak by bylo možno očekávat. Také samotná aplikace a následné zkušenosti učitelů a žáků s použitím žákovských korpusů ve vyučování nejsou příliš četné, a není proto možné ani z této oblasti vyvodit pro jazykové vyučování jednoznačné závěry. Ačkoliv je tedy možné říci, že otázka využití žákovských korpusů se jeví jako velmi podnětná a zajímavá, přesto stále převažují hypotézy a očekávání kladená na tento výzkumný a pedagogický nástroj nad konkrétními aplikačními výsledky.

Korpusoví lingvisté, kteří na vytváření žákovských korpusů pracují, jsou velmi optimističtí ohledně jejich užitečnosti pro výuku druhého/cizího jazyka, protože patří mezi vědce, kteří v současné době nejlépe mohou posoudit jejich informační potenciál; přesto to zcela jistě nebudou oni, komu se podaří žákovské korpusy uvést do vyučovací praxe. Proto je třeba, aby možnost využití žákovských korpusů byla poskytnuta vědcům, kteří budou schopni potenciál žákovských korpusů rozvinout.

Cílem kapitoly je představit dosavadní tradici využití žákovských korpusů ve světové aplikované lingvistice, naznačení jejich specifik ve srovnání s jinými typy korpusů a některých problémů spojených s jejich užitím. Výklad se zaměří na stručný

[176] Tono, Yukio. Using Learner Corpora in ELT and SLA Research. Paper presented at the Symposium on the Roles of Corpora in Language Teaching and Language Engineering of the 12th World Congress of Applied Linguistics (AILA), 1–6 August 1999, Tokyo, Japan.

popis současného stavu, na načrtnutí jejich obecného přínosu a limitů. Dále bude představen možný vztah žákovských korpusů a jazykového vyučování.

7.1. Žákovské korpusy – jejich definice s ohledem na vyučovací praxi[177]

Stejně jako národní korpusy bývají i žákovské korpusy obecně definovány jako *systematické digitalizované soubory autentických textů produkované žáky cizího jazyka*. Tato definice do značné míry vychází z charakteristiky korpusů národních, srov. např. definici F. Čermáka: *Korpusová data jsou dnes ve vztahu k jazyku charakterizovaná jako (1) typická, (2) aktuální, synchronní a věrná, (3) neselektivní, (4) objektivní a realistická, (5) dostatečná, (6) nenáhodně získaná a (7) získatelná a získávaná snadno a rychle.*[178] S ohledem na pedagogickou praxi je však třeba některé aspekty této definice vysvětlit a v určitých detailech zpřesnit.

Pro všechny typy korpusů je důležitá systematičnost jejich budování, tato systematičnost zajišťuje, že texty zařazené do korpusu byly vybrány na základě určitého počtu předem stanovených a přesně definovaných kritérií, většinou externího charakteru. Vzhledem k tomu, že žákovské korpusy shromažďují tzv. mezijazyk, je reprezentativnost a vyváženost žákovských korpusů jiného charakteru, než jak je tomu u korpusů národních. Charakter mezijazyka je možné označit jako dynamický, resp. neustále proměnlivý s ohledem na progresivní nabývání jazyka žáků, díky této charakteristice je třeba počítat s tím, že žákovský korpus nepředstavuje reprezentativní soubor homogenního jazyka vzhledem k určitému časovému období ve vývoji jazyka, ale značně heterogenní masu, kterou je třeba odkazovat k dílčím úrovním osvojování. Lze tedy říci, že žákovský korpus zachycuje mezijazyk vázaný na skupiny jinojazyčných mluvčích s ohledem na jejich mateřské jazyky.

Dalším slovem, které je třeba si v definici žákovských korpusů upřesnit, je jejich autenticita. Jazykový materiál získávaný pro žákovské korpusy není možné charakterizovat jako autentický, tj. přirozeně se vyskytující, přirozeně produkovaný, příp. produkovaný spontánně, jak to vyžaduje komunikační situace, ve smyslu autenticity materiálu korpusů národních. Do žákovských korpusů se autentické texty dostávají jen zcela výjimečně, neboť získání takových materiálů na všech úrovních mezijazykového vývoje je pravděpodobně nemožné. Jazyk žáků osvojujících si druhý jazyk je obvykle vázán na školské prostředí a jeho produkce je vždy do jisté míry řízena. Není proto možné hovořit o autenticitě v pravém slova smyslu a jazykový materiál je následně třeba analyzovat a interpretovat s vědomím této řízenosti.[179] (Podrobněji viz kapitola 2.)

Další definiční problém ve spojitosti s žákovskými korpusy může vznikat při vymezení termínu *žákovský*, který v českém prostředí činí celou definici poněkud nejasnou. Slovo *žák* je v anglicky mluvících zemích, ve kterých vzniká převážná většina žákovských korpusů, obvykle používáno pro označení jinojazyčného mluvčího učícího se cizí/druhý jazyk (L2) v zemi, kde se tímto jazykem hovoří jako prvním jazykem (L1). Poněkud méně typické je toto označení pro někoho, kdo se učí cizímu jazyku v zemi, kde tento jazyk není jazykem prvním. Mohli bychom tedy zavést rozdíl mezi typickým „žákovským" korpusem, tj. korpusem textů cizinců učících se L2 jako druhý jazyk v zemi, kde je L2 národním jazykem, a jinými typy korpusů nerodilých mluvčích (např. *the East African subcorpus of the International Corpus of English*[180]). Rozdíly tohoto typu by opět jistě přinesly rozdíly v korpusových datech získaných např. u cizinců učících se češtinu jako druhý jazyk v ČR a u cizinců učících se češtinu jako cizí jazyk např. v Rakousku.

V českém prostředí je dále třeba upřesnit, že význam slova *žák* je v terminologickém označení žákovský korpus oproti současnému úzu rozšířen nejenom na věkovou skupinu žáků základních a středních škol, tak jak je obvyklé, ale na všechny věkové kategorie i studijní statusy, tj. obsahově zahrnuje všechny jazyk se učící osoby bez ohledu na věk a institucionální příslušnost.

7.2. Rozpětí výzkumu žákovských korpusů

Žákovské korpusy jsou velmi speciální s ohledem na interpretaci dat, která obsahují; od analytiků totiž vyžadují mnohem širší oborovou základnu, než jaká je nezbytná pro výzkum korpusů národních. Výzkum žákovských korpusů leží na rozhraní přinejmenším čtyř vědních oborů: korpusové lingvistiky, jazykovědné teorie, osvojování druhého jazyka a výuky cizího jazyka. Každý z těchto oborů je důležitý a hraje svou roli ve vytěžení dat shromážděných v žákovských korpusech.

Znalost teorie SLA[181] je prerekvizitou pro interpretaci dat. Široká škála sociálních, kognitivních a psychologických faktorů, které hrají klíčovou roli v jazykovém učení, byla do hloubky studována v SLA a obeznámenost s výsledky těchto výzkumů pomáhá analytikům žákovský korpusů aby vyvodili správné interpretace (srov. např. žákovská produkce v Ellis, Barkhuizen, 2005).

[177] Srov. kapitola 1.

[178] Čermák, F. Korpus, informace a lingvistika. In: Přednášky z XLVIII. běhu LŠSS UK. Praha: Karolinum, 2005., s. 19–20.

[179] Tj. žákovské korpusy u textů uvádějí, za jakých okolností byly texty získávány, což určuje míru jejich elicitace, resp. autenticity.

[180] http://www.corpora4learning.net/resources/corpora.html.

[181] Second Language Acquisition.

Dobré základy znalosti lingvistické teorie jsou nezbytným předpokladem pro úspěšnou analýzu jazykových dat korpusů. Pro analýzu jazykového užití je ve světových žákovských korpusech úspěšně využíván funkční přístup (Meyer, 2002, s. 6), avšak současné studie ukazují, že to není jediná metodologická základna, o kterou je možné analýzy opřít. Z jiných jazykovědných přístupů, které byly doposud použity, je možné zmínit např. kognitivní přístup (Stefanowitsch, Gries, 2006). Stefanowitsch s Griesem využívají korpusová data pro zkoumání ustálenosti asociací mezi slovy a konstrukcemi v kognitivním rámci konstrukční gramatiky.

Nezbytná je i znalost cílů vyučování cizích jazyků, protože tato znalost vede k efektivnímu využití výzkumných výsledků pro pedagogickou praxi. Právě nové vyučovací přístupy, jako např. tzv. *Data Driven Learning* (DDT), by měly být pečlivě propracovány s ohledem na praktickou využitelnost a efektivitu vyučovacího procesu[182].

7.2.1. Studie opřené o žákovské korpusy

Obecně je možné konstatovat, že navzdory množství nesporně zajímavých poznatků vznikajících na základě analýz žákovských korpusů, je tento výzkum na samotném začátku, protože v tomto okamžiku sice již existuje poměrně velké množství studií postavených na materiálu žákovských korpusů, avšak výsledky těchto studií nejsou syntetizovány a zapojeny do nějakého širšího rámce. Cílem následující části je ukázat oblasti, na něž se dílčí studie opřené o žákovské korpusy zaměřují[183]. Vzhledem k tomu, že cílem textu není přinést kompletní přehled analyzovaných jevů v oblasti osvojování a učení druhého jazyka, byly v této části vybrány pouze ukázky možných analýz.

Většina studií publikovaná do této doby se opírá o ICLE korpusy a vychází z argumentativních esejů pokročilých žáků, tyto studie pokrývají už mnoho oblastí z různých rovin jazyka. Pokud bychom se pokusili vymezit oblasti, které jsou výzkumně frekventované, mohli bychom studie rozdělit do následujících kategorií.

7.2.1.1. Kategorie morfologicko-syntaktická

Studie v této oblasti se často zaměřují na dílčí užívání konkrétních jazykových prvků typu předložky, spojky, apod. Dále jsou zde frekventované studie o užívání různých typů vedlejších vět a otázek. Jinou výzkumně prosazovanou oblastí jevů jsou typy chyb s ohledem na úroveň osvojení jazyka podle SERR. V oblasti osvojování jazyka je pak často studován morfosyntaktický pozitivní a negativní transfer.

Příklady studií:

ABE, M. 2007. Grammatical errors across proficiency levels in L2 spoken and written English. *The Economic Journal of Takasaki City University of Economics*, 49, č. 3/4, s. 117–129.

BORIN, L. – PRÜTZ, K. (2004) New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In G. ASTON – S. BERNAREDINI – D. STEWART (eds.). *Corpora and language learners*. Amsterdam & Philadelphia: John Benjamins, s. 67–87.

BEKIOU, K. – DÍAZ, L. 2004. What the use of Childes can say in analysing second language acquisition data: The acquisition of Spanish simple past tenses by Greek L1 learners. In B. LEWANDOWSKA-TOMASZCZYK (ed.). *Practical applications in language computers (PALC 2003)*. Frankfurt: Peter Lang, s. 321–342.

BIBER, D. – REPPEN, R. 1998. Comparing native and learner perspectives on English grammar: a study of complement clauses. In S. GRANGER (ed.). *Learner English on computer*. London & New York: Addison Wesley Longman, s. 145–158.

7.2.1.2. Kategorie lexikologická

V oblasti popisu lexikální roviny jazyka vznikají studie popisující nadužívání, a naopak nedostatečně časté používání jednotlivých slov. Lexikologická oblast je v korpusových studiích hojně zastoupena, a to především z toho důvodu, že většina korpusů se opírá o žáky osvojující si angličtinu. Frázová složka tohoto jazyka je natolik důležitá, že je snadno pochopitelné, že právě popisy kombinatoriky slov, žákovské problémy s kolokačními spojeními a frazeologií v širokém slova smyslu budou ve výzkumném poli dominantní.

Příklady studií:

BACZKOWSKA, A. 2000. *The semantic analysis of word combinations in a learner EFL corpus*. Paper presented at the PALC'2009: Practical Applications in Language Corpora, University of Lodz.

CHEN, H. 1998. *Underuse, overuse, and misuse in Tawainese EFL learner corpus*. Paper presented at the First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, 14–16 December 1998, Hong Kong (China).

[182] Pro podrobnosti o využití DDT viz kapitolu 8.

[183] Rozsáhlá databáze bibliografických záznamů o textech přispívajících k poznání žákovského jazyka je uvedena na http://www.uclouvain.be/en-cecl-lcBiblio.html. V lednu 2012 obsahovala 600 bibliografických zápisů.

CROSS, J. – PAPP, S. 2008. Creativity in the use of verb + noun combinaitons by Chinese learners of English. In G. GILQUIN – S. PAPP – D. M. BELÉN (eds.). *Linking up contrastive and learner corpus research*. Amsterdam & Atlanta: Rodopi, s. 57–81.

MEUNIER, F. – GRANGER, S. (eds.) 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: John Benjamins.

COBB, T. 2006 Collocations in a learner corpus. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, *63*, č. 2, s. 293–295.

7.2.1.3. Kategorie diskurzu

Bez povšimnutí nezůstává ani nejvyšší jazyková rovina, rovina diskurzní. Zde vznikají studie jednak popisující rozdíly v komunikačních strategiích, jednak studie zabývající se spojitostí žákovských textů.

Příklady studií:

ASAO, K. 2002 Communication strategies of EFL learners: a corpus-based approach. In T. SAITO – J. NAKAMURA – S. YAMAZAKI (eds.). *English corpus linguistics in Japan*. Amsterdam & New York: Rodopi, s. 291–392.

BELZ, J. – VYATKINA, N. 2005. Learner corpus research and the development of L2 pragmatic competence in networked intercultural language study: the case of German modal particles. *Canadian Modern Language Review/Revue canadienne des langues vivantes*, *62*, č. 1, s. 17–48.

BOLTON, K. – NELSON, G. – HUNG, J. 2002. A corpus-based study of connectors in student writing: research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics*, 7, č. 2, s. 165–182.

CALLIES, M. 2006. *Information highlighting and the use of focusing devices in advanced german learner english. A study in the syntax-pragmatics interface in second language acquisition*. Marburg: Philipps-Universität.

7.2.1.4. Použití žákovských korpusů pro DDT

Ačkoliv využití metody *Data Driven Learning* patří mezi častá diskusní témata v oblasti výuky cizích jazyků, studií, které by se pokusily propojit žákovské korpusy s DDT, existuje poskrovnu.

Příklady studií:

LEECH, G. 1997. Teaching and language corpora. A convergence. In A. WICHTMANN – S. FLIGELSTONE – T. MCENERY – G. KNOWLES (eds.). *Teaching and Language Corpora* London: Longman, s. 1–23.

GRANGER, S. 1996. Exploiting learner corpus data in the classroom: Form-focused instruction and data-driven learning. Paper presented at TALC 1996, Lancaster, 9–12 August 1996.

GRANGER, S. – TRIBBLE, C. 1998. Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In S. GRANGER (ed.). *Learner English on Computer*. London: Longman, s. 199–209.

7.3. Potenciál a limity žákovských korpusů

7.3.1. Empirické základy SLA a žákovský korpus

Jedním z důležitých aspektů žákovských korpusů je přinést na výzkumné pole SLA širší empirickou základnu, než jakou mělo k dispozici doposud. Bez dostatečného množství průkazných dat není možné pokládat závěry za dostatečně věrohodné. Ford et al. (2005) poznamenávají, že ve vztahu k SLA nastává doba pro ověření dosavadních hypotéz na větší a lépe konstruované datové základně, podobné té, kterou má k dispozici výzkum osvojování prvního jazyka. Žákovské korpusy všeobecně obsahují data pocházející od stovek až tisíců žáků, také tematicky jsou jazykové podklady rozrůzněné a umožňují diverzifikovanější pohled na žákovský jazyk. Také je třeba si všimnout, že možná právě díky objemu dostupných dat se v jazykovém vyučování přesouvá váha od nejnižších úrovní výuky z tradičního zdůrazňování morfologické základny jazyků k upřednostňování výuky lexika, frázových konstrukcí, žánrové diverzifikaci diskurzu a mnoha dalším doposud zanedbávaným aspektům ve výuce cizích jazyků.

Všeobecná tendence ke zdůrazňování nutnosti velké datové základny v jazykovědě však nezbytně neplatí tak zcela absolutně ve výzkumu osvojování cizích jazyků. Pro korpusové lingvisty, jak poznamenává Sinclair, „*the whole point of assembling a corpus is to gather data in quantity*“ (Sinclair, 1995, s. 21), avšak pro výzkum osvojování a výuky cizích jazyků jsou požadavky na tvorbu korpusu mnohem složitější, čímž je od začátku limitována možnost získání tak rozsáhlé databáze, jako je tomu pro jazyky národní. Pro potřeby SLA je třeba sledovat mnoho dalších proměnných, které ovlivňují jazykovou produkci, než jen pouhou kvantitu, žákovské korpusy musí být nutně shromažďovány na základě velmi přesných předem daných kritérií, která

jsou uplatňována pro každého respondenta, od nějž jsou získávána data. V ideálním případě tak žákovský korpus umožňuje výzkum na základě různých proměnných (tj. nejenom na základě prvního jazyka žáka), těmito proměnnými mohou být např. využívané médium (mluvená vs. psaná produkce), typ úkolu, úroveň žáka, věk žáka.

7.3.2. Přínosy žákovských korpusů

Jednou z výhod žákovských korpusů je, že poskytují širokou materiálovou základnu pro analýzu. To je v oblasti výzkumů výuky a osvojování cizího jazyka velmi důležité, neboť mnoho výzkumů žákovského jazyka bylo donedávna vedeno především na experimentálních datech (např. typu multiple choice). Experimentální data mohou být v analýzách použita, jestliže se výzkum zaměřuje např. na abstraktní znalost jazykového jevu. Pro mnoho účelů je však důležité zjistit, co může student produkovat spontánně. Jedině při spontánní produkci se totiž může projevit zásadní rozpor mezi abstraktní znalostí systému a reálnou performancí žákovského jazyka. Vyvozovat závěry o tom, co může žák spontánně produkovat, pouze na základě experimentálních dat, nemůže být pokládáno za zcela spolehlivé.

Dalším nesporným pozitivem žákovských korpusů je, že umožňují více praktických výzkumů, zaměřených na dílčí jevy[184], které poskytují nikoliv separátně, ale v širším kontextu. Zatímco experimentální data jsou vhodná vždy jen pro výzkum omezeného množství jevů žákovského jazyka současně, žákovské korpusy dovolují sledování několika témat zároveň, resp. v součinnosti. Např. relativní frekvence různých typů chyb může být sledována i ve vzájemné souvislosti. Nadto není zcela nutné přistupovat ke korpusovým datům s předběžně zformulovanou hypotézou k určitému problému; díky tomu mohou být nové aspekty žákovského jazyka objeveny náhodně. Neposledním pozitivem žákovských korpusů je široké textové ukotvení jazykových jevů; právě jazykový kontext i metadata týkající se komunikačního ukotvení textů umožňují, aby na těchto textech mohly být studovány pragmatické otázky a otázky z oblasti diskursu, což zahrnuje např. komunikační strategie.

Protože podle definice je žákovský korpus systematický a je sestaven na základě určitého počtu přesně vymezených kritérií, může být na jeho základě analyzován také vliv jednotlivých kritérií na výslednou podobu textu a na jazykové jevy v textu obsažené. Např. jakýkoliv jev může být analyzován z hlediska úrovně znalosti cílového jazyka, z hlediska prvního jazyka studenta, typu textu, typu žákovského prostředí, ve kterém vznikl (přirozený, instruovaný), věku, pohlaví, délky osvojování jazyka, vlivu L3 atd.

[184] Příklady dílčích studií opírajících se o žákovské korpusy viz výše.

7.3.3. Limity žákovských korpusů

Bez ohledu na jednoznačný kladný potenciál mají žákovské korpusy přirozeně také své limity. V porovnání s pozitivním přínosem je však jejich počet skromnější, neboť nikdo nebude konstruovat výzkumný a pedagogický nástroj, aniž by se snažil o jeho maximální přínosnost. Ve výčtu limitů žákovských korpusů, jichž jsou si jejich tvůrci vědomi, kterým však nebylo možno se vyhnout, je např. jejich zaměření pouze na produktivní dovednosti, a tedy nemožnost jejich využití pro výzkum receptivních schopností studentů. Ačkoliv oba typy výzkumů jsou stejně důležité, není pravděpodobně možné úspěšně je propojit jedním výzkumným nástrojem.

I když žákovské korpusy poskytují velmi komplexní obrázek o individuálním žákovi a jeho jazykové produkci, přesto se i v této oblasti vyskytují jevy, které není možné s použitím žákovských korpusů stávajících typů zkoumat. Tímto jevem je např. zjišťování jistoty v používání určitého jazykového jevu ve specifickém kontextu.

Dalším omezením, které je platné v rámci celé korpusové lingvistiky, tj. není specifické pouze pro žákovské korpusy, je to, že není možné zkoumat jevy, které nejsou v korpusu obsaženy. Jestliže se určitá struktura/jazykový prvek nevyskytuje v textu, neexistuje způsob, jak zjistit, jestli ho student zná, nebo ne. Zvláště určité specifické jevy není možné v korpusech nalézt a je třeba je ověřovat experimentálním způsobem, který zajistí jejich elicitaci.

Také podrobnější investigace implicitních charakteristik určitého studenta je v rámci dat a metadat žákovských korpusů omezená. Není např. možné zkoumat motivaci použití určitých struktur. Podobně není možné přesně analyzovat roli učebního vstupu, ať už v podobě učebnice či jiných učebních materiálů; stejně tak i vliv užitých učebních metod působících na akvizici zůstává v žákovském korpusu pouze na pozadí a není možné jej přímo ověřovat. Ty je také možné zkoumat jen experimentálním způsobem, protože sám žák není schopen je reflektovat a i pro učitele by mohla být takováto reflexe neproveditelná a v zásadě subjektivní.

Ve výčtu jevů tohoto typu by byla i role interakce v jazykovém vyučování, tj. stimulu pro jazykové projevy, avšak tento jev je možné úspěšně zkoumat v mluvených žákovských korpusech, pokud jsou vhodně parametrizovány.

Dalším limitem, který je však způsoben pouze ne zcela dostatečnou rozpracovaností současných žákovských korpusů, je skutečnost, že žákovské korpusy existují jen pro malé množství jazyků (většina jich je zaměřena jen na angličtinu); stávající korpusy se omezují na sběr několika málo typů textů, na jednu vybranou jazykovou úroveň apod. (viz též kapitola 2).

Z tohoto výčtu limitů pravděpodobně vyplývá, že nejlepším postupem komplexních výzkumů v oblasti osvojování druhého jazyka je kombinace korpusové analýzy s experimentálním přístupem.

7.4. Žákovské korpusy a pedagogický materiál

7.4.1. ŽK a slovníky

Z nedostatku komplexních analýz vychází také ne zcela výrazný dopad žákovských korpusů na současné pedagogické materiály. Výjimkou jsou žákovské slovníky, pro něž byl také veden systematický výzkum. Prvním slovníkem, který zohledňoval data žákovských korpusů, byl *Longman Language Activator* (1993). Dále se o korpusový výzkum opírají slovníky *Longman Dictionary of Contemporary English* (1995); *Cambridge International Dictionary of English* (1995). Longman Learner Corpus byl použit k identifikaci nejčastějších žákovských chyb, které byly vyčísleny v tzv. *help-boxes* na konci každého hesla.

7.4.2. ŽK a gramatiky

Použití žákovských korpusů pro konstrukci gramatik pro cizince se doposud zásadním způsobem neprojevilo. Vzorem zdrojového využití žákovského korpusu pro gramatické účely by se mohl stát žákovský korpus TeleNex, který obsahuje texty čínských žáků učících se angličtinu a stal se zdrojem popisu gramatického systému angličtiny pro čínské žáky. Tzv. TeleGram[185] obsahuje informace o gramatice angličtiny, které jsou napsány pro všechny učitele angličtiny (avšak ne výlučně) v Hong Kongu. Gramatická deskripce angličtiny zdůrazňuje momenty, které jsou obtížné pro čínské žáky učící se anglický jazyk, jsou zde ukázány časté chyby a jsou vysvětleny na pozadí čínštiny.

7.4.3. Využití ŽK pro popis češtiny jako cizího jazyka

V tomto momentě je tedy pravděpodobně stále příliš brzy, abychom navrhli určité specifické postupy inkorporace výsledků studií žákovských korpusů do pedagogických materiálů. Pro češtinu by však mělo být přínosné už samo zmapování obtížných jevů pro jednotlivé skupiny žáků. Protože při materiálové prezentaci je velmi obtížné odhlédnout od popisu gramatického systému z hlediska rodilého mluvčího, žákovský korpus by měl být právě tím podkladem, který by tuto situaci mohl pomoci zásadně změnit.

Analýzy žákovských korpusů mohou ovlivnit především následující oblasti:

[185] http://www.telenex.hku.hk/telec/pmain/openreg.htm.

a) Mohou pomoci rozhodovat, které jevy by měly být ve výuce zvláště zdůrazňovány – jejichž explanace by měla být prohloubena. Mohou také pomoci odhalit jevy, které učebnice a materiály doposud vůbec nezahrnovaly. Nemusí se vždy jednat o jednotlivé gramatické prvky, často je třeba se zaměřit na rovinu lexikální, na frekventované fráze, typy spojení, apod. Příspěvkem v této oblasti selekce jazykových jevů, které by měly být vyučovány, by také mohlo být rozhodování v oblastech, co je pro žáky cizince (zvl. začátečníky) v užívání užitečné, tj. které jazykové elementy mohou úspěšně používat.
b) Výsledky korpusových studií mohou pomoci indikovat, jak určité jevy vyučovat. Např. Grangerová na základě svých výzkumů využívání časů cizinci vyvozuje, že časy je třeba vyučovat na úrovni textu, nikoliv na úrovni věty a že na úrovni pokročilých studentů je třeba časy studovat kontrastivním způsobem (Granger, 1999, s. 200).
c) Žákovské korpusy jsou vhodným prostředkem pro strukturaci gramatického systému z hlediska jinojazyčného mluvčího a mohou také pomoci determinovat pořadí, v jakém by měly být jevy vyučovány, aby odpovídaly určitým stadiím osvojení si jazyka.
d) Žákovské korpusy mohou být přímo využitelné pro čerpání příkladů typických chyb, jež je možno využít pro potřeby jazykového testování, pro které pomáhají vybírat oblasti obtížných jevů, nebo je s nimi možno pracovat přímo ve vyučování.
e) V neposlední řadě by na základě korpusových dat mohl začít být zohledňován i výchozí jazyk žáků, tj. L1. Žákovské korpusy, resp. subkorpusy skupin mluvčích s určitým výchozím jazykem, by tedy mohly pomoci vytvářet pedagogické materiály přihlížející k určitým L1, které jsou u cizinců učících se češtinu frekventované.

7.4.4. Žákovské korpusy a přímé nebo nepřímé pedagogické využití

Dalším rozdělením žákovských korpusů s ohledem na jejich pedagogické využití je de facto i způsob, kterým probíhá jejich sběr. Současným trendem je využití žákovských korpusů k nepřímým pedagogickým účelům, tj. žákovské korpusy nejsou používány přímo ve výuce jako studijní materiál. Jak bylo popsáno výše, žákovské korpusy jsou sestavovány pro akademické nebo komerční účely (tj. pro účely nakladatelství produkujících slovníky a gramatické příručky) za účelem validnějšího popisu tzv. mezijazyka a následně např. i pro sestavování vhodnějších výukových materiálů pro danou cílovou skupinu (tj. pro skupinu studentů s totožným profilem, který má skupina respondentů, na jejichž základě vzniká určitý žákovský korpus, např. výchozí jazyk, věk, úroveň apod.). Avšak v původní tradici byly korpusy sestavovány pro

přímé využití ve výuce. Tyto korpusy jsou shromažďovány učiteli jako součást jejich výukových aktivit, žáci jsou tedy zároveň korpusovými respondenty i uživateli korpusových dat. Tyto korpusy jsou samozřejmě pouze malého rozsahu. Ačkoliv bývají kvůli vágně zadaným parametrům a své nereprezentativnosti označovány jako *dirty corpora*[186], jejich výhodou oproti reprezentativním korpusům je, že jsou aktuální pro samotnou práci žáků. V posledních letech dokonce někteří jazykovědci upřednostňují tento typ malých korpusů před reprezentativními velkými korpusy, např. Mukherjee a Rohrbach (2006, s. 228) považují tento směr ve využití žákovských korpusů za jeden z nejslibnějších do budoucna. Domnívají se, že hlavním pozitivem tohoto přístupu je, že zaměření na výstup konkrétních žáků může do práce s korpusem přilákat mnohem více učitelů než velké korpusy postavené na „cizích" datech. Za velmi důležitý rys malých korpusů považují fakt, že prozkoumávání vlastní jazykové produkce motivuje žáky reflektovat používání jazyka a tak zvyšuje jejich citlivost na používání jazyka a umožňuje uvědomělou práci s osvojovanými jazykovými daty.

Jiným zajímavým příkladem užití žákovských korpusů, který v sobě zahrnuje možnost nepřímého i přímého využití žákovského korpusu, je tzv. *IWiLL*, tj. *Intelligent Web-based Interactive Language Learning*[187]. IWiLL je elektronické prostředí pro výuku jazyka, které umožňuje učiteli i žákům vytvořit a používat online databázi esejů tchajwanských žáků angličtiny. Přínosné by bylo už samotné online zpřístupnění žákovských textů, IWiLL však ještě navíc obsahuje texty, které jsou učiteli chybově tagovány (POS). Podle Wibleho tento korpus slouží pro přímé i nepřímé pedagogické účely: žákovská produkce je přímo zapojena do aktivit v průběhu vyučování a zároveň se tak naplňuje *Taiwan Learners' Corpus*[188].

7.4.4.1. ŽK korpusy pro přímé využití ve výuce

Doposud nejpřímějším a nejdůležitějším způsobem užití žákovských korpusů je identifikace specifických obtíží určité skupiny žáků a koncentrace na tyto jevy při výuce. Druhým způsobem práce se žákovskými korpusy (ať už malého či velkého rozsahu) je analýza chybovosti přímo ve výuce při přímém pedagogickém využití[189].

[186] Nečisté korpusy.

[187] Wible et al. (2001).

[188] Shi, R. H. H. Compiling Taiwanese Learner Corpus of English. Citováno z: http://www.google.cz/search?hl=cs&rlz=1T4SKPT_csCZ406CZ406&sa=X&ei=DQI5T5jZEObm4QTviKmhCw&ved=0CBsQvgUoAA&q=Taiwan+Lerners%C2%B4Corpus.&nfpr=1&biw=1366&bih=528&cad=cbv&sei=OwI5T4evJ4Si4gT05a2hCw 13. 2. 2012

[189] Vzhledem k tomu, že CzeSLCzeSL je velmi pečlivě strukturován v oblasti metadat, je vhodný i pro přímé využití ve výuce. Tj. je možné, aby učitel v korpusu selektoval data např.

V současné době se již nediskutuje o relevanci národních korpusů pro zlepšování jazykového vyučování. Je uznanou skutečností, že mohou lépe než pouhá intuice ukázat, co rodilí mluvčí píšou a říkají v určitých situacích. Pro jazykové vyučování však není pouze důležité vědět, co typicky používají rodilí mluvčí, ale také jaké jsou typické obtíže určité skupiny mluvčích při osvojování cílového jazyka. Žákovské korpusy poskytují základnu pro porovnání jazyka žáků s cílovým jazykem a umožňují ukázat tak na typologii obtížných jevů při osvojování cílového jazyka.

Jako nejpřímější a pravděpodobně nejdůležitější způsob užití žákovských korpusů se jeví identifikace specifických obtíží určité skupiny žáků, která umožňuje koncentrovat se na tyto jevy při přípravě výukových materiálů a při výuce samé. Druhým způsobem je analýza chybovosti přímo ve výuce, jak bylo naznačeno výše, při tzv. přímém pedagogickém využití žákovských korpusů. Poměrně tradiční metodickou námitkou je, že žák setkávající se s chybných textem, fixuje tento chybný text namísto textu (resp. struktur a jednotek) správného, bezchybného. Tato námitka je oprávněná zvláště na podprahových úrovních[190] osvojování jazyka. Pro vyšší úrovně se však práce s tzv. *negativní evidencí* jeví jako přínosná a podporující uvědomělé učení. S. Grangerová (Granger, 1996, s. 5) poukazuje na to, že využití negativní evidence je zvláště vhodné pro pokročilejší studenty, a to v případech, kdy dochází nebo již došlo k fosilizaci určitých forem. Přímé využití žákovského korpusu se na tomto poli přibližuje metodám výuky v podobě Data Driven Learning, avšak má v takových případech určité výhody před pouhým upozorňováním na daný druh chyb. Jednou z těchto výhod je, že jestliže žáci sami vyhledávají chyby a nedostatky, nebo spíše odlišnosti mezi jazykem žákovským a jazykem rodilých mluvčích, zvyšuje se jejich jazyková autonomie a obecně i jejich schopnost všímat si daných odlišností.

Takováto aplikace žákovského korpusu ve výuce je přínosná pro jevy, které již byly ve výuce probírány a opakovány, které však žáci užívají ne zcela uspokojivě, ať už s vysokou mírou nejistoty demonstrovanou v mluveném projevu, nebo s vysokou chybovostí.

V takovém případě pozitivum využití žákovských korpusů spočívá v tom, že umožní upozadit korektivní roli učitele. Namísto toho, aby učitel žáky opětovně upozorňoval na chyby v daném jevu a tyto jevy případně znovu vysvětloval, může využít pozitivní přístup a pomocí chybových žákovských dat nechává samotné žáky chyby vyhledávat a podle dosažené jazykové úrovně opravovat či vysvětlovat.[191] Je však třeba mít na zřeteli, že tato negativní evidence je ve výuce účinná pouze tehdy, když si je žák užité negativity vědom. Naopak je třeba zdůraznit, že není možné využít žákovský korpus jako materiál k vyhledávání v okamžiku, kdy se žáci učí jakýko-

od požadované skupiny žáků s výchozím L1, určité jazykové úrovně, atp.

[190] Tj. na úrovních A1 a A2 podle SERR.

[191] Leech (1997, s. 11) tento přístup označuje jako *divergent learning*.

liv nový jazykový jev, v takovém případě by chybové texty mohly zapříčinit chybné vyvozování pravidel či osvojování chybných struktur.

Je také třeba, aby zároveň s chybovými texty učitel žákům poskytl pozitivní evidenci jazykových dat, tato evidence může pocházet jak z umělých učebních materiálů, tak z korpusu národního. Učitel může předcházet fixaci negativní evidence jazykového jevu využitím cvičení, která by poskytla dostatečné množství korektních použití osvojovaných struktur, bezprostředně následující po práci s žákovským korpusem.

V neposlední řadě k práci s žákovskými korpusy přímo ve výuce je nutné zdůraznit, že i tato cvičení by měl učitel pro žáky pečlivě připravit. Tak jako je problematická on-line práce v hodině s korpusem národním, a to jak z důvodů technických, tak pro velké množství materiálu, ve kterém se student těžko orientuje, stejně problematické by pravděpodobně bylo i prohledávání korpusu žákovského. Učitel by tedy měl materiálovou evidenci sám dopředu připravit a žákům postoupit již vytvořený subkorpus obsahující chybový jev v požadovaném rozsahu a extrahovaný z relevantních textů.

7.5. Závěr a výhledy

Vznik korpusů zásadním způsobem proměnil přístup k jazyku a jeho užívání a vyučování. Časné publikace týkající se výzkumu žákovských korpusů explicitně zdůrazňovaly jejich kladný potenciál pro jazykové vyučování (Granger, 1993, Milton, Chowdhury, 1994), avšak doposud lze říci, že vzniklo zatím jen málo konkrétních pedagogických aplikací opřených právě o žákovské korpusy. Tento nedostatek může být jednoduše vysvětlen převažující analytickou prací, která samozřejmě nutně musí předcházet před návrhy pedagogických nástrojů opřených o žákovské korpusy.

Tato situace může být také částečně vysvětlena tím, že korpus není jednoduchým nástrojem a je třeba, aby se s ním učitelé naučili zacházet. Nelze očekávat, že bez zkušeností mohou získat informace, které jim korpus může poskytnout. Nutno podotknout, že nedostatkem vysokoškolské přípravy učitelů a lektorů je, že semináře korpusové lingvistiky někdy sice jsou zařazeny do univerzitních sylabů, 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 období vědeckého bádání na poli aplikované lingvistiky opřené o data žákovských korpusů.

8. Využití korpusových dat při výuce češtiny jako cizího jazyka

Pavlína Vališová

O metodě využívání korpusu v jazykové výuce (tzv. Data Driven Learning či Discovery Learning) píší korpusoví lingvisté a provádějí na toto téma své výzkumy již řadu let (převážně pro výuku angličtiny). Poukazují na to, jaký by korpus mohl mít přínos pro výukové materiály nebo pro výuku samotnou. Proč jsou však tato stanoviska přes nesporné výhody korpusových dat stále spíše teorií? Jednotlivým problémům spočívajícím na cestě vedoucí od korpusového výzkumu češtiny směrem k praktické jazykové výuce se věnuje sedmá kapitola. V této části se pokusíme poukázat na konkrétní případy využití korpusu v hodinách CJC a příklady doložíme výzkumem, který byl proveden mezi studenty češtiny pro cizince.

Český národní korpus (ČNK) existuje od devadesátých let dvacátého stolení a stal se zdrojem dat pro výzkum mnoha lingvistů. V roce 2009 byla vydána první gramatika zpracovaná na základě ČNK[192]; pro tvorbu výukových materiálů však byl zatím použit pouze omezeně. Překážkou frekventovanějšího využití korpusu může být jednak neznalost či nedostatečná znalost samotného korpusu a vyhledávání v něm, obtíže technického charakteru při zacházení s korpusovým manažerem; dále předpokládáme, že problematickou pro snadné využití korpusových dat pro didaktické účely může být i vysoká flektivnost češtiny, a to zvláště pokud se jedná o výuku češtiny pro cizince.

Jak překlenout propast mezi výzkumem a výukovou praxí je v tomto oboru častou otázkou. Avšak na důležitost tohoto kroku ukazují přednosti korpusu, zvláště to, že poskytuje reálná jazyková data, autentický jazyk, který není upravený pro pedagogické účely. Neexistuje jiný tak rozsáhlý zdroj reálného užití jazyka, než je korpus.

[192] Cvrček, V. a kol.: Mluvnice současné češtiny. Praha: Karolinum, 2010.

8.1. Korpus jako zdroj dat pro učebnice a gramatiky

Při tvorbě didaktických materiálů, ať již učebnic, gramatik nebo výkladových slovníků, se korpus může stát pomocníkem, jestliže řešíme, kdy a jaké gramatické jevy prezentovat studentům, nebo když potřebujeme příklady užití (slovní spojení nebo věty).

Kritéria při rozhodování, kdy a jaký gramatický prvek předkládat studentům, mohou být frekvenční, funkční, obsahová, kvalitativní a kvantitativní, synchronní a diachronní apod. Škodová a Štindlová (Škodová, Štindlová, 2007, s. 57) uvádějí jako tři nejdůležitější hlediska: 1. komunikační potenciál daného jevu (jeho využití v praxi), 2. frekventovanost jevu/tvaru v současném jazyce (výzkum v ČNK) a 3. podíl gramatického jevu na budování celkové jazykové kompetence. Hrdlička považuje za hlavní kritérium „podíl na utváření komunikační kompetence mluvčího" (Hrdlička, 2002, s. 74), přičemž komunikační kompetenci definuje jako „schopnost mluvčího úspěšně realizovat svůj komunikační záměr" (Hrdlička, 2002, s. 70). Tuto kompetenci chápe komplexně, nejen jako znalost jazykovou, ale jako souhrn jazykových, kulturních, společenských a situačních kompetencí. Frekvence daného gramatického jevu tedy ani nemůže být jediným hlediskem výběru jevu pro didaktické účely, avšak zcela jistě není nezanedbatelná. Pro jazykové vyučování je podstatné i rozdělení korpusů na psané a mluvené; frekvenční analýza těchto korpusů umožňuje rozdělení jazykových jevů typických pro mluvenou a psanou komunikaci a vymezení frekventovanosti těchto jevů v jednotlivých komunikátech.

Další možností využití korpusů je získání autentických textů pro výuku. I přes snahu autorů nových učebnic uplatnit ve výuce komunikační přístup, se stále setkáváme s formulacemi vět, které jsou komunikačně nepřirozené, a s vykonstruovanými texty zaměřenými na prezentaci vybraného penza gramatických jevů. Pouze materiály pro vysoce pokročilé studenty (B2, C1) zahrnují autentické texty, pro výuku nižších úrovní se převážně využívají pouze texty uměle vytvořené. Přitom student se s autentickými texty setkává každý den a i s minimální znalostí češtiny se v nich musí naučit orientovat. Samozřejmě, že simplifikace jazyka podle dané jazykové úrovně referenčního rámce je na místě, zvláště u tak vysoce flektivního jazyka, jakým je čeština. Otázkou však zůstává, zda zjednodušovat texty na úkor přirozenosti a také, zda není vhodnější, aby si student zvykl porozumět větě nebo textu, které obsahují i jiné pády a tvary, než které si už osvojil.

Pokud bychom pominuli výše zmíněné argumenty pro využití korpusu, stále zůstane jedna významná možnost jeho aplikace, kterou je výuka lexika, speciálně kolokací, které student nenajde v běžných překladových slovnících.

8.2. Data Driven Learning: Student jako výzkumník a objevitel

Jak tedy otevřít korpusu dveře do jazykové učebny? Vypůjčme si základní premisu zakladatelemetody Data Driven Learning, Tima Johnse, který ukázal důležitost korpusové analýzy: „*Research is too serious to be left to the researchers.*" (Johns, 1991). Jinými slovy, student jazyka také de facto provádí jazykový výzkum, stejně jako lingvisté objevuje pravidla a vzory jazykového užití. Johns navrhuje, aby studenti byli vedeni ke zkoumání, podobně jako když lingvisté zjišťují o svém mateřském jazyce fakta, která nebyla doposud objevena.

Obrat od deduktivní výuky směrem k induktivní zahrnuje celou řadu změn týkajících se role studenta a učitele. Deduktivní přístup je typický pro gramaticko-překladovou metodu, ve smyslu představení pravidla a následného procvičování. V induktivní výuce už učitel není prezentujícím odborníkem, ale organizátorem výuky a partnerem studenta. Student se naopak neučí o jazyce, ale učí se, jak se učit, a to díky cvičením, která vyžadují pozorování a interpretaci konkrétního užití jazyka (Bernardini, 2009).

DDL souvisí se současným pojetím jazykové výuky. Komunikační metoda, nebo řekněme její jednotlivé přístupy, došly k pochopení toho, že jazyk není jen sadou fonologických, gramatických a lexikálních pouček určených k zapamatování, ale živým organismem. Studenti cizího jazyka se již neučí o jeho jednotlivých principech proto, že tyto principy v daném jazyce existují, ale proto, že je mohou využít pro komunikaci v reálném životě (Nunan, 2006, s. 6–10). Gramatika je tak prezentována prostřednictvím nejrůznějších témat odrážejících reálný život, s čímž souvisí i zaměření na cílovou skupinu: studenty a jejich jazykové potřeby. Současné komunikační přístupy se vyznačují orientací na studenta, jeho potřeby a požadavky, učitel ustupuje do pozadí (Škvorová, 1992). Neznamená to však, že by se vyučovalo pouze to, co by student chtěl. Ale ani ztrátu funkční pozice učitele ve výuce. Učitel se z dominantní role přesouvá do role organizátora, poradce a také toho, kdo studenty pro činnost motivuje.

8.3. Typy cvičení vhodných pro češtinu

V následujícím oddíle uvedeme ukázky práce s korpusem připravené pro výuku češtiny jako druhého/cizího jazyka. Nejpodstatnější fází práce s ČNK byla předcházející příprava konkrétních úkolů. První otázkou bylo, jakou měrou zjednodušit pro studenty vyhledávání v korpusovém materiálu. Vzhledem k nárokům kladeným na uživatele při práci s vyhledávačem Bonito, co se týče znalosti gramatiky a lingvistic-

ké terminologie, jsme se nejdříve zaměřili na průzkum učebnic češtiny pro cizince z hlediska užívání jazykovědné terminologie, kterou by měli žáci ovládat.

Zjistili jsme, že v učebnicích se obvykle nachází směs českých termínů a mezinárodních ekvivalentů latinského původu. Toto nekonzistentní používání termínů v materiálech pro cizince dobře ilustruje např. nejčastější popis slovních druhů pojmy: *substantiva, adjektiva, zájmena, číslovky, slovesa, adverbia, prepozice, spojky.*[193] Mnoho učebnic také využívá angličtinu jako mediační jazyk, proto vysvětluje gramatiku nebo podává instrukce v angličtině. České termíny mohou některým studentům pomoci v pochopení významu termínu (např. *ženský rod – žena, množné číslo – mnoho*), pro některé studenty, hlavně z evropského areálu, jsou naopak srozumitelnější výrazy mezinárodní (např. femininum, plurál). Při srovnání s Manuálem ke korpusovému manažeru Bonito, kde je základní terminologie uvedena mezinárodními termíny (např. v případě slovních druhů, rodu, čísla apod.) a specifické podskupiny pouze českými výrazy (např. vztažná zájmena, číslovky násobné apod.), je zřejmé, že dotazy pro studenty je nutné co nejvíce zjednodušit takovým způsobem, aby využívali co nejmenší počet termínů (Osolsobě, Vališová, 2010).

Kvůli maximálnímu zjednodušení vyhledávání, tzn. tak, aby byly dotazy vhodné i pro uživatele bez zkušeností s využíváním korpusu, jsme se rozhodli aplikovat vyhledávání převážně pomocí *word* (viz otázky v tabulkách 1, 3 a 5). Znalost jednotlivých *tagů* jsme omezili na pasivní znalost, tj. poznání pádu, rodu a čísla při zobrazení *tag*. Kromě vyhledávání slova studenti používali pouze zobrazení *lemma* a *tag* a vyhledávání slova v různých korpusech. Vypracovali jsme dvě sady otázek – pro méně a více pokročilé, přičemž typy dotazů byly stejné a lišily se pouze úrovní češtiny (srov. tabulky č. 1, 3, 5 a 7). První skupina studentů byla na úrovni A1 až A2 podle SERR, druhá skupina na úrovni B2 až C1. Všichni studenti pracovali s korpusem poprvé. Pro studenty bylo nutné zpracovat manuál využívající množství obrázků a popisující jednoduchou češtinou vyhledávání určitého typu krok za krokem. Manuál jsme však spíše považovali za pomůcku sloužící k opětovnému vyhledávání při samostudiu.

Konkrétní úkoly k vyhledávání jsme rozdělili na čtyři okruhy. Studenti měli za úkol: 1. vyhledat základní tvar, 2. zjistit, jaký je rod, číslo nebo pád daného slova, 3. vyhledat frekvenci a rozhodnout, který ekvivalent patří do mluvené a který do psané češtiny, 4. zjistit kolokabilitu daného slova. V tabulkách č. 1–8 uvádíme příklady otázek.

Největšímu zájmu se těšily první typy úkolů, týkající se lemmatu, zřejmě i díky nejjednoduššímu hledání. Např. studenti dostali sadu minulých příčestí: *šel, jel* a *jedl* a jejich úkolem bylo zjistit základní tvary těchto sloves, tj. infinitivy. Využívali tedy korpus jako on-line slovník, který jim sice neposkytne překlad do jejich mateřského jazyka, ale v případě, že neznají gramatický tvar daného slova, mohou v korpusu zjistit jeho tvar základní, tj. nominativ nebo infinitiv. Základní tvar neboli lemma pak již bez problémů vyhledají ve svém dvojjazyčném slovníku. Pokud totiž z důvodu neznalosti deklinačních paradigmat či z důvodu alternací v základu slova nejsou schopni utvořit základní tvar, není pro ně ani možné dané slovo ve slovníku vyhledat. Tabulka č. 1 uvádí otázky prvního okruhu a tabulka č. 2 příklad vyhledávání, tedy konkrétní konkordanční řádky, které studenti uvidí.

Tabulka 1: Vyhledávání základního tvaru

Otázky pro začátečníky	Otázky pro pokročilé
Jaké jsou infinitivy od tvarů **jel, jedl** a **šel**?	Jaké jsou infinitivy od tvarů **vyňat** a **najat**?
Jaký je nominativ slova **psem**?	Jaký je nominativ slova **břiše**?

Tabulka 2: Příklad

otravuje tak pozdě v noci ? „ zaklel hajný a	< **šel/jít**>	k telefonu . Zvedl sluchátko a chtěl se pře
u mu oči a červený že je jak tulipán . Tak si	< **šel/jít**>	raději lehnout . Ráno bývá moudřejší veče
No jó , ty abys nerýpal , „ odtušil Tyčinka a	< **šel/jít**>	se uklidňovat před další disciplinou do úst
chod , „ zamumlal jenom . Do schodů jsem	< **šel/jít**>	zvolna , jelikož jsem si myslel , že nebude
m . Takže . . . v pátek měl volno a tvrdí , že	< **šel/jít**>	spát hrozně brzy . Prý už kolem osmé . Na
ký policajt a řekl mi , abych šel s ním , a já	< **šel/jít**>	před ním na chodník a představte si , že ta

Druhý okruh otázek, při kterých studenti vyhledali slovo, poté si zobrazili *tag* a zjistili jeho pád, rod nebo číslo, supluje to, co chybí běžnému dvojjazyčnému slovníku. Korpus by se tedy mohl stát jakýmsi doplňkem slovníku. Dalším pozitivem využití korpusu je to, že při zobrazení *tagu* studenti mohou zjistit další gramatické kategorie slova, popř. podle kontextu je možné i odhadnout jeho význam. V tomto

[193] Zkoumali jsme terminologii v těchto učebnicích:
Adamovičová, A. – Ivanovová, D. 2006. *Basic Czech I., II.* Praha: Karolinum.
Bischofová, J. a kol. 2008. *Čeština pro středně a více pokročilé.* Praha: Karolinum.
Čechová, E. – Remediosová, H. 2005. *Chcete mluvit česky?* Liberec: Harry Putz.
Froulíková, L. 2008. *Adam a Eva v Českém ráji.* Praha: Academia.
Froulíková, L. 2002. *Zahrada českého jazyka.* Praha: Academia.
Holá, L. 2000. *Czech Step by Step.* Havlíčkův Brod: Fragment.
Holá, L. 2006. *New Czech Step by Step.* Praha: Akropolis.
Holá, L. – Bořilová, P. 2009. *Česky krok za krokem 2.* Praha: Akropolis.
Hronová, K. 1998. *Čeština pro cizince.* Plzeň: Fraus.
Matula, O. 2007. *Český den. Kurz českého jazyka pro azylanty navazující na Manuál pro učitele českého jazyka pro cizince bez znalosti latinky.* Praha: Člověk v tísni o.p.s., Projekt Varianty.
Nývltová, D. – Štindlová, B. 2008. *Česky v Česku I., II.* Praha: Akropolis.
Parolková, O. 2004. *Czech for foreigners.* Praha: Bohemika.
Pintarová, M. – Režková, I. 1995. *Communicative Czech. Elementary Czech.* Praha: Univerzita Karlova.
Pintarová, M. – Režková, I. 1999. *Communicative Czech. Intermediate Czech.* Praha: Univerzita Karlova.
Štindl, O. 2008. *Easy Czech. Elementary.* Praha: Akronym.
Váchalová, S. 2003. *Survival Czech.* Voznice: Leda.

spatřujeme jeho největší přínos pro přímou práci s vyhledávačem, neboť vyhledat slovo a zobrazit si *lemma* či *tag* zvládnou po krátkém procvičení i začátečníci. A pro ně, protože ještě neumí vytvořit základní tvary slov, je korpus v tomto smyslu obzvláště přínosný. Studenti se navíc nemusí učit novou terminologii, postačí, když ví, že např. na třetí pozici je rod. Písmena M, I, F a N označující všechny rody v češtině, jež studenti znají, neboť se používají v téměř všech učebnicích češtiny. Podobně je to s pády, pokud se studenti učí podle mezinárodní terminologie (Nom, Gen, Dat apod.), není obvykle náročné přiřadit pády k číslům 1 až 7, protože v tomto pořadí bývají pády v tabulkách také ve většině učebnic.

Tabulka 3: Jaký je rod, číslo nebo pád?

Otázky pro začátečníky	Otázky pro pokročilé
Jaký rod má slovo **postel**?	Jaký rod má slovo **téma**?
Jaký rod má slovo **centrum**?	Jaký rod má slovo **pyré**?
Jaký rod má slovo **chleba**?	Jaký rod má slovo **noc**?

Tabulka 4: Příklad

o středních školách, a o chvíli, kdy	< **postel/NNFS1---A----** >	hořela , když jsem si do ní lehal , a
mraky, mu vrhl oknem světlo na	< **postel/NNFS1---A----** >	Uvědomil si , že zapomněl otevřít
pokoje, svlékl sako a natáhl se na	< **postel/NNFS1---A----** >	Teplý vzduch , vanoucí otevřeným
vřené. Zakryl baterku prsty, aby na	< **postel/NNFS1---A----** >	dopadal jen úzký paprsek světla ,
armáda stínů její vlastní smrtelnou	< **postel/NNFS1---A----** >	To nebezpečí bylo reálné , bez ohl
; nebral konce , na němž by čekala	< **postel/NNFS1---A----** >	Chvíli poté , co George Smiley od

Třetí sada otázek se týkala psané a mluvené češtiny. Studenti slova vyhledávali v reprezentativních korpusech SYN2000 a ORAL2008. Vhodné jsou zvláště díky tomu, že korpus psané češtiny SYN2000 obsahuje jen malý podíl beletrie, a tudíž se hodí pro výuku cizinců nejvíce, zatímco ORAL2008 je sociolingvisticky vyvážený korpus mluveného jazyka. Nevýhodou ORALu je jeho velikost – pouhý 1 milion slov. Z tohoto důvodu se často stává, že hledané slovo má velmi malý výskyt, přestože je poměrně frekventované, nebo se v korpusu nevyskytuje vůbec. Pokud se však slovo vyskytuje v obou korpusech, výsledky jsou velmi signifikantní, a toto vyhledávání se hodí např. pro vyhledávání spisovných a obecněčeských tvarů. Student se tak může rozhodnout pro vhodnější tvar při psaní eseje nebo dopisu.

Tabulka 5: Frekvence v mluvené a psané češtině

Otázky pro začátečníky	Otázky pro pokročilé
Jaké slovo má větší frekvenci: **mohu** nebo **můžu**? mluvená čeština: psaná čeština:	Jaké slovo má větší frekvenci: **sousedi** nebo **sousedé**? mluvená čeština: psaná čeština:

Jaké slovo má větší frekvenci: **děkuju** nebo **děkuji**? mluvená čeština: psaná čeština:	Jaké slovo má větší frekvenci: **otci** nebo **otcové**? mluvená čeština: psaná čeština:
Jaké slovo má větší frekvenci: **brzo** nebo **brzy**? mluvená čeština: psaná čeština:	Jaké slovo má větší frekvenci: **rybníkách** nebo **rybnících**? mluvená čeština: psaná čeština:

Tabulka 6: Příklad

SYN2000	ORAL2008
Počet výskytů **1830** > Query : "brzo"	Počet výskytů **83** > Query : "brzo"
Počet výskytů **9799** > Query : "brzy"	Počet výskytů **19** > Query : "brzy"

Poslední typ úkolů, vyhledávání kolokací, však vyžadoval mnohem větší úsilí. Informace nebyly na hlavní straně, bylo nutné v manažeru Bonito kliknout na horní liště na Konkordance, poté vybrat Statistiky a nakonec vybrat možnost Kolokace, což není jednoduchá cesta, pokud člověk nepoužívá korpus denně. Také množství čísel ve statistických údajích bylo pro studenty matoucí. Pokus alespoň částečně využít vyhledávání podle tagu (např. u otázky: *Jaké prefixy se pojí se slovesem konat*?), byť s pokročilými studenty, byl zcela neúspěšný. Byli zmatení nejen z morfologických značek, ale i z množství dat, zvláště statistických informací, ve kterých se nedokázali zorientovat.

U některých úkolů však stačilo prohlédnout jen několik konkordančních řádků, abychom našli odpověď. Například pokud dáme do vyhledávače sloveso *zajímat se*, většina vět obsahuje předložku *o*, a je tedy i pro začátečníky ihned zřejmé, jaká předložka se s tímto slovesem používá. Důležitý je tedy správný výběr konkordancí, které čteme. Pokud řešení úkolu není tak zřetelné jako v příkladu v tabulce č. 8, je nutná asistence učitele, který upozorní na konkordanční řádky skrývající důležitá data nebo odpověď.

Tabulka 7: Kolokabilita

Otázky pro začátečníky	Otázky pro pokročilé
Jaká je prepozice u slovesa **zajímat se**?	Jaký pád se pojí se slovesem **rozumět**?
Jaká je prepozice u slovesa **těšit se**?	Jaká předložka se nejčastěji pojí se slovesem **ptát se**?

Tabulka 8: Příklad

„on je džentlmen! On je Angličan!	< **Zajímá se**>	o vás! Jste nemocná , madame , celá
snadnější zkoušet rozumně mluvit a	< **zajímat se**>	o okolí. Sestra Markhamová se na ni
živý, nudný patron, nemá prý cenu	< **zajímat se**>	o něj. Myslela jsem na to , že se mu
do malé kanceláře a usadili se. „	< **Zajímáme se**>	o knihu, kterou jste si vypůjčila ze
ko kapsy. Na tom není nic špatného,	< **zajímat se**>	, zda máme na něco nárok. “ „ No, mě
zájmu mluvíte?“ „O svém vlastním,	< **Zajímám se**>	se o to z osobních důvodů a pan „ O
		o to z osobních důvodů a pan Wolfe

8.4. Zpětná vazba studentů

Výzkum byl proveden mezi vysokoškolskými studenty v přípravných ročních kurzech pro rusky mluvící na VUT v Brně (6 pokročilých studentů – úroveň B2), mezi studenty Evropských studií na univerzitě v Magdeburgu, kteří mají češtinu jako výběrový cizí jazyk (9 mírně pokročilých studentů – úroveň A2), a také mezi několika studenty češtiny ve firemních kurzech jedné IT společnosti v Brně (5 studentů – čtyři vysoce pokročilí – úroveň B2–C1, a jeden začátečník – A1). Všichni studenti věnovali korpusu jednu dvouhodinovou lekci. Zdůrazňujeme, že žádní z nich nebyli bohemisté, ale studenti učící se pouze praktický jazyk, neboť chceme poukázat na výhody využívání korpusu pro výuku nelingvistů.

Při prvotním vysvětlení, co korpus je, jak se dá využívat, a po krátkém předvedení, se zdvihla vlna nadšení, která většinou rychle opadla, jakmile se studenti začali pokoušet vyhledávat sami. Ideálním postupem při práci s korpusem byla samostatná práce studentů, avšak s asistencí učitele. Největším problémem při korpusovém vyhledávání bylo samotné připojení k internetu a nainstalování manažeru Bonito. Pokud bylo vše připraveno a s využitím učitelova vedení i vyhledáno, dalším úskalím bylo též velké množství dat (mnoho vět, neznámá slova), ve kterých se studenti ztráceli a opět vyžadovali učitelovo vedení, aby byli schopni se v konkordančních řádcích zorientovat. Většina studentů nakonec řekla, že se jim práce s korpusem líbila a že budou korpus využívat i v budoucnu pro studium češtině. Zda tomu tak je, však již nelze ověřit. Přesvědčili jsme se však tímto pokusem, že korpus pro studenty přitažlivý je, a na konkrétních cvičeních jsme ukázali, že data-driven úkoly jsou aplikovatelné i na češtinu. Jak překonat technické obtíže, aby se využívání korpusu stalo běžným, je tedy tématem do budoucna.

V tomto bodě je nutné poukázat na velký potenciál nového nástroje SyD – Korpusový průzkum variant,[194] který je přístupný on-line bez registrace, díky čemuž odpadá řada překážek. Také uživatelské prostředí je velice přátelské a vyhledávání nenáročné. První strana se podobá Googlu, obsahuje pouze okna k zadání klíčových slov. Studenti mohou vyhledávat 2–8 variant a po jednom kliknutí se jim zobrazí kolá̌cové grafy informující je o rozložení klíčových slov v psané (SYN2010) a mluvené češtině (ORAL 2006 a 2008). Dále mohou zjistit žánrové rozložení (opět zobrazeno graficky) a snad největší přínos vidíme v zobrazení nejčastějších kolokací pomocí tzv. *word clouds*; frekvence kolokací je zvýrazněna velikostí fontu a také barvou slov. Při kliknutí na kolokaci se zobrazí několik konkordančních řádků a můžeme tak vidět konkrétní větu nebo např. frázi, ve které se vyskytuje dané spojení slov. Nelze rozšířit kontext, což pro cizince však nemusí být nevýhodou. Nejdůležitější informace jsou zobrazeny graficky a více dat pro běžné vyhledávání a učení není potřeba.

Zmiňme také možné úkoly. Studenti mohou v SyDu vyhledávat nejen nejčastější kolokace, slovní spojení a fráze, ale také mohou podle kolokací zobecnit význam a uvědomit si rozdíl ve významovém odstínění dvou slov. Jako příklad uvedeme slova *díky* a *kvůli*. Nejčastější kolokace slova *díky* jsou: *bůh, dotace, podpora, vstřícnost, spolupráce*, zatímco u *kvůli* to je *nedostatek, zranění, podezření, krize* apod. Na první pohled, a to i bez čtení konkordancí, je tedy jasné, které slovo se používá spíše v pozitivním významu a které v negativním. U více pokročilých je pak možné začít třídit kolokace, např. při dotazu: *Co můžeme sebrat?* studenti vyhledávají mezi kolokacemi pouze objekty slovesa. Díky vyhledaným kolokacím, popř. čtením konkordancí, si pak uvědomí různé významy slovesa: *sebrat odvahu* (odvážit se), *sebrat peněženku* (ukrást) apod. Využití nástroje SyD, a to především práce s kolokacemi, je tedy výzvou pro další zkoumání sbližování korpusové lingvistiky a jazykové výuky.

8.5. Korpusová data přímo ve výuce, ale bez počítače

V závěru bychom chtěli shrnout, že největší překážky při korpusovém vyhledávání s využitím programu Bonito se studenty byly technického charakteru (připojení k internetu, instalace Bonita, instalace české klávesnice, náročnost vyhledávání apod.). Dalším negativem tohoto typu práce je ale též množství jazykových dat, která se studentům nabízejí. Studenti by zřejmě museli korpus využívat pravidelně, aby si hledání zautomatizovali – na což však v běžné výuce není čas. V testovacích podmínkách potřebovali stálou asistenci učitele, a to nejen kvůli vyhledávání, ale často byli ztraceni v množství dat a nedokázali selektovat správné příklady.

Důležitá jsou však nesporná pozitiva, která jsme díky práci se studenty prokázali. Dnešní studenti jsou zvyklí na rychlost, pokud něco neví, hledají informace na Googlu nebo v on-line slovnících. V tomto ohledu může být korpus do budoucna klíčovým doplňkem jazykové výuky. Zpětná vazba studentů byla vesměs velice pozitivní. Práce s korpusem je zajímala a bavila. Většině z nich vyhovovaly úkoly, co se týče úrovně, velká většina (15 z 20 studentů) uvedla, že je práce s korpusem zajímavá a že budou používat korpus ke studiu češtiny i v budoucnosti, viz tabulka č. 9.

[194] http://syd.korpus.cz/

Tabulka 9: Zpětná vazba studentů

Je hledání v korpusu těžké?	
Ano	**0** studentů
Ne	**8** studentů
Středně	**12** studentů
Je hledání v korpusu zajímavé?	
Ano	**15** studentů
Ne	**0** studentů
Středně	**5** studentů
Budete používat korpus při učení češtiny?	
Ano	**15** studentů
Ne	**1** student
Nevím	**4** studenti

Jaké je tedy řešení? Nabízí se dvě cesty: vytvořit uživatelsky přátelské prostředí pro snadné vyhledávání, nejlépe bez lingvistické terminologie a statistik (což již poskytuje nový nástroj SyD) nebo využívat při výuce cvičení, která obsahují pouze několik učitelem předem vybraných a případně i vytištěných konkordancí. V žádném případě bychom se však DDT přístupu ve výuce češtiny neměli vzdávat. V rámci výuky je přínosné uplatnit deduktivní i induktivní přístup. Pokud použijeme korpusová data jako doplňkový materiál ke klasické výuce, studenti se setkají s autentickým jazykem, aniž by zapnuli počítač. Jestliže učitel vybere ilustrativní příklady – věty z korpusu, studenti se mohou učit na základě korpusových dat i bez počítače. Konkordance je též možné upravit na modelové věty, tj. zkrátit či upravit je podle lexikografických pravidel, aby výuku neznesnadňovala málo frekventovaná či rušivá slovní zásoba (Vališová, 2011).

V učebnici (v rámci čtení nebo poslechu) se například objeví slovo, které studenti znají, ale v jiném kontextu. Internetové překladové slovníky, např. Slovník.cz nabízejí mnoho ekvivalentů, avšak bez vysvětlení. V takovém případě je vhodné využít ke specifikaci korpus. Podle učitelem předem vybraných vět studenti sami objeví různé významy slova. Viz tabulka č.10 s příklady autentických vět vybraných z korpusu za účelem procvičování polysémie.[195]

[195] Příklady v tabulkách 10 a 11 jsou vytvořeny na základě dat z korpusu SYN2000. Některé z vět v tabulce 10 jsou zjednodušeny.

Tabulka 10: Korpus a čtení/poslech

Jsem žák druhé **třídy**, ale tolik už přece vím.
Na hlavní **třídě** si všiml obchodu s pánskými oděvy.
Objednal si tedy sedadlo v první **třídě**.
My, Angličané střední **třídy.**
Zabočil s autem za hlavní **třídu** a rychle jel k mostu.
Ten nejspíš letí první **třídou**.
Poslouchala jsem pod oknem **třídy** asi deset minut.
Za jejich vilou byla široká **třída**, po které jezdily dvě tramvaje.
Martin vlastně s námi do **třídy** nechodil.
Sedíte v kupé druhé **třídy** s dvojicí neznámých.
Specialista první **třídy.**
Postoupila do poslední **třídy** gymnázia.

Při procvičování psaní nebo mluvení můžeme naopak využít korpus jako zdroj kolokací. Získáme množství dat, která mohou sloužit jako vodítko při produktivních typech úloh, tj. při psaní a mluvení, čímž si studenti rozšiřují slovní zásobu. Viz tabulka č. 11 se seznamem kolokací slova *vlak*.

Tabulka 11: Korpus a psaní/mluvení

Vlak
jede, přijede, zastaví, pojede, vyjíždí, vjíždí, jezdí, stojí, projíždí, odjíždí, rozjíždí se, pohne se, přejede, zmizí, vyjede, rozjede se, vykolejí, vezl, zůstane, uhání, zastavuje, uhání, blíží se, přepravuje, usmrtí
plný, naložený, půlnoční, směřující, jedoucí, tažený, označený, odjíždějící, jedoucí, projíždějící, odvážející, převážející

Na základě vlastní sondy (Vališová, 2011) jsme zjistili, že výukové materiály ve formě KWIC (cvičení na papíře, ale s klíčovým slovem uprostřed a neupravenými větami) jsou spíše vhodné pro vysoce pokročilé studenty jako doplňková cvičení, a to z důvodu obtížné slovní zásoby a také náročné orientace v kontextu, jímž je často neukončená věta. Studenty však bavilo pracovat se seznamy kolokací, např. třídit je podle významu nebo využívat kolokace při komunikačním cvičení. Pokud shrneme naše výsledky, domníváme se, že by bylo vhodné více začlenit do výuky práci s kolokacemi a lexikálními vzory, a v neposlední řadě se také věnovat výukovým materiálům vytvořeným z korpusových dat, ale upraveným, resp. pro snazší porozumění zkráceným na celé věty.

Vytvořit výukové materiály založené na korpusu jako doplňkový materiál k výuce je tedy další výzvou do budoucna. Zatím může korpus sloužit učitelům při vytváření vlastních cvičení. Není žádoucí zavrhovat tradiční materiály ani dělat ze studentů korpusové lingvisty, přesto však můžeme podporovat studenty v tom, aby sami objevovali jazyk na základě autentických dat.

Literatura

ATKINS, S. – CLEAR, J. – OSTLER, N. 1991. *Corpus design criteria.* Dostupné z WWW: http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf

BARTRAM, M. – WALTON, R. 1991 *Correction.* Stuttgart: Klett.

BEDŘICHOVÁ, Z. – ŠEBESTA, K. – ŠKODOVÁ, S. – ŠORMOVÁ, K. 2011. Podoba a využití korpusu jinojazyčných a romských mluvčích češtiny: CzeSL a ROMi. In F. Čermák (ed.). *Korpusová lingvistika Praha 2011.* Svazek 2 – Výzkum a výstavba korpusů. NLN, s. 93–104.

BEDŘICHOVÁ, Z. – ŠEBESTA, K. – ŠORMOVÁ, K. 2011. ROMi – první rozsáhlá databanka romského etnolektu češtiny. In *Lidé města*, 13, s. 160–163.

BELZ, J. – VYATKINA, N. 2005. Learner Corpus Analysis and the Development of L2 Pragmatic Competence in Networked Intercultural Language Study: The Case of German Modal Particles. *Canadian Modern Language Review,* 62, č. 1, s. 17–48.

BERNARDINI, S. 2009. Corpora in the classroom. An overview and some reflections on future developments. In K. AIJMER (ed.). *Corpora and Language Teaching (Studies in Corpus Linguistics).* Amsterdam: John Benjamins, s. 15–36.

BĚLIČ, J. 1950. K otázce vzniku nové spisovné češtiny. *Slovo a slovesnost,* 12, s. 9–15.

BĚLIČ, J. 1955. Nové údobí ve vývoji českého jazyka. *Naše řeč*, 38, s. 129–146.

BĚLIČ, J. 1958. Vznik hovorové češtiny a její poměr k češtině spisovné. In *Československé přednášky pro IV. mezinárodní sjezd slavistů v Moskvě.* Praha: Nakladatelství Československé akademie věd, s. 59–71.

BĚLIČ, J. 1959. Bojujme za upevnění a šíření hovorové češtiny. *Český jazyk a literatura*, 9, s. 433–441.

BĚLIČ, J. – HAVRÁNEK, B. – JEDLIČKA, A. – TRÁVNÍČEK, F. 1961. K otázce obecné češtiny a jejího poměru k češtině spisovné. *Slovo a slovesnost*, 22, s. 98–107.

BĚLIČ, J. – HAVRÁNEK, B. – JEDLIČKA, A. 1962. Problematika obecné češtiny a jejího poměru k češtině spisovné. *Slovo a slovesnost*, 23, s. 108–126.

BIRD, S. – LIBERMAN, M. 1999. Annotation graphs as a framework for multidimensional linguistic data analysis. In *Proceedings of the Workshop „Towards Standards and Tools for Discourse Tagging"*. Association for Computational Linguistics, s. 1–10. Dostupné z WWW: http://www.ldc.upenn.edu/acl/W/W99/W99-0301.pdf

BISCHOFOVÁ, J. – HRDLIČKA, M. 2007. *Sociokulturní minimum pro azylanty.* Praha: Ministerstvo školství, mládeže a tělovýchovy ČR, SOZE.

BOŘKOVCOVÁ, M. 2006. *Romský etnolekt češtiny*. Praha: Signeta.

BRILL, E. 1995. Transformation–based error–driven learning and natural language processing: A case study in part–of–speech tagging. *Computational Linguistics*, 21, č. 4, s. 543–566.

CARLETTA, J. C. 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22, č. 2, s. 249–254.

CARLETTA, J. C. – MCKELVIE, D. – ISARD, A. 2002. Supporting linguistic annotation using XML and stylesheets. In G. SAMPSON – D. MCCARTHY (eds.). *Corpus linguistics: readings in a widening discipline.* London & New York : Continuum Interpretations. Dostupné z WWW: http://homepages.inf.ed.ac.uk/jeanc/revised.but.like.readings-in-corpling.pdf

CARLETTA, J. et al. 2003. The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web* (3rd Workshop on NLP and XML). Budapest.

CORDER, S. P. 1974. Idiosyncratic Dialects and Error Analysis. In J. RICHARDS (ed.). *Error analysis: Perspectives on Second Language Acquisition.* Essex: Longman, s. 158–171.

CORDER, P. 1967. The Significance of Learners' Errors. *International Review of Applied linguistics.* Vol. 5, n. 4, s. 162–169.

CVRČEK, V. 2006. *Teorie jazykové kultury po roce 1945.* Praha: Karolinum.

CVRČEK, V. 2008. *Regulace jazyka a Koncept minimální intervence.* Praha: Nakladatelství Lidové noviny, Ústav Českého národního korpusu.

CVRČEK, V. a kol. 2010. *Mluvnice současné češtiny 1. Jak se píše a jak se mluví.* Praha: Karolinum.

ČERMÁK, F. 1996. Obecná a spisovná čeština: poměr, funkce a metodologie. In ŠRÁMEK, R. (ed.). *Spisovnost a nespisovnost dnes.* Brno: Pedagogická fakulta Masarykovy univerzity, s. 14–18.

ČERMÁK, F. 2005. Korpus, informace a lingvistika. In *Přednášky z XLVIII. Běhu LŠSS UK.* Praha: Karolinum, s. 19–20.

ČERVENKA, J. – KUBANÍK, P. – SADÍLKOVÁ, H. 2010. Sociolingvistický výzkum situace romštiny v České republice. In *Studie z aplikované lingvistiky/ Studies in applied linguistics*, 1, s. 167–170. Dostupné z WWW: http://www.linguistik.uni-kiel.de/sldr/stuff/sldr_mosel_handout.pdf

DANEŠ, F. 1979. Postoje a hodnoticí kritéria při kodifikaci. In J. KUCHAŘ (ed.). *Aktuální otázky jazykové kultury v socialistické společnosti.* Praha: Academia, s. 79–91.

DANEŠ, F. 1988. Pojem „spisovného jazyka" v dnešních společenských podmínkách. In R. BRABCOVÁ – F. ŠTÍCHA (eds.). *Dynamika současné češtiny z hlediska lingvistické teorie a školské praxe.* Praha: Pedagogická fakulta Univerzity Karlovy, s. 21–28.

DANEŠ, F. 1993. Čeština bez příkras a v plné kráse. *Český jazyk a literatura*, 43, s. 180–183.

DANEŠ, F. 1996. Preskripce – anebo „nechte svůj jazyk na pokoji"? In I. NEBESKÁ – A. MACUROVÁ (eds.). *Jazyk a jeho užívání.* Praha: Filozofická fakulta Univerzity Karlovy, s. 166–174.

DANEŠ, F. 1997. Situace a celkový stav dnešní češtiny. In F. DANEŠ a kol., *Český jazyk na přelomu tisíciletí.* Praha: Academia, s. 12–24.

DÍAZ-NEGRILLO, A. – FERNÁNDEZ-DOMÍNGUEZ, J. 2006. Error Tagging Systems for Learner Corpora. *Resla.* 19, s. 83–102.

DÍAZ-NEGRILLO, A. – MEURERS, D. – VALERA, S. – WUNSCH, H. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36, č. 1–2, s. 139–154.

DIPPER, S. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In R. ECKSTEIN – R. TOLKSDORF (eds.). *Proceedings of Berliner XML Tage*, s. 39–50. Dostupné z WWW: http://www.ling.uni–potsdam.de/~dipper/papers/xmltage05.pdf

DOOLITTLE, S. 2009. Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basierenden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenen – Architektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner. Magisterarbeit, HU-Berlin.

DULAY, H. – BURT, M. 1974. You can't learn without goofing. In J. C. RICHARDS (ed.). *Error analysis.* London: Longman, 1974.

DULAY, H. – BURT, M. – KRASHEN, S. 1982. *Language Two.* Oxford: OUP.

EDGE, J. 1994. *Mistakes and Correction.* London – New York: Longman.

ELLIS, R. 1994. *The Study of Second Language Acquisition*. Oxford: OUP.

ELLIS, R. – BARKHUIZEN, G. 2009. *Analysing learner language.* Oxford: OUP.

Encyklopedický slovník češtiny. 2002. KARLÍK, P. – NEKULA, M. – PLESKALOVÁ, J. (eds.). Praha: Nakladatelství Lidové noviny.

FITZPATRICK, E. – SEEGMILLER, M. S. 2004. The Montclair electronic language database project. In CONNOR, U. – UPTON, T. A. (eds.). *Applied Corpus Linguistics: A Multidimensional Perspective.* Rodopi, s. 223–238. Dostupné z WWW: http://chss.montclair.edu/linguistics/MELD

FORD, P. – JOHNSTON, B. – BRUMFIT, Ch. – MITCHELL, R. – MYLES, F. 2005. Practice learning and the development of students as critical practitioners: some findings from research. *Social Work Education*, 24, č. 4, s. 391–407.

GAC spol. s r. o. 2007. *Analýza postojů a vzdělávacích potřeb romských dětí a mládeže.* Dostupné z WWW: www.nros.cz/analyza-postoju-a-vzdelavacich-potreb-romskych-deti-a-mladeze.

GRANATH, S. 2004. Who benefits from learning how to use corpora? In J. Sinclair (ed.). *How To Use Corpora In Language Teaching*. Amsterdam: John Benjamins, s. 47–65.

GRANGER, S. 1993. The international Corpus of Learner English. In ARTS, J. et al. (eds). *English Language Corpora: Design, Analysis and Expoitation.* Amsterodam: Rodopi, s. 57–69.

GRANGER, S. 1998. The computer learner corpus: a versatile new source of data for SLA research. In S. GRANGER (ed.). *Learner English on Computer.* London: Longman, s. 3–19.

GRANGER, S. 1999. Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In H. HASSELGARD – S. OKSEFJELL (eds.). *Out ofCorpora: Studies in Honour of Stig Johansson.* Amsterdam: Rodopi, s. 1991–202.

GRANGER, S. 2003. Error-tagged Learner Corpora and CALL: A PROMising Synergy. *CALICO journal,* 20, č. 3, s. 465–480.

HANA, J. – ROSEN, A. – ŠKODOVÁ, S. – ŠTINDLOVÁ, B. 2010. Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*. Uppsala: Association for Computational Linguistics.

HAUSENBLAS, K. 1962. Styl jazykových projevů a rozvrstvení jazyka. *Slovo a slovesnost*, 23, s. 189–201.

HAUSENBLAS, K. – KUCHAŘ, J. a kol. 1979. *Čeština za školou.* Praha: Panorama.

HAVRÁNEK, B. 1932. Úkoly spisovného jazyka a jeho kultura. In B. HAVRÁNEK – M. WEINGART (eds.). *Spisovná čeština a jazyková kultura.* Praha: Melantrich, s. 32–84.

HENDRICH, J. a kol. 1988. *Didaktika cizích jazyků*. Praha: Státní pedagogické nakladatelství.

HOFFMANNOVÁ, J. 1997. *Stylistika a ... Současná situace stylistiky.* Praha.

MILTON, J. – CHOWDHURY, N. 1994. Tagging the interlanguagae of Chinese learners of English. In L. FLOWEDEW – K. K. TONG (eds). *Entering Text.* Hong Kong: The Hong Kong University of Science and Technology, s. 233–243.

HOMOLÁČ, J. 2009. *Internetové diskuse o cikánech a Romech.* Praha: Karolinum.

HRBÁČEK, J. 1994. *Úvod do studia českého jazyka*. Praha: Karolinum (2. vyd.).

HRBÁČEK, J. 1995. Hovorová čeština. In J. KUKLÍK (ed.). *Přednášky z XXXIV. a XXXV. běhu Letní školy slovanských studií.* Praha: Filozofická fakulta Univerzity Karlovy, s. 53–61.

HRDLIČKA, M. 2002. *Cizí jazyk čeština*. Praha: ISV.

HRDLIČKA, M. 2009. *Gramatika a výuka češtiny jako cizího jazyka. K prezentaci gramatiky českého jazyka v učebnicích češtiny pro cizince.* Praha: Karolinum.

HRDLIČKA, M. 2010. *Kapitoly o češtině jako cizím jazyku.* Plzeň: Vydavatelství ZČU.

HRDLIČKA, M. 1995. *Překladatelské miniatury.* Praha: Karolinum.

IZUMI, E. – UCHIMOTO, K. – ISAHARA, H. 2004. The NICT JLE Corpus Exploiting the language learners' speech diabase for research and education. *International Journal of The Computer, the Internet and Management*, 12, č. 2, s. 119–125.

IZUMI, E. – UCHIMOTO, K. – ISAHARA, H. 2005. Error Annotation for Corpus of Japanese Learner English. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005).* Korea, s. 71–80. Dostupné z WWW: http://acl.ldc.upenn.edu/I/I05/I05-6009.pdf

JAKOBSON, R. 1932. O dnešním brusičství českém. In B. HAVRÁNEK – M. WEINGART (eds.). *Spisovná čeština a jazyková kultura.* Praha: Melantrich, s. 85–122.

JAMES, C. 1998. *Errors in Language Learning and Use.* Longman.

JELÍNEK, M. 2001. Co je to jazyková chyba? In *Profesor Hauser jubilující.* Brno: Pedagogická fakulta Masarykovy univerzity, s. 75–85.

JELÍNEK, M. 2007. Purismus. In J. PLESKALOVÁ – M. KRČMOVÁ – R. VEČERKA – P. KARLÍK (eds.). *Kapitoly z dějin české jazykovědné bohemistiky.* Praha: Academia, s. 540–579.

JELÍNEK, T. 2008. Nové značkování v Českém národním korpusu. *Naše řeč*, 91, s. 13–20.

JELÍNEK, T. – PETKEVIČ, V. 2011. Systém jazykového značkování současné psané češtiny. In *Korpusová lingvistika Praha 2011, sv. 3: Gramatika a značkování korpusů.* Praha: Nakladatelství Lidové noviny/Ústav českého národního korpusu, s. 154–170.

JOHNS, T. 1991. Should You Be Persuaded. Two samples od data-driven learning matherials. In T. JOHNS – P. KING (eds.). *Classroom Concordancing*. English Language Research Journal 4. Birmingham: Birmingham University, s. 1–13.

KOPEČNÝ, F. 1949. Spisovný jazyk a jeho forma hovorová. *Naše řeč,* 33, s. 14–22.

KORČÁKOVÁ, J. 2004. *Chyba a učení cizím jazykům.* Hradec Králové: Gaudeamus.

KUČERA, K. 1990. *Český jazyk v USA.* Praha: Univerzita Karlova.

LARSEN-FREEMAN, D. – LONG, M. H. 1992. *An Introduction to Second Language Acquisition Research.* London/New York : Longman.

LEECH, G. 1997. Introducing corpus annotation. In R. GARSIDE – G. LEECH – A. MCENERY (eds.). *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman, s. 1–18.

LEŃKO-SZYMAŃSKA, A. 2004. Demonstratives as anaphora markers in advanced learners' English. In G. ASTON – S. BERNARDINI – D. STEWART (eds.). *Corpora and Language Learners*. Amsterdam: Benjamins, s. 89–107.

LÜDELING, A. – WALTER, M. – KROYMANN, E. – ADOLPHS, P. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005 Conference, 14–17 July.* Birmingham. Dostupné z WWW: http://www.corpus.bham.ac.uk/pclc/#corpora

LÜDELING, A. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In P. GROMMES – M. WALTER (eds.). *Fortgeschrittene Lernervarietäten*. Niemeyer: Tübingen, s. 119–140.

MACHOVÁ, S. 2000. Dvě předložky vedle sebe. *Naše řeč,* 83, s. 30–34.

MATHESIUS, V. 1932. O požadavku stability ve spisovném jazyce. In B. HAVRÁNEK – M. WEINGART (eds.). *Spisovná čeština a jazyková kultura.* Praha: Melantrich, s. 14–31.

MEURERS, D. 2005. On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua,* 115, č. 11, s. 1619–1639.

MEURERS, D. 2009. On the Automatic Analysis of Learner Language. Introduction to the Special Issue. *CALICO Journal.* 26, č. 3, s. 469–473. Dostupné z WWW: http://www.sfs.uni-tuebingen.de/~dm/papers/meurers-09.pdf

MEYER, Ch. F. 2006. *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.

MUKAŘOVSKÝ, J. 1932. Jazyk spisovný a jazyk básnický. In B. HAVRÁNEK – M. WEINGART (eds.). *Spisovná čeština a jazyková kultura.* Praha: Melantrich, s. 123–156.

MUKHERJEE, J. – ROHRBACH, J. M. 2006. Rethinking applied corpus linguistics from a language-pedagogical perspective: new departures in learner corpus research. In B. KETTEMAN – G. MARKO (eds.). *Planing, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*. Frankfurt/MaIn Peter Lang, s. 205–232.

NICKEL, G. 1989. *Some controversies in present day error analysis. IRAL,* 27, s. 293–305.

NICHOLLS, D. 2003. The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference, 28–31 March.* Lancaster, s. 572–581. Dostupné z WWW: *ucrel.lancs.ac.uk/publications/cl2003/papers/nicholls.pdf*

NORRISH, J. 1987. *Language learners and thein errors.* London: Macmillan.

NUNAN, D. 2006. *Task-Based Language Teaching*. Cambridge: Cambridge University Press.

OSOLSOBĚ, K., VALIŠOVÁ, P. 2010. Tagset korpusů ČNK z hlediska předpokládané znalosti gramatické terminologie u nerodilých mluvčích (Možnosti a meze využívání korpusů češtiny pro nerodilé mluvčí). In I. DOMINIKOVÁ – M. LACHOUT (eds.). *Lingua terminologica*. Praha: MUP, s. 141–156.

PAJAS, P. – ŠTĚPÁNEK, J. 2006. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of LREC 2006 Workshop on Merging and Layering Linguistic Information*. Genoa: ELRA.

PALENČÁROVÁ, J. – ŠEBESTA, K. 2006. *Aktivní naslouchání při vyučování: rozvíjení komunikačních dovedností na 1. stupni ZŠ*. Praha: Portál.

RASTELLI, S. 2009. Learner Corpora without Error Tagging. *Linguistic online,* 38, č. 2. Dostupné z WWW: http://www.linguistik-online.com/38_09/rastelli.html

RASTELLI, S. – FRONTINI, F. 2008. SLA meets FLT research: the form/fiction split in the annotation of Learner Corpora. In *Proceedings of TaLC 8.* Lisabon, s. 446–451.

REUER, V. – KÜHNBERGER, K. U. 2005. Feature Constraint Logic and Error Detection in ICALL Systems. In P. BLACHE – E. STABLER (eds.). *Proceedings of the 5th International Conference on the Logical Aspects of Computational Linguistics (LACL 2005).* Lecture Notes in Artificial Intelligence 3492, Springer, s. 255–270.

REZNICEK, M. – KRUMMES, C. – HIRSCHMANN, H. – LÜDELING, A. – ENSSLIN, A. – CHAN, J. W. – ZELDES, A. – KRAUSE, T. – ZIPSER, F. 2010b. Dass wenn man etwas will, muss man dafür arbeiten – Zielhypothesen im Lernerkorpus Falko1. *31. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Postersession der Sektion Computerlinguistik,* 25. 2. 2010. Dostupné z WWW: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/standardseite#tools

REZNICEK, M. – WALTER, M. – SCHMID, K. – LÜDELING, A. – HIRSCHMANN, H. – KRUMMES, C. 2010a. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 1.0.* Humboldt-Universität zu Berlin.

RICHTER, M. 2010. *Pokročilý korektor češtiny*. Diplomová práce. Praha: Matematicko-fyzikální fakulta Univerzity Karlovy v Praze.

RICHTEROVÁ, K. 2011. Korpus DEAF: psané texty českých neslyšících. *Studie z aplikované lingvistiky/Studies in applied linguistics,* 1, s. 65–78.

RINGBOM, H. 1987. *The role of the first language in foreign language learning.* Clevedon & Philadelphia: Multilingual Matters.

ROGATCHEVA, S. 2009. "I've only found the answer a few days ago:" aspect use in Bulgarian and German EFL writing. In C. PRADO-ALONSO – L. GÓMEZ-GARCÍA – I. PASTOR-GÓMEZ – D. TIZÓN-COUTO (eds.). *New Trends and Methodologies in Applied English Language Research. Diachronic, Diatopic and Contrastive Studies,* Frankfurt: Peter Lang, s. 255–278.

RUSÍNOVÁ, Z. 1995. Spisovná a obecná čeština. In J. JANČÁKOVÁ – M. KOMÁREK – O. ULIČNÝ (eds.). *Spisovná čeština a jazyková kultura 1993.* Praha: Filozofická fakulta Univerzity Karlovy, s. 57–60.

RYCHLÝ, P. 1998–2003. *Bonito – grafické uživatelské rozhraní systému Manatee*, Verze 1.49. Dostupné z WWW: http://korpus.cz/bonito.

SAMARIN, W. J. *Field linguistics. A guide to linguistic fieldwork.* New York.

SEEGMILLER, M. S. – FITZPATRICK, E. 2003. Practical aspects of corpus tagging. In B. Lewandowska-Tomaszczyk (ed.). *Palc 2001: Practical Applications in Language Corpora.* Peter Lang Pub Inc. Dostupné z WWW: http://chss.montclair.edu/linguistics/MELD/Lodzpaper.pdf

SGALL, P. 1960. Obichodno-razgovornyj češskij jazyk. *Voprosy jazykoznanija*, 9, s. 11–20.

SGALL, P. 1996. Uživatel spisovného jazyka a hyperkorektnost. In R. ŠRÁMEK (ed.). *Spisovnost a nespisovnost dnes.* Brno: Pedagogická fakulta Masarykovy univerzity, s. 59–63.

SGALL, P. – HRONEK, J. 1992. *Čeština bez příkras*. Praha: H§H.

SCHACHTER, J. – CELCE-MURCIA, M. 1977. Some reservations concernig error analysis. *TESOL Quarterly,* 11, č. 4, s. 441–451.

SCHMID, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing.* Dostupné z WWW: http://www.ims.uni–stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

SCHMIDT, T. 2004. EXMARaLDA – ein System zur computergestützten Diskurstranskription. In A. MEHLER – H. LOBIN (eds.). *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*. Wiesbaden: Verlag für Sozialwissenschaften, s. 20–218.

SCHMIDT, T. 2001. The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse. In *Proceedings of the IRCS Workshop On Linguistic Databases, 11–13 December 2001.* Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania, s. 219–227. Dostupné z WWW: http://www.exmaralda.org/files/IRCS_Paper.pdf

SELINKER, L. 1972. Interlanguage. *IRAL*, 10, č. 3, s. 209–231.

SCHMIDT, T. 2009. Creating and working with spoken language corpora in EXMARaLDA. *LULCL II: Lesser Used Languages & Computer Linguistics II,* s. 151–164.

SINCLAIR, J. 1995. From theory to practice. In G. LEECH – G. MYERS – J. THOMAS (eds.). *Spoken English on Computer.* Harlow: Longman, s. 99–112.

SINCLAIR, J. 1996. *EAGLES. Preliminary recommendations on Corpus Typology.* EAG–TCWG–CTYP/P. Dostupné z WWW: http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html

SKALKOVÁ, J. 1999. *Obecná didaktika.* Praha: ISV.

Společný evropský referenční rámec pro jazyky. Jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme. Olomouc: Council for Cultural Co-operation, Education Committee, Modern Languages Division, Univerzita Palackého v Olomouci.

STARÝ, Z. 1995. *Ve jménu funkce a intervence.* Praha: Karolinum.

STEFANOWITSCH A. – GRIES, S. TH. 2006. *Corpora in Cognitive Linguistics.* Berlin: Mouton de Gruyter.

STICH, A. 1969. Současné úkoly jazykové kultury. *Naše řeč,* 52, s. 155–166.

STICH, A. 1979. K pojmu jazykové kultury a jeho obsahu. In J. KUCHAŘ (ed.). *Aktuální otázky jazykové kultury v socialistické společnosti.* Praha: Academia, s. 98–108.

STRITAR, M. 2009. Slovene as a Foreign Language: The Pilot Learner Corpus Perspective. *Slovenski jezik – Slovene Linguistic Studies,* 7, s. 135–152. Dostupné z WWW: http://kuscholarworks.ku.edu/dspace/bitstream/1808/5274/1/8Stritar.pdf

STÜHRENBERG, M. et al. 2006. Multidimensional markup and heterogeneous linguistic resources. In *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing.* Trento: Italy. Dostupné z WWW: http://www.aclweb.org/anthology/W/W06/W06-2715.pdf

ŠATAVA, L. 2001. *Jazyk a identita etnických menšin: možnosti zachování a revitalizace.* Praha: Cargo Publishers.

ŠEBESTA, K. 2011. Akviziční korpusy. In *Minulost, přítomnost a budoucnost v jazyce a v literatuře. Ústí nad Labem 1.–3. 9. 2010.* Ústí nad Labem: PF UJEP.

ŠEBESTA, K. 2010. Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky/Studies in applied linguistics,* 2, s. 11–33.

ŠEBESTA, K. 1999. *Od jazyka ke komunikaci.* Praha: Karolinum.

ŠKODOVÁ, S. – ŠTINDLOVÁ, B. 2007. Modifikace principů přímé metody pro potřeby výuky gramatiky češtiny jako cizího jazyka. In J. ČEMUSOVÁ – B. ŠTINDLOVÁ (eds.). *Sborník Asociace učitelů češtiny jako cizího jazyka (AUČCJ) 2006–2007.* Praha: Akropolis.

ŠKODOVÁ, S. – ŠTINDLOVÁ, B. – HANA, J. – ROSEN, A. 2011. Víceúrovňová anotace českého žákovského korpusu. In *Korpusová lingvistika Praha 2011, sv. 3: Gramatika a značkování korpusů.* Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu, s. 208–225.

ŠKVOROVÁ, D. K principům komunikativní metody ve vyučování jazyků. *Časopis pro moderní filologii.* 1992, 74, č. 1, s. 89–95.

ŠOTOLOVÁ, E. 2008. *Vzdělávání Romů.* Praha: Nakladatelství Karolinum.

ŠTÍCHA, F. 1995. K pojetí spisovnosti. In J. JANČÁKOVÁ – K. KOMÁREK – O. ULIČNÝ (eds.). *Spisovná čeština a jazyková kultura 1993.* Praha: Filozofická fakulta Univerzity Karlovy, s. 57–60.

ŠTINDLOVÁ, B. 2011. Evaluace chybové anotace navržené pro žákovský korpus češtiny. SALi, 2011, č. 2, s. 37–60. Praha: FF UK.

ŠTINDLOVÁ, B. 2011. *Manuál pro přepis psaných materiálů.* Dostupné z WWW: http://utkl.ff.cuni.cz/~rosen/public/transkripce.pdf, http://utkl.ff.cuni.cz/~rosen/public/transkripce_doplnek.

ŠTINDLOVÁ, B. 2011. *Evaluace chybové anotace v žákovském korpusu češtiny.* Disertační práce. Praha: Filozofická fakulta Univerzity Karlovy v Praze.

ŠTINDLOVÁ, B. – ŠKODOVÁ, S. – HANA, J. – ROSEN, A. 2011. CzeSL – an error tagged corpus of Czech as a second language. PALC 2011 – Practical Applications in Language and Computers, Lodž 13.–15. dubna 2011. Výběr z příspěvků vyjde v nakladatelství Peter Lang v edici Łódź Studies in Language.

ŠTINDLOVÁ, B. – ROSEN, A. 2012. *Návod k anotaci chybového korpusu,* verze 5. Dostupné z WWW: http://utkl.ff.cuni.cz/~rosen/public/anotace.pdf.

THOMAS, J. 2006. Using Corpora in Language Teaching and Learning. In *Teaching English with Technology, A Journal for Teachers of English,* 6, č. 1.

TOGNINI-BONELLI, E. 2001. *Corpus Linguistics at Work.* Amsterdam/Philadelphia: Benjamins.

TONO, Y. 2003. Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 Conference.* Lancaster: United Kingdom, s. 800–809. Dostupné z WWW: http://www.scribd.com/doc/8254550/Learner-Corpora

TONO, Y. 1999. Using Learner Corpora in ELT and SLA Research. Paper presented at the *Symposium on the Roles of Corpora in Language Teaching and Language Engineering of the 12th World Congress of Applied Linguistics* (AILA), 1–6 August 1999, Tokyo, Japan.

ULIČNÝ, O. 1994. K teorii mluveného jazyka. In D. Davidová (ed.). *K diferenciaci současného mluveného jazyka.* Ostrava: Filozofická fakulta Ostravské univerzity, s. 19–25.

ULIČNÝ, O. 1996. Čeština devadesátých let dvacátého století. In R. ŠRÁMEK (ed.). *Spisovnost a nespisovnost dnes*. Brno: Pedagogická fakulta Masarykovy univerzity, s. 59–63.

ULIČNÝ, O. 1998. K článku prof. P. Sgalla „Neochuzujme spisovnou češtinu". *Český jazyk a literatura*, 49, s. 35–39.

VAN ELS, T. – BONGAERTS, T. – EXTRA, G. – VAN OS, C. – JANSSEN-VAN DIETEN, A. 1984. *Applied linguistics and the learning and teaching of foreign languages.* London: Edward Arnold.

VAN ROOY, B. – SCHÄFER, L. 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In D. ARCHER – R. RAYSON – A. WILSON – T. McENERY (eds.). *Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), 28–31 March 2003.* Lancaster: UCREL, Lancaster University, s. 835–844. Dostupné z WWW: http://www.corpus4u.org/upload/forum/2005092023174960.pdf

VALIŠOVÁ, P. 2009. *Korpus jako zdroj systémového popisu české konjugace v učebnicích češtiny jako cizího jazyka*. Diplomová práce na FF MU. Dostupné z WWW: https://is.muni.cz/auth/th/75420/ff_m_b1/?fakulta=1421;obdobi=4703;studium=499045

VALIŠOVÁ, P. 2011. Výukové materiály založené na korpusu. In Čermák, F. *Korpusová lingvistika Praha 2011. 2 Výzkum a výstavba korpusů*. Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu, s. 313–323.

WAIBEL, B. 2008. *Phrasal verbs. German and Italian learners of English compared.* Saarbrücken: VDM.

WEINBERGER, U. 2002. *Error Analysis with Computer Learner Corpora: A corpus-based study of errors in the written German of British University Students.* Diplomová práce. Lancaster University.

WIBLE, D., at all. 2003. A webbased EFL writing environment: integrating information for learners, teachers, and researchers. In WIDDOWSON, H. *Defining Issues in English Language Teaching*. Oxford: OxfordUniversity Press, s. 297–315.

ZELDES, A. – HIRSCHMANN, H. – LÜDELING, A. 2009. Multilevel Learner Corpora. *Workshop on Automatic Analysis of Learner Language, 10–14 March 2009 (AALL'09), CALICO ,09*. Arizona State University.

Český národní korpus

Český národní korpus – ORAL2008. (2008) Praha: Ústav Českého národního korpusu FF UK. Dostupné z WWW: http://korpus.cz

Český národní korpus – SYN2000. (2000) Praha: Ústav Českého národního korpusu FF UK. Dostupné z WWW: http://korpus.cz

Český národní korpus – Korpus SyD. Korpusový průzkum variant. (2010) Praha: Ústav Českého národního korpusu FF UK. Dostupné z WWW: http://syd.korpus.cz

Český národní korpus (2000–2010). Praha: Ústav Českého národního korpusu FF UK. Dostupné z WWW: http://korpus.cz

Medailonky autorů

Mgr. Zuzanna Bedřichová (doktorandka Ústavu českého jazyka a teorie komunikace Filozofické fakulty Univerzity Karlovy) se věnuje využití jazykových korpusů ve výuce, zejména pak zkoumání chybovosti romských žáků na základě ROMi, což je také téma její disertační práce. V projektu CZ.1.07/2.2.00/07.0259 je koordinátorkou sběru a zpracování jazykových dat, podílela se zejména na vytváření databanky ROMi.

RNDr. Milena Hnátková, CSc., (Ústav teoretické a komputační lingvistiky Filozofické fakulty Univerzity Karlovy) je matematická lingvistka specializující se na oblast kolokací v češtině. V projektu CZ.1.07/2.2.00/07.0259 se zabývá řízením anotátorů a lingvistickou anotací žákovských textů.

Doc. PhDr. Milan Hrdlička, CSc., (Ústav bohemistických studií Filozofické fakulty Univerzity Karlovy v Praze, Katedra českého jazyka a literatury Fakulty pedagogické Západočeské univerzity v Plzni) je bohemistou, jehož hlavní vědecký zájem se soustředí na výuku češtiny pro cizince a otázky lingvodidaktického popisu a prezentace češtiny pro jinojazyčné mluvčí. V projektu CZ.1.07/2.2.00/07.0259 zastřešuje sběr textového materiálu pro žákovský korpus.

Mgr. Tomáš Jelínek (Ústav teoretické a komputační lingvistiky Filozofické fakulty Univerzity Karlovy) je matematický lingvista, který se zaměřuje zejména na oblast automatické syntaktické analýzy a morfologické disambiguace českých korpusových textů. V projektu CZ.1.07/2.2.00/07.0259 je autorem komplexu programů pro automatickou identifikaci chyb v žákovských textech.

Mgr. Petr Jäger (Ústav teoretické a komputační lingvistiky Filozofické fakulty Univerzity Karlovy) je programátorem lingvistického softwaru. V projektu CZ.1.07/2.2.00/07.0259 je mj. autorem systému Speed pro řízení a správu zpracovávaných žákovských textů.

Doc. RNDr. Vladimír Petkevič, CSc., (Ústav teoretické a komputační lingvistiky Filozofické fakulty Univerzity Karlovy) je matematický lingvista, který se zabývá především morfologickým a syntaktickým značkováním českých korpusových textů. V projektu CZ.1.07/2.2.00/07.0259 se zabýval především organizací a řízením prací na přípravě korpusu emendovaných a lingvisticky značkovaných textů.

Ing. Alexandr Rosen, Ph.D., (Ústav teoretické a komputační lingvistiky Filozofické fakulty Univerzity Karlovy) je matematický lingvista, který je specialistou na gramatické formalismy, teorii syntaktické analýzy a paralelní korpusy. V projektu CZ.1.07/2.2.00/07.0259 je zejména spoluautorem chybové taxonomie a manuálů pro anotaci a přepis žákovských textů.

Prof. PhDr. Karel Šebesta, CSc., (Ústav českého jazyka a teorie komunikace Filozofické fakulty Univerzity Karlovy, Praha; Katedra českého jazyka a literatury, Fakulta přírodovědně-humanitní a pedagogická, Technická univerzita v Liberci), lingvista, bohemista, badatelsky a pedagogicky se věnuje aplikované lingvistice a didaktice jazyka. Řídí budování souboru akvizičních korpusů AKCES, v projektu CZ.1.07/2.2.00/07.0259 je hlavním koordinátorem.

Mgr. Svatava Škodová, Ph.D., (Katedra českého jazyka a literatury, Fakulta přírodovědně-humanitní a pedagogická, Technická univerzita v Liberci) je bohemistka, jejíž hlavní specializací je výuka češtiny pro cizince a žákovské korpusy češtiny. V projektu CZ.1.07/2.2.00/07.0259 byla garantem inovací studia v oboru čeština jako cizí jazyk a podílela se na vytváření chybové taxonomie.

Mgr. Kateřina Šormová (Ústav českého jazyka a teorie komunikace Filozofické fakulty Univerzity Karlovy) zpracovává disertační práci na téma čtenářská gramotnost romských žáků. V projektu CZ.1.07/2.2.00/07.0259 je koordinátorkou rozšiřování a emendace jazykové databanky ROMi.

Mgr. Barbora Štindlová, Ph.D., (Katedra českého jazyka a literatury, Fakulta přírodovědně-humanitní a pedagogická, Technická univerzita v Liberci) je bohemistka, jejíž hlavní specializací je výuka češtiny pro cizince a žákovské korpusy češtiny. V projektu CZ.1.07/2.2.00/07.0259 je spoluautorkou chybové taxonomie a manuálů pro anotaci a přepis žákovských textů.

Mgr. Pavlína Vališová (Ústav českého jazyka, Filozofická fakulta Masarykovy univerzity, Brno) je bohemistkou, jejíž odborný zájem se soustředí na korpusovou lingvistiku. Zabývá se využitím korpusů při výuce češtiny jako cizího jazyka, a to jak pro tvorbu výukových materiálů z autentických textů, tak i pro práci studentů-cizinců přímo s korpusovými nástroji.

Název	Čeština – cílový jazyk a korpusy
Editoři	Karel Šebesta, Svatava Škodová
Určeno pro	studenty bohemistiky a obecné lingvistiky
Vydavatel	Technická univerzita v Liberci
Schváleno	Rektorátem TU v Liberci dne 26. 4. 2012, čj. RE 30/12
Vyšlo	v květnu 2012
Počet stran	168
Vydání	první
Tiskárna	Vysokoškolský podnik Liberec, s.r.o., Hálkova 6, Liberec
Číslo publikace	55-028-12

Tato publikace neprošla redakční ani jazykovou úpravou.

ISBN 978-80-7372-848-9