Univerzita Karlova v Praze

Filozofická fakulta

Ústav českého jazyka a teorie komunikace

Filologie – Český jazyk

Barbora Štindlová

EVALUACE CHYBOVÉ ANOTACE V ŽÁKOVSKÉM KORPUSU ČEŠTINY

Evaluation of Error Mark-Up in a Learner Corpus of Czech

Disertační práce

Vedoucí práce – prof. PhDr. Karel Šebesta, CSc.

2011

Prohlašuji, že jsem disertační práci napsala samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Barbora Štindlová

V Praze dne 29. 4. 2011

Pro mou babičku Věru Erbanovou

PODĚKOVÁNÍ

Děkuji svému školiteli prof. Karlu Šebestovi za možnost podílet se na dobrodružství, jakým je budování žákovského korpusu češtiny nerodilých mluvčích, za podporu při hledání cesty ke zpracování tak obsáhlého tématu a za poskytování cenných rad a připomínek.
Neméně a s vděčností děkuji Alexandru Rosenovi a Jirkovi Hanovi za otevření dveří do nových lingvistických světů, za mnohahodinové konzultace a trpělivé zodpovídání mých otázek.
Svatavě Škodové děkuji za osobní i odbornou podporu, kterou mi po celou dobu poskytovala, a bez jejíhož přátelství by tato práce vznikla jen stěží.
Rovněž děkuji studentům Katedry českého jazyka a literatury FP TU v Liberci a Ústavu českého jazyka a teorie komunikace FF UK v Praze za jejich obětavou práci při testování anotace navržené pro žákovský korpus češtiny CzeSL.

Zvláštní poděkování za bezbřehou trpělivost, pochopení a trvalou podporu při psaní práce patří Ondřejovi, Ondráškovi, Tereze a samozřejmě mým rodičům.

Autorka

Název práce:	**Evaluace chybové anotace v žákovském korpusu češtiny**
Autor:	Barbora Štindlová
Ústav:	Ústav českého jazyka a teorie komunikace, Filozofická fakulta, Univerzita Karlova
Vedoucí disertační práce:	prof. PhDr. Karel Šebesta, CSc.

Abstrakt:

Předkládaná práce se obecně týká tématu češtiny jako cizího jazyka a částečně zasahuje do oblasti korpusové lingvistiky, neboť se věnuje problematice žákovských korpusů, především pak otázkám jejich chybového značkování a možnostem evaluace anotačních schémat. Žákovské korpusy se staly významným zdrojem pro poznání žákovského mezijazyka a významným stimulem pro různé oblasti studia a výuky cizího, resp. druhého jazyka. Jsou využívány zejména pro kontrastivní srovnávání jazyka rodilých a nerodilých mluvčích, resp. srovnávání žákovských mezijazyků a pro tzv. počítačem podporovanou chybovou analýzu žákovského jazyka. Pro tento typ analýzy má zcela zásadní důležitost tzv. chybové značkování. Chybové značkování je u každého korpusu, pokud jej používá, založeno na chybové typologii, jejíž vymezení je v mnoha teoretických aspektech problematické. Z toho důvodu je důležitým krokem při výstavbě žákovského korpusu zhodnocení spolehlivosti a validity navrženého anotačního schématu.

Disertační práce se zaměřuje především na technické aspekty a specifické problémy při elektronizaci rukopisů, na možnosti chybové anotace projevů nerodilých mluvčích a problematiku její evaluace. Zároveň však věnuje značný prostor i metodologii, koncepci a účelu budování žákovských korpusů, protože téma korpusu nerodilých mluvčích a jeho využití je v českém prostředí relativně nové a je vhodné jej podrobněji představit. V první části (A) jsou stručně shrnuty základní přístupy k otázkám nabývání cizího, resp. druhého jazyka a podrobněji představeny proměny teorie chyby v jazyce nerodilých mluvčích. V části (B) předkládám shrnutí aktuálního stavu problematiky a uvádím podrobný přehled existujících korpusů jazyka nerodilých mluvčích založený na dotazníkovém šetření a podrobné analýze dostupných žákovských korpusů. Třetí část práce (C) představuje budovaný žákovský korpus češtiny nerodilých mluvčích (CzeSL) a soustředí se především na problematiku přepisu dat. Čtvrtá část disertace (D) se zabývá evaluací konceptu chybové anotace navržené pro žákovský korpus CzeSL. Pro zhodnocení spolehlivosti anotačního schématu byl zvolen výpočet tzv. koeficientu mezianotátorské shody kappa. Výsledky měření mezianotátorské shody, analýza anotačních problémů, návrhy jejich řešení a zhodnocení anotačního schématu včetně chybové taxonomie jsou jedním z hlavních výsledků této práce.

Klíčová slova:	žákovský korpus, chybová anotace, elektronizace rukopisů, mezianotátorská shoda, mezijazyk

Title: **Evaluation of Error Mark-Up in a Learner Corpus of Czech**

Author: Barbora Štindlová

Department: Institute of Czech Language and Theory of Communication, Faculty of Arts, Charles University in Prague

Supervisor: prof. PhDr. Karel Šebesta, CSc.

Abstract:

The thesis deals with the topic of Czech as a second language, while introducing methods of corpus linguistics as applied to texts produced by language learners. The context is the process of building and exploiting a learner corpus, with a focus on its error mark-up and options for evaluating the annotation scheme.

Learner corpora have become a major resource for investigating a learner interlanguage and a significant incentive for many different types of research and teaching of second/foreign languages. They are used mainly for contrastive studies of native and non-native speakers, i.e. for contrastive interlanguage analysis, and for computer-aided error analysis of the learner language. This kind of analysis is crucially dependent on the type and quality of the error mark-up. In every error-annotated corpus the error annotation is based on an error typology, which is necessarily problematic from a number of theoretical aspects. Evaluation of the reliability and validity of the annotation scheme design is therefore an important step in the build-up of a learner corpus.

The thesis is concerned primarily with the technical aspects and specific issues involved in the digitization of hand-written texts, with options for the error annotation of non-native speakers' language, and with the issues of its evaluation. At the same time, a significant amount of space is devoted to the questions of methodology, architecture and purpose of the compilation of learner corpora, because the topic of a non-native speakers' corpus and its exploitation in the Czech environment is quite recent and thus a more detailed introduction is justified.

In the first part (A), several major approaches to the issues of foreign/second language acquisition are briefly summarized and the developments in the theory of error in non-native speakers' language are presented in more detail. In part B, a summary of the current state of the field is presented together with an overview of existing corpora of non-native speakers' language, the result of a questionaire-based research and a detailed analysis of available learner corpora. The third part (C) presents a learner corpus of non-native speakers' Czech (CzeSL), focusing on the issues of text transcription. In the fourth part (D), the error annotation scheme proposed for CzeSL is subjected to evaluation. To assess the reliability of the annotation scheme a measure of inter-annotator agreement – the coefficient kappa – is used. The measured results of the inter-annotator agreement, the analysis of the problematic points in the annotation scheme, and the evaluation of the scheme, including the error taxonomy, represent some of the main assets of the present thesis.

Keywords: learner corpus, error annotation, text transcription, inter-annotator agreement, interlanguage

OBSAH

SEZNAM TABULEK

SEZNAM GRAFŮ

ÚVOD

Předkládaná práce se obecně týká tématu češtiny jako cizího jazyka a částečně zasahuje do oblasti korpusové lingvistiky, neboť se věnuje problematice žákovských korpusů, především pak otázkám jejich chybového značkování a možnostem evaluace anotačních schémat.

Žákovské korpusy se staly významným zdrojem pro poznání žákovského mezijazyka a významným stimulem pro různé oblasti studia a výuky cizího, resp. druhého jazyka. Jsou využívány především ve dvou směrech. Za prvé jde o kontrastivní srovnávání jazyka rodilých a nerodilých mluvčích na pozadí korpusu národního, resp. srovnávání žákovských mezijazyků. Druhý přístup, tzv. počítačem podporovaná chybová analýza, vychází z metodologie původní chybové analýzy a zabývá se studiem žákovských chyb. Pro kontrastivní zkoumání žákovského jazyka není bezprostředně nutná chybová či jiná anotace a lze pracovat s holým (tj. neanotovaným) žákovským korpusem. Pro počítačem podporovanou chybovou analýzu má však zcela zásadní důležitost tzv. chybové značkování. Tato chybová anotace znamená přiřazení značek se striktně definovaným významem jednotlivým chybným výrazům, umožňuje následné vyhledávání v korpusu (např. při analýze nadužívání či podužívání konkrétních výrazových prostředků) a je základem pro další výzkumy v oblasti cizojazyčného učení a vyučování. Chybové značkování je u každého korpusu, pokud jej používá, založeno na chybové typologii, jejíž vymezení je v mnoha teoretických aspektech problematické. Z toho důvodu je důležitým krokem při výstavbě žákovského korpusu zhodnocení spolehlivosti a validity navrženého anotačního schématu.

Pro budování a zpracování žákovských korpusů jsou zásadní následující klíčové oblasti:

A. metodologie, koncepce a účel korpusu;
B. sběr dat a jejich povaha;
C. převod dat do elektronizované podoby, včetně technických aspektů;
D. anotace správní, lingvistická a chybová (včetně zhodnocení její spolehlivosti);
E. softwarové vybavení a aplikace užívané pro získávání dat z korpusu.

Ve své disertační práci se zaměřuji především na body C a D, konkrétně na technické aspekty a specifické problémy při elektronizaci rukopisů, na možnosti chybové anotace projevů nerodilých mluvčích a problematiku její evaluace. Zároveň však věnuji značný prostor i bodu A, nejen z toho důvodu, že s následujícími body úzce souvisí, ale zejména proto, že téma žákovského korpusu a jeho využití je v českém prostředí relativně nové a je vhodné jej podrobněji představit.

STRUKTURA PRÁCE

V první části (A) jsou stručně shrnuty základní přístupy k otázkám nabývání cizího, resp. druhého jazyka a podrobněji představeny proměny teorie chyby v jazyce nerodilých mluvčích. Tato část slouží jako metodologický základ pro principy budování chybové taxonomie a chybové anotace v žákovských korpusech obecně, a v korpusu češtiny nerodilých mluvčích zvlášť. Zároveň je také bází pro koncept ověřování validity této chybové anotace.

V části (B) předkládám shrnutí aktuálního stavu problematiky a uvádím podrobný přehled existujících korpusů jazyka nerodilých mluvčích založený na dotazníkovém šetření a podrobné analýze dostupných žákovských korpusů. Korpusy jsou srovnávány na základě parametrů výstavby, tj. sběru materiálu, metadat, anotace apod. Dále v této části zevrubně analyzuji přístupy k chybové anotaci, výhody a nevýhody jednotlivých anotačních typů a možnosti jejich aplikací.

Třetí část práce (C) představuje budovaný žákovský korpus češtiny nerodilých mluvčích (CzeSL) a soustředí se především na problematiku přepisu dat. Jsou zde zmíněny principy přepisu rukopisů podle zásad Text Encoding Initiative (TEI), a zároveň také pravidla pro přepis uplatňovaná v Korpusu soukromé korespondence (KSK). Dále v této části předkládám původní návrh pravidel pro přepis (rukou psaných) textů nerodilých mluvčích, který byl vybudován pro specifické potřeby žákovského korpusu CzeSL, resp. pro potřeby akvizičních korpusů skupiny AKCES.

Čtvrtá část disertace (D) se zabývá evaluací konceptu chybové anotace navržené pro žákovský korpus CzeSL. Pro zhodnocení spolehlivosti anotačního schématu byl zvolen výpočet tzv. koeficientu mezianotátorské shody kappa. Evaluace byla aplikována na vzorek textů (9848 slov),

který byl anotován proškolenými anotátory – studenty Ústavu českého jazyka a teorie komunikace FF UK v Praze a Katedry českého jazyka a literatury TU v Liberci. Výsledky měření mezianotátorské shody, analýza anotačních problémů, návrhy jejich řešení a zhodnocení anotačního schématu včetně chybové taxonomie jsou jedním z hlavních výsledků této práce.

Cíle předkládané disertační práce jsou:

1) zmapování hlavních teoretických přístupů k problematice nabývání cizího, resp. druhého jazyka, které bude následně využito například pro potřeby výuky v rámci modulů prohloubené specializace pro studijní programy Specializace v pedagogice a Filologie na FP TU v Liberci;[1]
2) aktuální podrobný přehled světových žákovských korpusů a analýza jejich anotačních formátů, včetně návrhů chybových taxonomií;
3) představení způsobu transkripce dat navržené pro žákovský korpus CzeSL;
4) návrh postupu pro zhodnocení spolehlivosti anotace projevů nerodilých mluvčích češtiny a anotačního schématu určeného pro žákovský korpus CzeSL, analýza problémů a příklady řešení;
5) překladový terminologický slovník jako základ budoucího výkladového slovníku pro potřeby oboru čeština jako cizí jazyk.

TERMINOLOGICKÁ POZNÁMKA

V práci používám množství termínů, které jsou v běžné v zahraniční literatuře věnující se tématu nabývání druhého, resp. cizího jazyka a cizojazyčného vyučování, a v českém prostředí jsou užívány především v didaktice cizích jazyků. Velká část těchto termínů nemá standardizovaný český ekvivalent a řada z nich dokonce není jednoznačně definována.

Abychom předešli nedorozumění, považuji za nutné vymezit zde základní termíny, se kterými v disertační práci operuji.

Podrobný seznam pojmů, které jsou pro tuto práci zásadní, je uveden v příloze v podobě překladového terminologického slovníku. Pro snazší orientaci zároveň v textu vždy uvádím i anglický ekvivalent prezentovaného termínu.

[1] Viz http://www.c2j.cz/attachments/078_Koncepce_inovace.pdf.

Základní termíny

první jazyk, mateřský jazyk

Termínem *první jazyk (first language, L1*) obvykle označujeme jazyk, jejž si jedinec osvojí jako první v pořadí. V tomto smyslu se jedná o *jazyk mateřský (mother tongue).* Zároveň se však tento termín může vztahovat na jazyk dominantní, tj. ten, kterým jedinec mluví nejlépe, nejčastěji, příp. nejraději. Většinou je rozlišování termínu *první* a *mateřský jazyk* irelevantní, u multilingválních mluvčích však může v průběhu života docházet k proměnám v chápání toho, co je jejich *první jazyk.*

cizí jazyk, druhý jazyk

Terminologicky se obvykle rozlišuje pojem *cizí jazyk (foreign language, FL)* jako jazyk nabývaný v prostředí, kde se tímto jazykem nemluví (např. studium angličtiny v neanglicky mluvících zemích) a *druhý jazyk (second language, L2)* jako jazyk osvojovaný v přirozeném prostředí, tj. kde je tento jazyk oficiálním komunikačním prostředkem (např. tzv. „imigrantská angličtina"). Někteří autoři (např. Ellis, 1994) však chápou termín *druhý jazyk* jako nadřazený a označují jím jakýkoli nemateřský jazyk, který se jedinec učí poté, co si osvojil jazyk mateřský. Pro potřeby této práce mezi oběma termíny nerozlišuji a užívám je libovolně.

cílový jazyk

Termínem *cílový jazyk (target language, TL)* se označuje jazyk, kterého jedinec nabývá, jehož osvojení je cílem učebních aktivit, v němž chce být mluvčí schopen komunikovat.

mezijazyk, interlanguage

Termínem *mezijazyk,* příp. také *interlanguage,* se označuje tzv. žákovský jazyk, tj. jazyk nerodilých mluvčích. Selinker (1972), který termín poprvé použil, jím charakterizoval obecný rys žákovského jazyka být samostatným systémem, který má status přechodného systému mezi jazykem prvním a jazykem cílovým.

žák, žákovský jazyk, žákovský korpus

V kontextu korpusů shromažďujících jazyk nerodilých mluvčích pomíjíme tradiční konotace, které výraz *žák (learner)* v češtině má, a chápeme jej ve shodě se zahraniční terminologií jako osobu učící se cizí jazyk, a to bez ohledu na věk, příp. jiné sociologické proměnné.
Žákovským jazykem (learner language) pak rozumíme obecně jazyk nerodilého mluvčího.
Termín *žákovský korpus*[2], tj. korpus jazyka nerodilých mluvčí, byl zvolen v souladu s tradičně užívanými světovými termíny *learner corpus,* resp. *Lernerkorpus.* V českém prostředí se lze setkat i s variantním názvem *studijní korpus,* viz Čermák a Schmiedtová (2004).

jazykový vstup, jazykový výstup

Termín *jazykový vstup,* příp. *vklad (input, exposure)* charakterizuje souhrnně způsob působení jazyka na žáka. V kontextu jazykového vyučování jsou za tzv. *jazykový vstup* považovány učební činnosti a materiály. Hypotéza srozumitelného jazykového vstupu *(comprehesible input hypothesis)* byla poprvé formulována S. Krashenem (1981).
Jazykový výstup (output) je chápán souhrnně jako žákova produkce. Hypotézu jazykového výstupu *(output hypothesis)* formulovala poprvé M. Swainová (srov. 1995).

osvojování jazyka

Osvojování jazyka (acquisition) je chápáno jako obdoba učení dítěte mateřskému jazyku a jeho charakteristikou je, že není uvědomováno.

učení se jazyku

Termín *učení se jazyku (learning)* reflektuje vědomý proces směřující k jazykovým pravidlům, obvykle má institucionalizovanou podobu.

[2] Termín byl poprvé použit v návrhu projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk* v roce 2008.

nabývání jazyka, jazyková akvizice

Oba termíny, tj. *nabývání* i *akvizice,* užíváme jako zastřešující pro pojmy předchozí (tj. osvojování a učení).

chybová anotace, značkování (tagování), emendace

Anotací se obecně rozumí proces, ale i výsledek přiřazení příslušené sekundární informace (metainformace) k originálnímu textu. Vhodnou a spolehlivou anotací (externí, lingvistickou, lemmatizací apod.) se korpus pro uživatele mnohonásobně zhodnocuje. O tzv. *chybové anotaci* (neboli *chybovém značkování,* příp. *tagování*) se hovoří při označování chyb v textech nerodilých mluvčích podle zvolené chybové taxonomie. Pro potřeby této práce užívám pojem *chybová anotace* jako souhrnný, zahrnující obě anotační aktivity, tj. *značkování* (anotaci v užším slova smyslu) a přímou opravu daného chybného výrazu, tj. *emendaci.* Zároveň však také na některých místech pracuji s pojmem *anotace* v užším slova smyslu, tj. označujícím pouze přiřazení příslušené chybové značky (tagu).

Základní zkratky[3]

CA	*contrastive analysis*	kontrastivní analýza
CEA	*computer aided error analysis*	počítačem podporovaná chybová analýza
CIA	*contrastive interlanguage analysis*	kontrastivní analýza mezijazyka
EA	*error analysis*	chybová analýza
FEAT		anotační program vyvinutý pro chybové značkování češtiny nerodilých mluvčích
FLA	*foreign language acqusition*	nabývání, příp. osvojování cizího jazyka
FLL	*foreign language learning*	učení se cizímu jazyku
FLT	*foreign language teaching*	výuka cizího jazyka
IAA	*inter-annotator agreement*	mezianotátorská shoda
SERR	*Common European Framework of Reference for Languages*	Společný evropský referenční rámec pro jazyky
SLA	*second language acquisition*	nabývání, příp. osvojování druhého jazyka
TEI	*Text Encoding Initiative*	Iniciativa pro kódování textu
UG	*universal grammar*	univerzální gramatika
XML	*Extensible Markup Language*	rozšiřitelný značkovací jazyk
κ	*kappa coeficient*	koeficient kappa

[3] V případě, že v českém odborném diskurzu chybí vhodná zkratka, používám v disertační práci odpovídající mezinárodně standardizovanou formu. Domnívám se, že zavádět nové české překladové zkratky není pro potřeby této práce vhodné.

1 TEORIE CHYBY V JAZYCE NERODILÝCH MLUVČÍCH

Hlavním tématem naší práce je ověřit funkčnost navržené chybové anotace pro žákovský korpus češtiny.[4] Ústředním pojmem pro problematiku značkování žákovského jazyka a pro validaci tohoto značkování je pojem jazykové chyby v projevu nerodilých mluvčích. V následující kapitole se soustředíme na proměny ve vnímání chyby v jazyce nerodilých mluvčích v souvislosti s proměnami teoretických přístupů k analýzám nabývání cizího jazyka. Stručně shrneme vývoj této disciplíny od poloviny dvacátého století, a pak se zaměříme na vymezení pojmu 'chyba' v nemateřském jazyce a na vývoj konceptu chyby v návaznosti na modelování teorie nabývání druhého, resp. cizího jazyka. Podrobněji se budeme zabývat chybovou analýzou, resp. definicí, klasifikací a evaluací chyb v projevech nerodilých mluvčích, protože jádro naší práce se věnuje problematice žákovského korpusu a jeho chybové anotaci, resp. chybové taxonomii. Chceme tedy vytvořit bázi, na jejímž základě budeme moci jednotlivé typy taxonomií zhodnotit a vzájemně porovnat.

1.1 Teorie nabývání druhého, resp. cizího jazyka (SLA)

Typickým rysem v oblasti nabývání druhého, resp. cizího jazyka v posledních padesáti letech je velký rozptyl teoretických hledisek. Od poloviny minulého století bylo předloženo mnoho různorodých teorií[5], epistemologicky ukotvených především v realismu a empirismu (méně již v relativismu). Tato skutečnost vychází z faktu, že nabývání cizího jazyka je komplexní proces a pro porozumění tomuto procesu je nutná spolupráce odborníků z mnoha vědních oblastí. Není úkolem této práce podrobně rozebrat dílčí teoretické přístupy k SLA. Protože však v následujících oddílech hodlám vymezit pojem chyba v kontextu nabývání cizího jazyka a cizojazyčného vyučování a zmapovat chápání chyby ve vztahu k hlavním teoretickým konceptům SLA (a následně i konceptům cizojazyčného vyučování, dále jako FLT), zmíním zde stručně několik teorií, které měly, resp. mají podle mého názoru zásadní vliv na podobu zkoumání nabývání druhého jazyka. V tomto přehledu vycházím především ze studií Larsen-Freeman*ové* (1997), Doughty*ové* a Longa (2004), Jordana (2004), Van Pattena a Williams*ové* (2007), Lightbown*ové* a Spada*ové* (2007).

Z historického hlediska můžeme vymezit tři hlavní vědecká paradigmata, která ovlivnila výzkumy SLA, resp. FLL a FLT. Jedná se o behaviorální, kognitivně-komputační a tzv.

[4] K tomu viz dále v této práci, kap. 8.
[5] Larsen-Freeman*ová* a Long (1991: 227) uvádějí, že jich bylo navrženo nejméně čtyřicet.

dialogickou perspektivu (prezentovanou také jako hermeneutickou, sociálněkognitivní či kulturní) v analýzách nabývání cizího jazyka. V zásadě lze říci, že většina bádání v dané oblasti aplikuje kognitivně-komputační hledisko, naopak tzv. dialogický přístup je často chápán jako „nevědecký". Srov. Johnson*ová* (2004: 9).

Behavioristicky orientovaná teorie nabývání druhého jazyka chápe jazyk jako strukturu a akvizici jako formování jazykového chování *(habit formation).*[6] Akcentuje roli (jazykového) prostředí a vstupu *(inputu),* resp. jazykové instrukce, podceňuje naopak interní mechanismy, tj. mentální procesy budování jazykové znalosti. Takové nahlížení na jazykové učení podnítilo vznik kontrastivní analýzy zaměřující se primárně na interferenční vliv prvního jazyka na jazyk cílový, která měla velký vliv nejen na teorii SLA, ale také na podobu FLT. Ke kontrastivní analýze viz dále v této práci, oddíl 1.5. V rámci tohoto přístupu je akcentována především orální produkce. Experimentální metody a konstrukty jako podmiňování, posilování a potrestání (v případě jazykového učení oprava), které behavioristický přístup přinesl, jsou pro některé analýzy SLA využitelné i v současnosti.

Nedostatky v behavioristickém chápání jazykové akvizice vedly k hledání alternativního teoretického rámce. Nové paradigma bylo orientováno mentalisticky a dalo základ formování teorie mezijazyka. Lze vymezit dva subsměry, resp. dvě verze tohoto přístupu: starší hypoteticko-deduktivní přístup a novější informačně procesuální, komputační. Na přelomu šedesátých a sedmdesátých let minulého století poukazovali někteří badatelé na skutečnost, že žákovský jazyk je systematický a chyby v něm jsou dokladem pravidly řízeného chování[7] (Corder, 1967; Nemser, 1971; Selinker, 1972). Z tohoto náhledu se vyvinula koncepce mezijazyka, resp. předpokladu, že studenti cizího jazyka internalizují mentální gramatiku, přirozený jazykový systém, který se dá popsat jazykovými pravidly a principy.[8] Zásadní vliv na změnu paradigmatu v SLA měla Chomského (1959) kritická recenze Skinnerova textu „Verbální chování", která fakticky byla obecnou kritikou behaviorismu (resp. empirismu) a ve které Chomsky poprvé představil ideu biologické predispozice k jazyku[9]. Ačkoli se Chomsky nikdy

[6] Vedle behaviorismu byla hybnou silou v tomto přístupu k jazykovému učení strukturní lingvistika, resp. deskriptivismus.

[7] Viz dále zde, oddíl 1.4.

[8] Odmítnutí behaviorismu v kontextu SLA souvisí s tzv. kognitivním obratem v psychologii, resp. lingvistice, a s nástupem nativismu, resp. mentalismu. Viz dále např. White*ová,* 2004: 19.

[9] Později definovanou jako tzv. LAD (*Language Acquisition Device),* modul, který využívá jazykový materiál z okolí a buduje na základě vrozených mechanismů jazykovou kompetenci mluvčího.
Chomsky (1959: odd. V): „These abilities indicate that there must be fundamental processes at work quite independently of "feedback" from the environment. … are important factors, as is the remarkable capacity of the child to generalize, hypothesize, and "process information" in a variety of very special and apparently highly complex ways which we cannot yet describe or begin to understand, and which may be largely innate, or may

nezaměřoval na problematiku nabývání cizího jazyka, jeho revoluční názory a návrh nových metod zkoumání jazyka, které významně proměnily podobu moderní lingvistiky, značně ovlivnily i teorii SLA. Podle Chomského nedokázala behavioristická teorie zodpovědět tzv. logický problém jazykové akvizice, tj. fakt, že dítě nabývající mateřský jazyk (příp. cizinec nabývající druhý jazyk) má více informací o struktuře osvojovaného jazyka, než k jakým se na daném stupni vývoje mohlo dostat učením v souvislosti s jazykovým vkladem. Zároveň Chomsky (1972) kritizuje koncept kontrastivní analýzy, resp. představu totálního explicitního popisu jazykové struktury pro účely komparace a explanace.

Teorie univerzální gramatiky (UG), kterou Chomsky postuloval, vychází z pozorování, že jazykový vliv okolí není dostatečným vysvětlením jazykové akvizice. První období výzkumů role UG v SLA se zaměřovalo především na zjišťování, zda a do jaké míry je UG k dispozici studentovi cizího jazyka, resp. zda je jeho mezijazyk vázán principy univerzální gramatiky. Jinými slovy, zda je gramatika mezijazyka řízena obdobným způsobem jako gramatika rodilého mluvčího. Výsledky provedených experimentů jsou nejednoznačné. Zatímco např. White*ová* (2004) vidí UG jako nejvhodnější teoretický rámec pro popis SLA, protože studenti cizího jazyka mají k UG bezprostřední přístup, jiní autoři (Bley-Vroman, 1983; Schachter*ová*, Jackendoff)[10] akceptují roli UG v akvizici prvního jazyka, ale poukazují na to, že koncept UG je pevně spjat s hypotézou kritického období, a proto není jako teoretický základ výzkumů SLA přijatelný. Bley-Vroman (1990), Schwartz*ová* a Sprouse (2000) aj. převádějí na rovinu učení druhému jazyku tzv. logický problém jazykové akvizice a uvažují o tom, že důvodem disproporce mezi jazykovým vstupem (tj. inputem) a výstupem (tj. outputem) je vliv prvního jazyka, tj. jistým způsobem zprostředkované působení UG. Od devadesátých let minulého století se debata o roli UG v nabývání cizího jazyka přesunula k výzkumům povahy mezijazyka, resp. k otázkám odlišnosti gramatiky žákovského jazyka a gramatiky rodilých uživatelů cílového jazyka. Základním tématem je otázka počátečního stavu *(initial state)* interlanguage, tj. zda a příp. jak je počáteční mezijazyk utvářen pod vlivem gramatiky mateřského jazyka. Srov. např. White*ová* (2004: 30–33). K problematice mezijazyka viz dále v této práci, oddíl 1.4.

Náhled na jazyk jako vrozenou schopnost, chápání jazykového učení jako kreativního procesu a postupné zpracovávání jazykového vstupu pomocí vrozených mechanismů je základem teorie SLA, kterou představil na přelomu sedmdesátých a osmdesátých let S. Krashen (1981). Jednalo

---

develop through some sort of learning or through maturation of the nervous system. The manner in which such factors operate and interact in language acquisition is completely unknown.“

[10] SCHACHTER, J. On the issue of completeness in second language acquisition. *Second Language Research,* 1990, vol. 6, no. 1, s. 93-124.
JACKENDOFF , R. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press, 2002.

se o první teorii, která byla zpracována cíleně pro problematiku SLA a je dodnes jednou z nejvlivnějších teorií v tomto oboru, ačkoli její závěry si vysloužily velkou kritiku (srov. např. White*ová,* 1987; McLaughlin, 1978; Cook, 1991 aj.). Krashen vymezil svůj model jazykové akvizice pěti hypotézami: 'hypotéza jazykového osvojování – učení' rozlišuje mezi vědomým jazykovým učením, zaměřeným na akvizici forem a pravidel, a podvědomým osvojováním (cizího) jazyka, tj. procesem, který se uplatňuje při nabývání mateřštiny; 'hypotéza přirozeného pořádku' tvrdí, že gramatické struktury jsou osvojovány v předvídatelném sledu a 'hypotéza srozumitelného jazykového vstupu' *(comprehensible input)* řeší vztah mezi vstupem a akvizicí jazyka; 'hypotéza monitoru' (tzv. *Monitor Model)* chápe funkci vědomého učení jako monitorujícího, resp. (sebe)korekčního prostředku; 'hypotéza afektivního filtru' definuje metaforickou bariéru, která brání v nabývání jazyka, i když je k dispozici adekvátní jazykový input.

Vedle rozvíjení nativistických konceptů v závislosti na revidovaných chomskyánských teoriích (řízení a vázání, minimalismus) odrážejí ostatní současné teorie nabývání cizího jazyka kognitivní, resp. vývojovou perspektivu v lingvistice. Mohli bychom říci, že v nynějších výzkumech SLA se projevuje jisté napětí mezi dvěma vůdčími přístupy, z nichž jeden můžeme klasifikovat jako psychologický (resp. psycholingvistický) a druhý jako sociokognitivní (resp. sociolingvistický). Jak uvádějí Taron*eová* a Swain*ová* (1995: 167), tam, „kde psycholingvisté mají tendenci zaměřovat se při výzkumu na izolovaná individua, sociolingvisté preferují chápat individua jako reprezentanty jazykové komunity, ve které se vyskytují."

Psychologizující přístupy k SLA studují mentální struktury a procesy, které se vztahují k nabývání cizího jazyka. Jak jsme již zmínili výše, je studium psychologických aspektů jedním z hlavních přístupů k současnému uvažování o SLA a vedlo ke vzniku mnoha akvizičních modelů, které se zabývají především otázkami pozornosti, paměti, plynulosti, resp. mentálních reprezentací a způsobu zpracovávání informací (kognitivně orientovaný přístup k uvažování o SLA viz např. Skehan, Gass*ová* aj.).[11] Tyto modely odmítají ideu vrozené jazykové predispozice, resp. myšlenku vrozeného neurologického modulu pro jazykovou akvizici. Vlivným směrem bádání v této sféře je konekcionismus, který vychází z předpokladu, že informace lze členit na elementy, mezi nimiž existují spoje, které mají různou závažnost, resp. sílu. Tyto elementy (uzly) vytvářejí konekcionistické, neuronové sítě umožňující paralelní zpracovávání údajů. V kontextu SLA to znamená, že žáci vystavovaní dostatečně frekventovanému jazykovému vstupu postupně budují znalost jazyka vytvářením pevné sítě mezi jednotlivými elementy, resp.

[11] SKEHAN, P. *A cognitive approach to language learning.* Oxford: OUP, 1998.
GASS, S. *Input, interaction and the second language learner.* Mahwah, NJ: L. Erlbaum, 1997.

jazykovými rysy. Spoje mezi jednotlivými uzly jsou v závislosti na jazykovém vstupu posilovány nebo zeslabovány. Tj. jazykové učení je nahlíženo jako simultánní zpracovávání zkušenosti, jejíž opakování tyto spoje upevňuje. Toto pojetí je podpořeno předpokladem, že jazyk je osvojován alespoň částečně po blocích větších, než je samostatné, izolované slovo, a že věty se při produkci nutně netvoří po slovních jednotkách (srov. Ellis, 2003, Lightbown*ová* a Spada*ová,* 2006: 41). Studie vycházející z tohoto přístupu se však úzce zaměřují na nabývání znalosti lexika a gramatických morfémů. Modelem nabývání cizího jazyka úzce navazujícím na konekcionismus je tzv. kompetiční model (*competition model,* viz Bates*ová* a McWhinney, 1989), který se vymezuje vůči principům UG a akvizici cizího jazyka chápe jako proces učení založený na univerzáliích kognitivního procesu. Bere v úvahu nejen jazykovou formu, ale také význam a užití a jako základní principy teorie definuje vkladem řízené učení (*input-driven learning)* založené na síle tzv. podnětů (*cues*), kterými jsou v jazyce signalizovány určité funkce. Síla těchto podnětů nemusí být v prvním a druhém jazyce stejná (např. slovosled, který signalizuje vztahy participantů apod.).

Diskuze o dalším směřování analýz v SLA se rozvinula v návaznosti na text Firtha a Wagnera (1997: 285)[12], kteří odmítli příliš „individualistický a mechanický" psycholingvistický přístup, jež pohlíží na žáka jako inherentně defektního mluvčího, a navrhli „rekonceptualizaci SLA jako sociokognitivního a kontextuálně orientovaného procesu". Řada psycholingvisticky zaměřených badatelů jejich návrhy následně podrobila ostré kritice (např. Long, Gass*ová* aj.). Sociokognitivní teorie navazují na výzkumy ze sedmdesátých let, které se objevily jako alternativa behavioristického přístupu k SLA i jako odpověď na etablování teorie UG. Jedná se např. o Schumannův (1978) *akulturační model* a jeho propracovanější variantu v podobě *nativizačního modelu* R. Andersena (1985), který chápe SLA jako proces adaptace na novou kulturu, determinuje stupeň sociální a psychologické vzdálenosti mezi kulturou žákovského a cílového jazyka a zkoumá proces pidžinizace při nabývání druhého jazyka, který podle něj nastává v případě, když žák při akulturaci selže. Dalším přístupem je teorie přizpůsobení *(accomodation theory)* Howarda Gilese akcentující roli motivace; diskurzní teorie *(discourse analysis)* Hatch*ové* odrážející se v pozdějším pojmu sociálního myšlení (viz např. Atkinson); a *model proměnlivé kompetence* Taron*eové,*[13] pokoušející se osvětlit variabilitu mezijazyka a

[12] FIRTH, A., WAGNER, J. On discourse, communication, and (some) fundamental conceptsin SLA. *Modern Language Journal,* 1997, vol. 81, s. 285-300.

[13] GILES, H., TAYLOR, D. M., BOURHIS, R. Towards a theory of interpersonal accommodation through language: some Canadian data. *Language in Society,* 1973, vol. 2, s. 177-192.
HATCH, E. *Discourse and Language Education.* Cambridge: Cambridge University Press, 1992.
TARONE, E. On the variability of interlanguage systems, *Applied Linguistics,* 1983, vol. 4, s. 143-163.

proces jazykového učení, který následně domýšlejí Ellis, Widowson, Bialystok*ová.* Badatelé aplikující sociálně orientovaný přístup k analýzám SLA jsou v současnosti často kritizováni za přílišný teoretický pluralismus a relativismus. Srov. Matsuoka a Evans (2004).

Současné chápání jazykového učení jako komplexního, nelineárního, dynamického a proměnlivého procesu (srov. Larsen-Freeman*ová,* 2002: 35) vede ke vzniku nových teorií nabývání cizího jazyka. Jednou z aktuálních a vědeckou veřejností diskutovaných koncepcí je teorie komplexnosti ve vývoji jazykových znalostí[14] (také nazývaná teorie chaosu, *Chaos/complexity theory – C/CT*), kterou postulovala Larsen-Freeman*ová*[15] jako adaptaci původně fyzikálního přístupu a kterou v modifikované formě prezentovali i Verspoor*ová,* Lowie a De Bot (2007) jako tzv. teorii dynamických systémů *(Dynamic Systems Theory - DST).* Tento přístup chápe vývoj jazykové znalosti jako nelineární, adaptativní, interaktivní a nevratný (v tom smyslu, že jednoduchá opětovná aplikace nějaké procedury nemusí mít vždy stejný efekt), vyplývající z vnější interakce i vnitřní reorganizace zároveň (Verspoor*ová,* Lowie a De Bot, 2007: 2). Praktickou aplikací C/CT, resp. DST je tzv. komplexní adaptivní systém. Částečně odlišným směřováním v souvislosti s úvahami o jazykovém učení jako komplexním, organickém procesu je teorie fraktálového modelu SLD, která chápe dosavadní přístupy k SLA jako nedostatečně fragmentární a snaží se postihnout všechny proměnné procesu jazykového učení a jejich vzájemné vztahy (Menezes*ová*, 2004).

Nový teoretický rámec odmítající „kognitivní jednostrannost" v přístupu k nabývání druhého jazyka, zaměřenou zejména na výzkum vztahu mentálních procesů a jazykové kompetence, nabízí také Johnson*ová* (2004), která přistupuje k problematice SLA z hlediska performance, obhajuje tzv. dialogický přístup[16] a úzce navazuje na Vygotského sociokulturní teorii a Bachtinův koncept dialogičnosti významu a jeho pojetí heteroglosie.

ATKINSON, D. Toward a sociocognitive approach to second language acquisition. *Modern Language Journal*, 2002, vol. 86, s. 525-545.

[14] Nabývání jazyka se chápe ve vývojové perspektivě, proto zastánci tohoto přístupu dávají přednost užívání termínu 'vývoj znalosti cizího jazyka' *(second language development – SLD)* před termínem 'akvizice cizího jazyka' (SLA), protože tento termín lépe demonstruje množství dynamicky propojených procesů, včetně akvizice i ztráty. Srov. Verspoor*ová,* Lowie a De Bot (2007: 27).

[15] Stať Larsen-Freeman*ové* (2002) je některými odborníky označována jako počátek nové vlny v analýzách SLA, resp. jako počátek nástupu nového paradigmatu. Srov. např. Menezes*ová* (2004) aj.

[16] Tzv. dialogický přístup k SLA upřednostňuje kvalitativní výzkum před kvantitativním, zaměřuje se spíše na individuální specifičnost při nabývání cizího jazyka než na výzkum vedoucí k všeobecně platné generalizaci zjištění. Zdůrazňuje důležitost sociálních, kulturních, politických, historických a institucionálních kontextů.

1.2 Definice chyby

Chyby jsou nedílnou součástí procesu nabývání druhého, resp. cizího jazyka a stejně tak i cizojazyčného vyučování a fenomén chyby je dlouhodobě jedním z ústředních zájmů badatelů z obou oblastí. Vymezení chyby v produkci nerodilých mluvčích[17] však není snadné, protože chápání chyby v kontextu jazykové akvizice a jazykového vyučování je značně subjektivní (srov. např. Shaughnessy*ová,* 1977; Lennon, 1991; James, 1998). Jednoznačné definování chyby je problematické, protože podle řady dílčích výzkumů (James, 1977; Hughes a Lascaratou*ová*, 1982; Davies*ová,* 1983; McCretton*ová* a Rider, 1993) rodilí mluvčí jsou při identifikaci a evaluaci chyb v projevech cizinců nejednotní a v názorech na to, co je chyba a jaká je míra její závažnosti, se projevuje výrazná rozrůzněnost.[18] V analýzách SLA a FLT je chyba standardně definována jako odchylka od normy cílového jazyka. V souvislosti s tímto obecným vymezením se však nabízí otázka, jaká varieta cílového jazyka by měla sloužit jako norma pro klasifikaci chyb v jazyce nerodilých mluvčích. Spisovná podoba psaného jazyka je nevhodným modelem pro popis žákovy orální produkce. Z obdobného důvodu však není přijatelné definovat jako normu cílového jazyka jeho standardní podobu mluvenou. Definice chyby v současných výzkumech nabývání cizího jazyka tedy odkazují většinou na produkci nerodilého mluvčího, která se odchyluje od „správné" verze cílového jazyka, jejíž normou je nejednoznačně vymezená tzv. „norma (dospělého) rodilého mluvčího", k pojmu viz např. Faerch a Casper*ová* (1987, s. 125). Druhou komplikací při definování pojmu chyba v jazyce nerodilého mluvčího je již zmíněný fakt, že převážná většina vysvětlení vychází ze zaměření na chyby v produkci. Chyby v recepci jsou při vymezování výše uvedeného pojmu opomíjeny, protože „je velmi obtížné přiřadit příčinu receptivního selhání neadekvátní znalosti konkrétního syntaktického rysu vyskytujícího se v nepochopené výpovědi", Corder (1974: 125). Srov. i Maicusi et al. (1999–2000).

[17] V práci se zaměřujeme pouze na problematiku chyby v jazyce nerodilých mluvčích a záměrně upouštíme od srovnávání s chybami vyskytujícími se v prvním jazyce. Přesto jsme si vědomi vlivu, který má na kategorizaci jazykových defektů v procesu SLA chápání chyb jako „přechodných forem" ve vývoji osvojování jazyka u dětí nabývajících první jazyk a chyb jako pouhých „přeřeknutí"v jazyce dospělých rodilých mluvčích. Viz i Ellis (1994).

[18] Srov. např. studii Hughese a Lascaratou*ové* (1982), kteří zkoumali shodu při identifikaci chyb a při hodnocení jejich závažnosti. Hodnotiteli v této sondě byli rodilí a nerodilí učitelé cílového jazyka a zároveň také rodilí mluvčí – neučitelé. V návaznosti na obdobný výzkum C. Jamese (1977) Hughes a Lascaratou*ová* potvrdili, že rodilí mluvčí jsou při identifikaci chyb i při posuzování míry jejich závažnosti benevolentnější než nerodilí učitelé daného jazyka. Zároveň se lišila i hodnocení rodilých mluvčích. Zajímavým zjištěním bylo, že „správné" věty, které autoři excerpovali z *Oxford Advanced Learner's Dictionary of Current English* a vložili je mezi chybné příklady produkce cizojazyčných studentů, označilo značné množství hodnotitelů jako chybné (např. *Neither of us feels quite happy* označili jako chybnou konstrukci 3 nerodilí a 3 rodilí učitelé angličtiny a 5 rodilých ne-učitelů z celkového počtu 30 respondentů.)

Pro ilustraci zmíněného problému, tj. obtížné vymezitelnosti pojmu chyba v jazyce nerodilých mluvčích, uvádím několik rozdílných definicí žákovské chyby. Někteří autoři prezentují ostře formulovaná vymezení žákovské chyby, která však v žádném případě nejsou jednoznačná. Např. Chun et al. (1982: 538) definuje chybu v projevu nerodilého mluvčího jako „použití jazykových jednotek takovým způsobem, který podle pokročilých uživatelů *(fluent users)* indikuje chybu nebo nedostatečné učení". Liski a Puntanen (1983:227) postulují, že *„[ch]*yba se objeví vždy, když mluvčí selže při aplikaci modelů a variant promluv, které používají vzdělaní lidé v současných anglicky mluvících zemích." Jiní badatelé dávají přednost mírnějším definicím, např. Corder (1973: 26) používá „označení chyba pro popis takového znaku žákovského projevu, který se *[v širokém slova smyslu – BŠ]* liší od projevu rodilého mluvčího." Dulay*ová* et al. (1982: 130) vnímají chybu jako „defektní stránku žákovské řeči", která „se odchyluje od nějakého vybraného standardu jazykové performance dospělých." Lennon (1991) vymezuje chybu jako „lingvistické formy nebo kombinace forem, které by ve stejném kontextu a za obdobných podmínek produkce rodilý 'protějšek' *[uvozovky BŠ]* s velkou pravděpodobností nepoužil." Užším způsobem vymezuje chybu Brown (2000: 222), když uvádí, že chyba indikuje „zřetelnou deviaci od gramatiky dospělých rodilých mluvčích." Poněkud odlišnou definici chyby přináší Piepho (1973: 20): „Chyba není odchylkou od jazykové normy, nýbrž zkušenost, že se nějaký komunikační záměr nedá prosadit zvolenými prostředky." James (1998: 1) navrhuje stručnou, ale výstižnou definici chyby jako „nezdařeného kusu jazyka" *(unsuccessful bit of language).*

V této práci nechci prezentovat žádnou novou, „správnou" definici žákovské chyby. Pro své potřeby se přikláním k výše uvedenému vymezení Dulay*ové* et al. (ibid.). V následujících oddílech podrobněji pojednáme problematiku defektnosti projevů nerodilých mluvčích a přiblížíme proměňující se chápání chyby v jednotlivých modelech SLA. Zaměříme se přitom především na kontrastivní analýzu, chybovou analýzu a koncept mezijazyka.

1.3 Koncept chyby

V historii výzkumů nabývání druhého, resp. cizího jazyka a cizojazyčného vyučování jsou chyby v jazyce nerodilých mluvčích chápány jako jazykové jevy odchylující se od lingvistických pravidel a standardního užívání cílového jazyka, od jeho „mainstreamové" varianty.[19] Tradičně

[19] Mainstreamová varianta cílového jazyka, tj. ta, kterou mluví jeho rodilí mluvčí a/nebo členové etnoligvistické komunity, s níž je primárně asociován. (Termín jsem si vypůjčila od V. Elšíka, 2006.)

byly vnímány negativně, jako „nechtěné formy“,[20] které indikují selhání v jazykovém vyučování i učení.

V kontextu behaviorismu, který byl základem teorie učení v polovině dvacátého století,[21] je chyba nazírána jako jev, kterému je třeba se vyhnout, protože provedení, ale i pouhá registrace chyby mají funkci negativního, tj. nežádoucího zpevnění. Srov. i Choděra (2006: 162). Do konce šedesátých let minulého století převažoval v souvislosti s výzkumy nabývání cizího jazyka názor, že jazyková akvizice je primárně záležitostí formování nových jazykových návyků. Velká část chyb pak byla chápána jako přetrvávání návyků z mateřského jazyka v novém jazyce. Z toho důvodu se množství výzkumů aplikované lingvistiky soustředilo na komparaci prvního a cílového jazyka, která měla umožnit anticipaci a vysvětlení žákovských chyb. Ačkoli behavioristický přístup k SLA a FLT, resp. kontrastivní analýza, byly mnohokrát podrobeny kritice, je nesporné, že významným přínosem tohoto konceptu je přinejmenším změna perspektivy v pohledu na původ chyb v jazyce nerodilých mluvčích, zdůraznění interferenčních efektů mateřského a druhého jazyka a přesun explanace chyby z jazykového vyučování na jazykový transfer.

Od konce šedesátých let se náhled na akvizici cizího jazyka proměňuje. Nabývání cizího jazyka se začíná chápat jako kreativní proces konstruování systému, kterým žáci testují své hypotézy o cílovém jazyce. Chyby v žákovském jazyce nejsou vnímány jako povrchové deviace, tj. nepřínosně negativně, nýbrž jako důkazy žákovy snahy rekonstruovat pravidly řízený jazykový systém a jako možnost sledovat proces zpracovávání jazykových dat jinojazyčnými mluvčími (Corder 1967, Nemser 1971, Selinker 1972). Jinými slovy, jazyk nerodilých mluvčích není nadále považován za defektní a nedokonalý, ale začíná se zkoumat jako legitimní jazykový systém, který prochází logickými, systematickými fázemi osvojování a který se tvořivě přizpůsobuje aktuálnímu jazykovému okolí. Srov. např. Brown 1987, Littlewood 1984, Ellis 1994, James 1998 aj. Nerodilí mluvčí v procesu zkoušení, chybování a ověřování svých jazykových hypotéz aktivně konstruují soubory pravidel v návaznosti na data, jimž jsou vystaveni, a postupně přizpůsobují svůj jazyk standardu jazyka rodilých mluvčích.

[20] Viz GEORGE, H. V. *Common errors in language learning.* Newbury HP: Rowley, 1972.
[21] Resp. teorie operantního podmiňování a postupné aproximace B. F. Skinnera (viz *Verbal Behavior*, Copley Publishing Group, 1957.)

1.4 Mezijazyk

Žákovský jazyk je v odborné zahraniční literatuře označován jako *interlanguage,* resp. jako mezijazyk v literatuře domácí. Termín zavedl L. Selinker (1972), který jím charakterizoval obecný rys žákovského jazyka být samostatným systémem, který má status přechodného systému mezi jazykem prvním a jazykem cílovým (srov. s Weinreichovým (1953) termínem 'interlingvální', zde poznámka 25). Selinker ve své studii navazuje na práce dvou předních odborníků v oblasti SLA, W. Nemsera a P. Cordera. Nemser (1971) v souvislosti s fonologickou a fonetickou analýzou jazyka nerodilých mluvčích hovoří o tzv. aproximativním systému. Tímto termínem akcentuje další charakteristický rys žákovského jazyka, a to trvalé směřování k cílovému jazyku. Corder (1967) používá pro žákovský jazyk termín idiosynkratický dialekt, který zdůrazňuje konotaci, že žákovský jazyk je jedinečný vzhledem k jednomu určitému individuu, tj. pravidla žákovského jazyka daného jedince jsou příznačná a typická pouze pro tohoto jedince.

Od přelomu šedesátých a sedmdesátých let jsou tedy žáci ve své jazykové produkci interpretováni pomocí nového paradigmatu, které umožňuje, aby byli nahlíženi ne jako pasivní příjemci cílového jazyka, ale jako aktivní jednotky, které gramatická pravidla samy vytvářejí na základě už osvojených pravidel mateřského jazyka a postupně je adaptují směrem k cílovému jazyku. Tato teorie tzv. mezijazyka umožňuje uchopit chybu jednak jako žákův interpretační prvek, tj. jako prvek, který ukazuje na to, jak pravidlům rozumí, a jednak jako konstrukční prvek, tj. jako prvek, který mu umožňuje posunovat se v internalizaci gramatických pravidel směrem k cílovému jazyku. V teorii mezijazyka nejsou tedy chyby signály neúspěšné produkce, ale jsou chápány jako vnější sledovatelná evidence, že se žákův jazyk vyvíjí. Srov. Brown, 1987: 168. Hypotéza mezijazyka vedla k zásadnímu zlomu ve výzkumech nabývání druhého jazyka, především v přístupu ke kontrastivní analýze. Jazykové učení je od této chvíle vnímáno jako proces, který konstruuje mezijazyk, resp. tzv. přechodnou kompetenci[22] reflektující dynamickou povahu žákova vyvíjejícího se systému. Důraz je kladen na studium chyb v psaných a mluvených projevech jinojazyčných mluvčích. Bezchybná produkce totiž neposkytuje téměř žádné informace o aktuálním stavu vývoje mezijazyka žáků, jak uvádí Brown (1987: 169), je to jen informace o cílovém systému jazyka, který už si žák osvojil.

V následujících oddílech podrobněji představíme některé metody, které se používají pro analýzu žákovského jazyka. Výběr prezentovaných metod (kontrastivní analýza, chybová analýza,

[22] K termínu viz Corder (1967), příp. Powell (1998: 4).

analýza přirozené posloupnosti akvizice morfémů a frekvenční analýza) byl řízen tématem této práce, tj. problematikou žákovských korpusů, na jejichž základě došlo k oživení prvních dvou zmíněných přístupů, a další dva byly v souvislosti s existencí žákovských korpusů výrazně rozvinuty. Vycházíme přitom především ze studií Jamese (1998), Ellise a Barkhuizena (2009), Browna (1987) a Richardse (1974).

1.5 Kontrastivní analýza *(contrastive analysis, CA)*

Idea jazykového transferu, která má ve vývoji úvah o nabývání druhého/cizího jazyka a o cizojazyčném vyučování výsadní roli, prošla v posledních padesáti letech několika podstatnými proměnami. V padesátých letech minulého století vrcholila její popularita jako nejdůležitějšího faktoru SLA i v přístupech k jazykovému vyučování. Na konci šedesátých let byla role transferu upozaděna, protože žákovské chyby se nadále nepovažovaly za nedostatky ovlivněné působením mateřského jazyka studentů, nýbrž za výsledky kreativního procesu budování znalosti jazyka. Někteří autoři dokonce pod vlivem univerzalistických výkladů existenci transferu při nabývání cizího jazyka odmítali. V současné době je však plně přijímána role transferu jako jednoho z několika různých faktorů ovlivňujících jazykové učení i vyučování.

Myšlenka působení mateřského jazyka na jazyk studovaný, resp. koncept kontrastivní analýzy v SLA, není nijak nový a vychází z triviálního pozorování, že skupiny studentů s konkrétním jazykovým pozadím dělají určité typy chyb, a z představy, že by pro výukový proces bylo přínosné, kdyby učitel mohl v souvislosti se zaměřenými didaktickými materiály predikovat, jakým obtížím musí konkrétní skupina studentů při studiu cizího jazyka čelit. Model kontrastivní analýzy se vyvinul z chápání nabývání cizího jazyka jako budování jazykového chování a stal se v polovině minulého století hlavním přístupem k analýzám SLA a FLT.[23] Metodologicky byl pevně ukotven v behavioristické psychologii a strukturní lingvistice, resp. v deskriptivismu. Ve světovém kontextu lze uvažovat o kontrastivní analýze dvojího typu. V zámořské tradici se kontrastivní analýza etablovala po druhé světové válce jako významný exponent aplikované lingvistiky zaměřené na oblast jazykového vyučování, resp. na nabývání cizího jazyka. Proto také Fisiak (1993) postuluje pro tento směr pojmovou inovaci „aplikovaná kontrastivní analýza“.

[23] Kontrastivní přístup byl obecně přijímaným konceptem lingvistickému výzkumu první a části druhé poloviny dvacátého století (studium diachronního vývoje jazyků, výzkum dialektů, jazykové typologie, překladu, výzkum jazykových univerzálií apod.).

V evropské tradici má kontrastivní analýza, resp. kontrastivní lingvistika, dlouhou historii a je chápána spíše teoreticky jako jazykovědná disciplína bez pedagogických implikací.[24]

Z behavioristického hlediska je jazyková akvizice produktem budování návyku k řečovému chování. Nabývání druhého jazyka je pak procesem překonávání „návyku k prvnímu jazyku" novým návykem k cílovému jazyku. Hlavní překážkou úspěšného nabývání cílového jazyka je v kontextu tzv. hypotézy kontrastivní analýzy *(contrastive analysis hypothesis, CAH)*[25] interference[26] systému mateřského jazyka se systémem cílového jazyka. To znamená, že podle kontrastivního přístupu k SLA je hlavním zdrojem chyby v produkci i recepci cílového jazyka mateřský jazyk studentů. Srovnání založené na systematické strukturní analýze těchto jazyků mělo poukázat na mezijazykové rozdíly a shody, umožnit definovat problémové a bezproblémové oblasti akvizice cílového jazyka a na základě vymezených kontrastů umožnit predikovat většinu chyb, které student v cizím jazyce udělá.[27] A následně pak také stanovit, co a s jakou intenzitou by se mělo vyučovat. Přístup kontrastivní analýzy předpokládal, že problémové jevy v cílovém jazyce mohou být anticipovány a chybám lze předcházet, tzn. koncepty jazykového vyučování lze modelovat tak, aby nedocházelo k fixaci nesprávného jazykového chování. Chyby v jazyce nerodilých mluvčích jsou v kontextu kontrastivní analýzy považovány za nežádoucí defekty cílového jazyka, jež je v procesu jazykového učení vždy třeba korigovat. Kontrastivní analýza zdůrazňuje interferenční efekty mateřského jazyka a cizího jazyka. Tento náhled je však značně zúženým pojetím interference, chápané ve smyslu proaktivní inhibice,[28] a zároveň ignoruje intralingvální vlivy na učební proces; chyby, jejichž explanace není ukotvena v komparaci jazykových jevů, tj. nejsou interferenční, tento metodologický přístup podceňuje.

[24] Významně rozvinuli kontrastivní přístup k analýze jazyka i někteří členové Pražského lingvistického kroužku, V. Mathesius, B. Trnka, J. Vachek.
[25] Tzv. hypotézu kontrastivní analýzy formuloval Robert LADO ve svém stěžejním díle *Linguistics Across Cultures* (Ann Arbor, MI: University of Michigan Press, 1957, s. 2): „Those elements that are similar to his native language will be simple for him, and those elements that are different will be difficult. The teacher who has made a comparison of the foreign language with the native language of the student will know better what the real learning problems are…" Podstatnou inspirací pro Ladův koncept jazykového transferu a interference je analýza mezijazykových vlivů od Uriela WEINREICHA (*Languages in Contact: Findings and Problems*. New York, 1953).
[26] Pojmy *interference* a *transfer* jsou přejaty z teorie učení (srov. např. UNDERWOOD, B. J. Interference and forgetting. *Psychological Review,* 1957, vol. 64, s. 49–60.).
[27] Jednotlivé kroky kontrastivní analýzy: formální popis prvního a cílového jazyka; výběr jazykových oblastí a jevů pro detailnější komparaci; srovnání, tj. identifikace podobností a rozdílů; predikce chyb. Srov. WHITMAN, R. Contrastive analysis: Problems and procedures. *Language Learning,* 1970, vol. 20, s. 191–197.
[28] K teorii interference viz např. POSTMAN, L., STARK, K. Role of response availability in transfer and interference. *Journal of Experimental Psychology,* 1969, vol. 79, no. 1, s. 168–177. Nebo TARALLO, F., MYHILL, J. Interference and natural language in second language acquisition. *Language Learning,* 1983, vol. 33, s. 55–76.

1.5.1 Kritika kontrastivní analýzy

Kontrastivní analýza čelila v praxi takovému množství obtíží, že byla na místě otázka, zda je skutečně reálně aplikovatelná (srov. Wardhaugh, 1970: 124). Diskuzi vyvolávala už sama základní procedura kontrastivní analýzy, tj. srovnávání jazyků. Hamp (1968)[29] konstatuje, že kontrastivní analýza nemá etablována jednoznačná a spolehlivá kritéria pro jazykovou komparaci. James (1971) však připomíná široce užívaná kritéria překladové ekvivalence *(translational equivalence)*, která postulovali Halliday, McIntosh a Strevens.[30] Pokus o formalizaci predikce chyb v rámci kontrastivní analýzy představili také Stockwell, Bowen a Martin[31] ve své tzv. hierarchii obtížnosti *(hierarchy of difficulty)*, kterou aplikovali na angličtinu a španělštinu, předpokládali však její obecnou využitelnost. Modifikaci této hierarchie navrhl později Eckman (1977) v hypotéze diferenciálu příznakovosti *(Markedness Differential Hypothesis)*, která předpokládá, že příznakové jednotky jsou akvizičně náročnější než nepříznakové. Ačkoli tzv. teorie příznakovosti nabízí sofistikovanější metody predikce obtížnosti, resp. chyb v jazyce nerodilých mluvčích, stále je podstatně závislá na subjektivní klasifikaci. Srov. Brown (1987: 160). Predikce obtíží, kterým musí čelit student při akvizici cílového jazyka, plynoucí ze srovnání systému prvního jazyka se systémem druhého jazyka, se na začátku sedmdesátých let jevila jako do jisté míry nerealistická pseudoprocedura. Zároveň se ukázalo, že predikované obtíže se v některých případech, především na rovině syntaktické, neobjevily, resp. objevily se chyby, které nebyly predikovány. Srov. Gradman (1971), Whitman a Jackson (1972), Zobl[32], Dušková (1984), Odlin.[33]

Wardhough (1970) proto odděluje silnou, prediktivní variantu kontrastivní analýzy *(strong version)* od její redukované podoby *(weak version)*, která má explanační povahu a klade významně menší nároky na kontrastivní teorii. Tzv. redukovaná verze kontrastivní analýzy vychází z dokladů opakovaných chyb v jazyce nerodilých mluvčích a využívá tyto doklady

[29] HAMP, E. P. What Contrastive Grammar Is Not, If It Is. *Georgetown Monograph no. 21*. Washington: Georgetown University Press, 1968.
[30] HALLIDAY, M. A. K., MCINTOSH, A., STREVENS, P. *The Linguistic Sciences and Language Teaching*. London: Longmans, 1964.
[31] STOCKWELL, R. P., BOWEN, J. D., MARTIN, J. W. *The Grammatical Structure of English and Spanish.* Chicago: University of Chicago Press, 1965.
[32] ZOBL, H. The formal and developmental selectivity of L1 influence on L2 acquisition. *Language Learning*, 1980, vol. 30, s. 43-57.
[33] Ačkoli je tato skutečnost důvodem výrazné kritiky metodologie kontrastivní analýzy, Schachter*ová* (1971) nabízí alternativní zdůvodnění: silná verze kontrastivní analýzy (tzv. *a priori)* je v případě predikování možných obtíží při nabývání druhého/cizího jazyka neutrální ve vztahu k recepci a produkci žákovského jazyka. Neužití příslušného jazykového jevu, který by měl působit akviziční obtíže, lze také chápat jako problém recepce, který se odráží v ne-produkci.

k analýzám shod a rozdílů mezi prvním a cílovým jazykem.[34] Na rozdíl od tzv. silné verze kontrastivní analýzy se cele soustředí na produkci žákovského jazyka, ztrácí svůj prediktivní charakter a následně se stává součástí chybové analýzy. Oller a Ziahosseiny (1970: 184) v návaznosti na závěry Brierea (1968) a myšlenku hierarchie obtížnosti Stockwella et. al. (1965) postulují tzv. umírněnou, střední variantu kontrastivní analýzy *(moderate version)*, která tvrdí, že zásadní obtíže při akvizici cílového jazyka způsobují ne rozdíly mezi jazyky, ale naopak podobnosti, které se liší pouze jemnými distinkcemi, tj. interference může být větší v případě, že učené jevy jsou podobné odpovídajícím jevům v mateřském jazyce, než v případě, kdy jsou rysy prvního jazyka od rysů cílového jazyka zcela odlišné.[35] Svými závěry podtrhují zobecňující povahu učení a otvírají tak cestu výzkumům tzv. intraligválních chyb.[36] K obdobnému závěru došli např. i Buteau, Di Pietro[37], Dušková (1984).

1.6 Chybová analýza *(error analysis, EA)*

V souvislosti s nástupem nativismu, který chápe jazykové učení jako jakékoli jiné učení, tj. mimo jiné jako proces, který v sobě zahrnuje chybování, se etabloval nový přístup ke zkoumání žákovských chyb, tzv. chybová analýza. Tento přístup byl přímou reakcí na nedostatečnost kontrastivní analýzy pro objasnění principů nabývání cizího jazyka, resp. pro vysvětlení chybovosti projevů nerodilých mluvčích.[38] Jak jsem uvedla výše, chybová analýza navazuje metodologicky na redukovanou verzi kontrastivní analýzy. Jsou zde však dva základní rozdíly: za prvé chybová analýza na rozdíl od kontrastivní analýzy nesleduje žákovský jazyk jako nedokonalou verzi cílového jazyka, to znamená, že zatímco redukovaná verze kontrastivní analýzy stále vychází z vnímání chyby jako nežádoucího defektu, chybová analýza dává žákovské chybě nový status klíče k porozumění procesu jazykové akvizice. Za druhé chybová

[34] Toto rozdělení striktně odmítá např. James (1998: 180), který za kontrastivní analýzu považuje pouze proceduru prediktivní, tj. silnou verzi CA. Tzv. 'diagnostickou' verzi CA, tj. redukovanou, chápe jako součást chybové analýzy a nazývá ji analýzou transferu *(transfer analysis)*.

[35] Problematiku negativního vlivu podobnosti struktur na učení a používání jazyka zmiňuje na počátku dvacátého století již maďarský psycholog P. Ranschburg ve svých statích o dyslexii, např. *Ähnlichkeit biem Lernen,* 1905. Srov. také tzv. *podobnostní útlum* (NAKONEČNÝ, M. *Encyklopedie obecné psychologie.* Academia, Praha, 1997, s. 209.).

[36] Je otázka, zda lze tuto tzv. umírněnou verzi považovat za variantu kontrastivní analýzy, či zda by nebylo případnější chápat ji jakou novou teorii v přístupu k žákovskému jazyku. Srov. i Majid Hayati (1997).

[37] BUTEAU, M. F. Students' errors and the learning of French as a second language – a pilot study. *IRAL*, r. 8, 1970, s. 133-145.
DI PIETRO, R. J. Kurze orientierende Bemerkungen zur Untersuchung sprachlicher Verschiedenheit. In *Reader zur kontrastiven Linguistik.* Ed. G. Nickel. Frankfurt am Main: Athenaum, 1972, s. 136-146.

[38] Kontrastivní analýza byla částečně úspěšná v oblasti fonologické, syntaktické, sémantické a lexikální interference lze predikovat s mnohem menším úspěchem, protože podstatným faktorem na těchto rovinách je kognitivní koordinace (tj. myšlení, zpracování, ukládání, vyvolávání apod.).

analýza nedává chyby do souvislosti s komparací cílového a mateřského jazyka, resp. nevysvětluje chyby v jazyce nerodilých mluvčích prizmatem prvního jazyka studentů. Srovnává naopak žákovskou podobu osvojovaného jazyka s formami cílového jazyka.

Corder (1967:163n.) ve svém přelomovém článku „The Significance of Learners' Errors" a textech následujících[39] vymezil v několika bodech základy chybové analýzy. Nabývání druhého jazyka dává Corder do vztahu s nabýváním prvního jazyka a žákovské chyby chápe obdobně jako chyby v mateřském jazyce dětí, tj. jako odchylky, které mají svou logiku v rámci jazykového systému, v němž se nerodilý mluvčí, resp. dítě, pohybuje. Na základě pozorování a zpětné vazby testuje žák hypotézy o podobě osvojovaného jazyka a posouvá svůj aktuální jazykový systém směrem k odpovídajícímu cílovému jazyku. Výzkumy nabývání žákovského jazyka, v předchozím období úzce zacílené na vnější aspekty procesu jazykové akvizice, tj. především na jazykové vyučování, by se podle Cordera měly soustředit na jazykové učení. Analýzy externích vlivů by měly být upozaděny ve prospěch zkoumání vnitřních mentálních procesů žáka, jinými slovy bádání se má méně zaměřovat na povahu jazykového vkladu a výrazněji sledovat povahu zpracování tohoto vkladu (tzv. žákův příjem, *intake*). Chyby jsou pak v této souvislosti chápány jako důkazy žákova tzv. zabudovaného sylabu *(built-in sylabus)*, který (spolu)určuje posloupnosti při jazykové akvizici.

Corder také navrhuje oddělit chyby systémové *(errors)* od chyb nesystémových *(mistakes)* a při analýzách SLA se zabývat pouze problematikou nedostatků systémových, které jsou součástí tzv. přechodné kompetence *(transitional competence)*, již Corder chápe jako reprezentanta autonomního jazykového systému s vlastními pravidly. Tento žákovský jazykový systém pak nazývá *idiosynkratický dialekt.* Corderův náhled na povahu chyby a žákovský idiosynkratický dialekt se v hlavních rysech shoduje s tím, co Selinker (1972) definuje jako mezijazyk, tj. mentální gramatiku, kterou žák konstruuje na rozdílných úrovních akvizičního procesu.

Žákovské chyby, resp. jejich analýza je důležitá nejen pro badatele v oblasti SLA, ale také pro pedagogy a samotné studenty. Jak uvádí Corder (1967: 167) může analýza chyb pomoci při výzkumech SLA objasnit proces, jímž je cizí jazyk osvojován, a registrovat druhy akvizičních strategií, které žáci uplatňují. V pedagogické praxi umožňuje sledování chyb učiteli reflektovat, které jevy si žák osvojil a které nikoli, a současně evaluovat výukové materiály a techniky. V kontextu jazykového učení jsou žákovské chyby ve spojení s odpovídající zpětnou vazbou prostředkem k objevování jazykových pravidel cílového jazyka, tj. k ověřování hypotéz mezijazyka.

[39] Texty z přelomu šedesátých a sedmdesátých let minulého století byly souhrnně publikovány v roce 1981 v CORDER, P. *Error Analysis and Interlanguage.* Oxford: Oxford University Press, 1981.

1.6.1 Neznalost cílového jazyka[40]

Předmětem chybové analýzy je tedy výzkum chyb v jazyce nerodilých mluvčích, resp. analýza žákovského jazyka ve vztahu k cílovému jazyku. James (1998: 62) upřesňuje, že jde o studium tzv. lingvistické ignorance, tj. žákovy neznalosti cílového jazyka. Tato neznalost se v jazyce nerodilých mluvčích projevuje dvěma způsoby. Za prvé jde o tzv. strategii vyhýbání se *(avoidance),* která ze své povahy nemůže být předmětem chybové analýzy (viz kritika Schachter*ové* a Celce-Murcia*ové,* 1977). Za druhé se jedná o využívání kompenzačních strategií, substitucí, které jsou inherentní součástí žákovského mezijazyka. Tento „substituční jazyk", jak jej nazývá James (1998: 63), zahrnuje i žákovské chyby. Již v oddíle 1.2 jsme však uvedli, že jednoznačně definovat chybu v žákovském jazyce není snadné. Pro charakterizaci nedostatků v projevech nerodilých mluvčích lze aplikovat kritérium tzv. gramatičnosti *(grammaticality),* ať již je chápána dichotomicky, či gradientně, a/nebo kritérium přijatelnosti *(acceptability)* jako hlediska, které reflektuje intenci mluvčího a hodnocení recipienta.[41] K pojmům gramatičnost a přijatelnost v souvislosti s problematikou SLA viz např. James (1998:64n.).[42] Pokud je upřednostněno kritérium gramatičnosti, je chyba definována jako porušení (gramatických) pravidel cílového jazyka. Volba tohoto kritéria se může jevit jako výhodná a objektivní, jeho aplikace však nemusí být snadná. Otázkou je především výběr variety jazyka, která by měla sloužit jako referenční norma. Výběr variety podstatně ovlivňuje především klasifikaci nedostatků fonologických a sémantických, které jsou v procesu chybové analýzy také obvykle zahrnovány pod široce vymezený pojem 'gramatiky'. Kritérium přijatelnosti reflektuje náhled účastníka komunikace a kontextuální ukotvenost výpovědi nerodilého mluvčího. Ve shodě s dalšími odborníky (např. Lyons[43]) chápu gramatičnost, resp. gramatickou správnost jako základní předpoklad přijatelnosti, ačkoli např. Borsley (1991)[44] uvažuje o existenci negramatických, ale z pragmatického hlediska přijatelných konstrukcí. Tato myšlenka by v rámci teorie nabývání cizího jazyka zasloužila podrobnější rozpracování.

[40] K termínu 'ignorance' viz např. James (1998: 62).
[41] K pojmům 'gramatičnost' a 'přijatelnost' viz dále např. CHOMSKY, N. *Aspects of the theory of syntax.* Cambridge: MA, 1965.
Vyhraněnou polarizaci obou kritérií při snaze o definici pojmu 'chyba' v jazyce nerodilých mluvčích, jak ji prezentují např. Ellis a Barkhuizen (2009: 56), nepovažujeme v kontextu SLA za vhodnou.
[42] James uvádí kromě výše uvedených kategorií vymezujících pojem lingvistické neznalosti nerodilých mluvčích ještě správnost *(correctness)* a „divnost" *(strangeness).*
[43] LYONS definuje gramatičnost jako „that part of the acceptability of utterances which can be accounted for in terms of the rules". In *Introduction to Theoretical Linguistics*. London: Cambridge University Press, 1968.
[44] BORSLEY, R. D. *Syntactic Theory: A Unified Approach.* London: Edward Arnold, 1991.

1.6.2 Chyby systémové *(errors)* a nesystémové *(mistakes)*[45]

Corder (1967)[46] odděluje *chyby/mistakes* jako v jistém smyslu přeřeknutí od *chyb/errors* jako systémových nedostatků, aplikuje Chomského distinkci a spojuje *chyby/mistakes* s performančním selháním a *chyby/errors* se selháním v kompetenci. Brown (1987) obdobně vymezuje termín *chyba/mistake* jako referující k samotnému užití, tj. k oblasti performance, zatímco *chyba/error* je pro něho systémová chyba, resp. nedostatek v osvojení systému cílového jazyka.[47] Tzv. nesystémové chyby v corderovském pojetí tedy nevyplývají z deficitu jazykové kompetence, ale jsou důsledkem momentálního selhání v produkci jazyka, problémem zpracování, který tkví např. v nedostatku automatizace, limitovanosti paměti apod. a brání mluvčímu použít osvojenou znalost cílového jazyka. Tato zaváhání, přeřeknutí, aktuální negramatičnosti a jiná performanční pochybení se objevují nejen v řečové produkci cizince, ale i v produkci rodilého mluvčího. Mluvčí je schopen nesystémové chyby odhalit a v případě potřeby je korigovat. Nesystémové chyby nemají v procesu jazykového učení žádnou relevanci (srov. Corder 1967: 167). Tzv. systémové chyby na druhou stranu se v řeči nerodilého mluvčího objevují opakovaně jako nezáměrně deviantní případy indikující nedostatek znalosti a jsou integrální součástí žákova mezijazyka. Jako chyby je lze uvažovat pouze v souvislosti s externí normou, což je v případě analýz osvojování jazyka norma rodilého mluvčího. Např. jestliže

[45] Vzhledem k tomu, že při diskusi o chybě v žákovském jazyce dochází k výraznému terminologickému nedostatku při porovnání češtiny a angličtiny, je třeba ozřejmit, jakým způsobem pracuji s označením *chyba* vzhledem k anglickým termínům *mistake* a *error.* V souladu s terminologií české verze SERR (Společný evropský referenční rámec pro jazyky, 2001) užívám pro termín *mistake* český ekvivalent *nesystémová chyba,* pro termín *error* pak pojem *systémová chyba.* V oddíle 1.6.2 však zároveň používám pro potřeby jednoznačného vysvětlení obou termínů i varianty *chyba/mistake* a *chyba/error*.

[46] Ačkoli byl rozdíl v chápání *mistakes* a *errors* v novodobé debatě o charakteru žákovských chyb představen na konci šedesátých let (Corder, 1967), idea tohoto rozdílu není nová a není ani omezena na analýzu angličtiny jako druhého jazyka (srov. německé *Fehler* a *Irrtum*, francouzské *faute* a *erreur).*
Zároveň je třeba předeslat, že v případě prezentace charakteristiky žákovských chyb došlo u Cordera k významnému posunu. Oproti polarizaci *error – mistake* (1967), jak ji uvádíme zde, v textu z roku 1971 (v naší práci cit. z roku 1974: 161), tj. po revizi přístupu k chybové analýze, Corder nazývá původní *mistakes* chybami *(errors)* a původní chyby hodnotí jako rysy žákova mezijazyka (resp. idiosynkratického dialektu), tj. jako „nechybové" ID formy.

[47] Na Cordera navazují další autoři podrobnějším rozlišením obecných typů žákovských chyb. Edge (1989: 11) používá termín *mistake* jako obecný termín zahrnující všechny typy chyb v jazyce nerodilých mluvčích (tj. totéž, čemu James (1998: 77) říká deviace) a dále vymezuje přeřeknutí (tj. *slips),* chyby, které může žák opravit sám; nedostatky (tj. *errors),* chyby pramenící z nedostatečné internalizace naučeného pravidla a pokusy (tj. *attempts),* chyby vyplývající primárně z neznalosti pravidla. James (1998: 83n.) prezentuje jazykové deviace nerodilých mluvčích jako lapsy (tj. *slips),* jednoznačná přeřeknutí nebo přepsání; nedostatky (tj. *mistakes),* které mohou být po upozornění sebekorigovány; chyby (tj. *errors),* které nelze bez dodatečného učení sebeopravit a solecismy, primárně ve smyslu hyperkorektního vyjádření. Norrish (s. 8) odlišuje chyby (tj. *errors),* jako systematické odchylky od cílového jazyka; nedostatky (tj. *mistakes),* kolísavé odchylky pramenící v nedostatečném osvojení pravidla a přeřeknutí (tj. *lapses),* vznikající nedostatečnou koncentrací. Další rozvíjející kategorizace viz např. i Hammerly. V ruské tradici rozlišuje např. Menčinskaja tzv. chyby poučné, které do jisté míry odpovídají corderovským *mistakes,* a chyby hloupé, jež by bylo možno ztotožnit s *errors.*
HAMMERLY, H. *Fluency and Accuracy: Toward balance in language teaching and learning.* Clevedon: Multilingual Matters, 1991.
NORRISH, J. *Language learners and their errors.* London: Macmillan, 1987.
MENČINSKAJA, A. N. 50 let sovetskoj psichologii obučenija. *Voprosy psichologii,* 1967, no.5, s. 71-88.

student češtiny jako cizího jazyka pronese větu *„Já budu koupit dům“*, reflektuje pravděpodobně svou kompetenční úroveň, na níž tvorba budoucího času v češtině vyžaduje použití auxiliáru *být*, avšak nereflektuje diferenci tvorby futura u dokonavých a nedokonavých sloves. V tomto případě se žák pravděpodobně dopustil systémové chyby, tj. chyby, která poukazuje na úroveň kompetence v cílovém jazyce.

Odlišení obou typů chybování není bezproblémové. Test sebekorekce, který by měl indikovat tzv. chyby nesystémové, je nejednoznačným kritériem, např. v případě, když žák není schopen ve svém projevu chybu identifikovat, ale pokud je mu chybný tvar ukázán, dokáže jej adekvátně opravit. Stejně tak je problematický vztah mezi znalostí, resp. kompetencí a tzv. systémovou chybou. James (1998: 80) odmítá striktní polarizaci znalost vs. neznalost a uvažuje o tzv. částečné znalosti *(partial knowledge)*[48], která konvenuje s myšlenkou vyvíjejícího se mezijazyka. V mnoha případech navíc rozlišení mezi nesystémovou a systémovou chybou není možné. Např. ze samotné chyby *„já kupovám“*, pokud se žák v zápětí sám neopraví, není možné zjistit, zda si doposud příslušné konjugační paradigma neosvojil, zatímco s prézentním konjugačním typem -á se již seznámil, nebo zda se jedná o pouhou momentální záměnu.

Dichotomii nesystémová a systémová chyba je věnována ve výzkumech SLA, resp. v analýzách mezijazyka značná pozornost a řada autorů Corderovu klasifikaci (viz zde pozn. 47) domýšlí. A to i navzdory faktu, že oddělování obou typů je komplikované, často subjektivní a dosah tohoto teoretického konstruktu na výukovou praxi je minimální. Jak např. uvádí Ellis (1985:68), distinkce mezi nesystémovou a systémovou chybou je reálně nepozorovatelná.

1.6.3 Proces chybové analýzy

Standardně je proces chybové analýzy rozfázován do několika dílčích kroků (viz Corder 1974, Ellis 1994 aj.). Jde o (1) sběr dat, tj. vzorků žákovského jazyka; (2) identifikaci chyb; (3) popis chyb, tzn. především jejich klasifikaci a kategorizaci; (4) explanaci chyb; (5) evaluaci chyb, která se ale obvykle vymezuje jako samostatná disciplína navazující na chybovou analýzu.

1.6.3.1 Sběr dat

Podoba dat sbíraných pro chybovou analýzu žákovského jazyka je ovlivňována mnoha různorodými faktory. Ty lze rozčlenit na faktory týkající se žáka, jazyka a produkce. Míra podrobnosti evidence těchto faktorů závisí na typu vzorku a cíli chybové analýzy.

[48] Shodně např. s Shaughnessy*ovou*, 1977: 190; Johnsonem, 1988: 90; Ellisem, 1994: 51 a dalšími.

1.6.3.2 Identifikace chyb

Identifikovat chybu v projevu nerodilého mluvčího znamená – v návaznosti na definici chyby, kterou jsem uvedla výše – porovnat produkci nerodilého mluvčího s tím, co by v daném kontextu a za obdobných podmínek produkoval rodilý mluvčí. Takové srovnání lze provést pomocí rekonstrukce zkoumaného vzorku, jak by byl prezentován rodilým mluvčím. Tento rekonstrukční krok je však sám o sobě problematický, protože obvykle nemáme k dispozici tzv. autoritativní interpretaci, tj. směrodatnou rekonstrukci chybového textu ověřenou u nerodilého autora projevu.

V příkladu (1) uvádíme tři možné rekonstrukce původní žákovské věty.[49]

(1) *Cítil jsem si dobře, protože čerstvý vzduch.

A. *Cítil jsem se dobře, protože byl čerstvý vzduch.*
B. *Cítil jsem se dobře díky čerstvému vzduchu.*
C. *Na čerstvém vzduchu jsem se cítil dobře.*

Všechny uvedené rekonstrukce jednoznačně reflektují chybu ve formě zvratného zájmena, neshodují se však již v identifikaci chyby v syntaktické struktuře věty. Interpretace A. detekuje chybu jako chybějící predikát (určitou formu slovesa) ve vedlejší větě, v B. je chyba určena jako chybně, tj. zde vedlejší větou, vyjádřený komplement, v C. je navíc identifikován i chybný slovosled. Problém je, že nedokážeme bez konzultace s autorem chybového textu (a v mnoha případech ani při diskuzi s ním) dovodit, co přesně chtěl nerodilý mluvčí vyjádřit.

Jinou komplikací související s identifikací chyb je rozhodování, do jaké míry je žákovská konstrukce chybná, resp. nepřijatelná. Ellis a Barkhuizen (2009: 59) zavádějí pojmy ‘prostá chyba’ *(absolute error)* a tzv. nepreferované formy *(dispreffered form)* a navrhují, že by se chybová analýza měla zaměřovat pouze na první zmíněný typ, protože identifikace tzv. nepreferovaných forem je při posuzování přijatelnosti vyjádření výrazně subjektivní.

Corder (1974) představuje základní model identifikace chyb v cizím jazyce a činí zásadní rozdíl mezi chybami zjevnými *(overt)* a skrytými *(covert).* Promluvy obsahující zjevnou chybu jsou na první pohled negramatické na větné rovině, resp. v češtině je tato chyba zjevná na rovině syntagmatu *vidím mamince* a někdy i níže na rovině samotného slova *vidím tatínce.* Skryté chyby jsou chyby, které sice vytváří gramaticky správně utvořenou větu, avšak jejich chybovost je interpretovatelná až v rámci komunikačního kontextu. Například věta *Jsem student* je zcela

[49] Všechny příklady, které uvádím v této práci, jsou autentické a pochází ze vzorku dat určeného k pilotní anotaci. Viz dále kapitola 9.

gramaticky správná, objeví-li se však jako odpověď na otázku A: *Co děláte večer?* B: **Jsem student* je zcela chybná a ukazuje na žákovu chybu v recepci, totiž na očekávání otázky *Co děláte? – Jsem student.*

1.6.3.3 Popis chyb a chybové taxonomie

Deskripce žákovských chyb se zabývá srovnáváním toho, jakým způsobem se formy produkované nerodilými mluvčími liší od forem produkovaných jejich rodilými protějšky (srov. Lennonova (1991) definice žákovské chyby). Tj. podle Cordera (1974: 128) je popis chyby v zásadě komparací dat původních chybových projevů nerodilých mluvčích a dat provedené rekonstrukce. Účelem tohoto kroku chybové analýzy je především potvrzení intuice rodilého mluvčího o defektnosti žákovského projevu explicitní konkretizací chyb. Zároveň je popis chyb nezbytným předpokladem pro kvantifikační a statistické analýzy a v nespolední řadě slouží fáze popisu chyb k vytvoření systému chybové kategorizace, založené na odpovídajícím lingvistickém přístupu.

Pro účely chybové analýzy je rozpracováno několik systémů klasifikací chyb odrážejících různorodá hlediska v přístupu k podobě žákovského jazyka, pro které se standardně užívá název *chybové taxonomie.* Obecně platí, že taxonomie by měla být organizována na základě určitých základních kritérií, která reflektují pozorovatelná, objektivní fakta o jevech, jež mají být kategorizovány, viz James (1998: 102). Dulay*ová* et al. (1982: 146n.) ve své práci pojednávají čtyři základní typy chybových taxonomií – A. taxonomii odrážející povrchovou realizaci, B. taxonomii založenou na lingvistických kategoriích, C. komparační chybovou taxonomii a D. taxonomii na základě komunikačního efektu. První dva typy lze charakterizovat jako čistě deskriptivní klasifikace, tzv. komparativní taxonomie jistým způsobem slučuje krok deskripce a explanace chyb, tzv. taxonomie komunikačního efektu se naopak dotýká evaluace žákovských chyb.

1.6.3.3.1 Taxonomie podle povrchové realizace

Tzv. taxonomie dle povrchové realizace (*surface strategy*)[50] klasifikuje chyby podle strukturních deformací žákovských projevů. Na základě tohoto konstruktu jsou chyby standardně klasifikovány jako tzv. vynechání (*omission*), přidání (*addition*), užití chybné formy (*misformation*) a chybný slovosled (*misordering*).

[50] Termín *surface strategy* uvedli Dullay*ová* et al. (1982), James (1998) užívá alternativní pojem *target modification,* protože považuje původní termín za zavádějící, zakládající nebezpečí tzv. srovnávacího omylu (*comparative fallacy*).

Vynechání se vymezuje jako absence obligatorního jazykového prvku, např. **jmenuju Namib, *Adam student.*

Tzv. chybu přidání klasifikujeme, je-li v promluvě použit nějaký nepotřebný nebo chybný element, např. **To je moje maminka. Ona se jmenuje Hana*. Dulay*ová* et al. (1982: 156) uvádějí, že tento typ chyby je prezentován „přítomností [jazykového] prvku, který se v korektní podobě jazykového projevu nesmí vyskytnout", a zároveň tvrdí, že se této chyby dopouštějí především pokročilejší uživatelé cizího jazyka. Zároveň definují tři podtypy, tj. dvojité označení *(*To je muž, kterého jsem viděl ho.)*, přílišné zpravidelnění[51] *(*Nenašli jsou to.)* a prosté přidání. Domnívám se, že uvedená definice je příliš úzká. Podle mého názoru lze pod tento typ chyby zařadit i užití nadbytečného elementu, který nemusí být z gramatického hlediska nutně klasifikován jako chybný.

Užití chybné formy (*misformation*) indikuje chybu ve formě slovního tvaru nebo syntaktické struktury (např. **Spám dlouho*). Dulay*ová* et al. (1982: 156) člení tuto chybovou kategorii dále na přílišné zpravidelnění[52], užívání protoforem *(*Vidíš já?)* a alternujících forem, tj. záměnné užívání různých tvarů (např. u osobních zájmen). Domníváme se, že vydělování chyby typu 'přílišné zpravidelnění' zvlášť u přidání a zvlášť u užití chybné formy je problematické. Navíc Dulay*ová* et al. (ibid.) nestanovují jednoznačná kritéria pro vzájemné odlišení tohoto typu chyby. Do kategorie užití chybné formy se standardně řadí i tzv. chybný výběr *(misselection),* např. **Ztratil jsem silnici (→cestu),* příp. je toto kritérium uváděno separátně.

Dalším zástupcem tzv. povrchově realizovaných chyb je špatný slovosled, např. **Jak jmenuješ se?*

James (1998: 111) navrhuje zařadit ke kritériím taxonomie podle povrchové realizace ještě tzv. blendy (*blends,* příp. také jako kontaminace, hybridizace), chybu ve výběru jedné ze dvou sémanticky obdobných struktur, např. **Mám rád tancuju.*

1.6.3.3.2 Taxonomie podle lingvistických kategorií[53]

Druhým typem deskriptivní taxonomie je klasifikace chyb na základě lingvistických kategorií, jejichž definice vychází z popisu cílového jazyka. Výhodou tohoto typu kategorizace chyb je fakt, že vychází z ukotveného, zavedeného popisného rámce a umožňuje proto účinnou

[51] V tomto případě je zpravidelnění chápáno jako užší vymezení problému nadměrného zobecnění.
[52] Srov. obdobnou kategorii u přidání.
[53] Jiný typ lingvisticky orientované chybové taxonomie, než o které hovoříme níže, představuje Yang (2010). Člení domény do tří skupin, na doménu substanční, textovou a diskurzivní. Návazně pak odděluje chyby v kódování, které zasahují fonologickou a grafologickou rovinu; chyby kompoziční a chyby v porozumění, které se týkají roviny lexikálně-gramatické; a chyby formulační a ve zpracování, jež se vztahují k rovině diskurzu.

pedagogickou aplikaci. V ideálním případě by lingvistická kategorizace chyby měla být hierarchická a zahrnovat informaci o jazykové rovině, v jejímž areálu je chyba lokalizována (tj. rovina fonologická, grafologická, gramatická, lexikální, textová nebo rovina diskurzu). Dále o slovním druhu zasaženého výrazu, ovšem problém při určování této subkategorie souvisí s úvahou, zda v případě, že se v tomto směru odlišuje rekonstrukce od původní podoby projevu, upřednostnit při klasifikaci rekonstrukční hypotézu nebo chybný originál. Zároveň by měla lingvistická taxonomie zahrnovat i specifikaci gramatické kategorie, která je chybou zasažena (např. číslo, čas, osoba apod.). James (1998: 105) uvádí, že by lingvistická taxonomie měla kromě výše uvedených atributů poskytovat také bližší informaci o úseku textu, na němž je chyba realizována *(rank,* např. morfém, slovo, syntagma, věta, souvětí). Problém aplikace této chybové taxonomie může nastat v případě, že chyba zasahuje rovinu lexikální, textovou, příp. diskurzivní, protože nepanuje obecná shoda ohledně dílčích subkategorií v rámci těchto domén. Při úvahách o povaze této deskriptivní chybové taxonomie zdůrazňují Ellis a Barkhuizen (2009: 60), že lingvistická kategorizace žákovských chyb by měla být řízena daty, tedy chybovými texty (tzv. *data-driven* přístup), tj. není podle nich cílem vybudovat plně vypracovanou taxonomii založenou na kategoriích deskriptivní gramatiky, ale je nutné reflektovat konkrétní chyby vyskytující se v analyzovaném projevu, resp. modelovat taxonomii podle nich.

V souladu s Jamesem (1998) považuji za výhodné kombinovat výše uvedené kategorizace chyb do souhrnné bidimenzionální taxonomie (viz v této práci dále, kapitola 6). Domnívám se však, že u obou zmíněných typů popisných chybových taxonomií je na místě uvažovat o jejich platnosti pro analýzu žákovského mezijazyka. Chápání cílového jazyka jako referenčního bodu totiž ve své podstatě odmítá představu mezijazyka jako svébytné jazykové variety.

1.6.3.4 Explanace chyb

Určení zdroje žákovské chyby je pravděpodobně nejdůležitějším krokem chybové analýzy, přinejmenším je-li chybová analýza využívána ve výzkumech nabývání cizího jazyka. V odborných studiích se vysvětlení chyby odděluje od jejího popisu, pouze v některých případech se oba tyto kroky chybové analýzy slučují.[54] Pro vysvětlení žákovských chyb se obvykle používá diagnosticky orientovaná kategorizace chyb, tzv. komparační taxonomie.[55] Tradičně se vymezují tyto čtyři kategorie: (1) interlingvální chyby na základě mezijazykového transferu; (2)

[54] Podrobnější úvahu o smyslu oddělování popisu a vysvětlení žákovské chyby viz např. James (1998: 173n.).
[55] Příp. někdy nazývaná jako výkladová klasifikace chyb.

intralingválnı́ chyby na základě cílového jazyka; (3) učebnı́ chyby (tj. plynoucı́ z kontextu výuky); (4) chyby vyplývajı́cı́ z uplatňovaných komunikačnı́ch strategiı́. Srov. i Dulay*ová* et al. (1982: 163n.), Brown (1987: 177n.), Koutiv*ová* a Storch (1989: 410n.), James (1998: 173n.), Ellis a Barkhuizen (2009: 62n.) aj. V českém prostředı́ pak Hendrich (1988: 365n.).

1.6.3.4.1 Interlingválnı́ chyby

Zdrojem interlingválnı́ch chyb je vliv mateřského jazyka nerodilých mluvčı́ch. Prvnı́ jazyk ovlivňuje podobu mezijazyka dvěma způsoby: jedná se za prvé o vliv pozitivnı́, tzv. jazykový transfer, za druhé jde o vliv interferenčnı́ (tzv. negativnı́ transfer). Interferenčnı́ chyby v projevech nerodilých mluvčı́ch majı́ kořeny v odlišnosti struktur v prvnı́m a cílovém jazyce, resp. v nepřı́tomnosti obdobných jazykových struktur v žákově mateřském jazyce. Projevujı́ se buď neúspěšným transferem z prvnı́ho do cílového jazyka (tzv. zjevná interference / *intrusive interference),* nebo neužitı́m, resp. neučenı́m těch struktur cílového jazyka, které absentujı́ v prvnı́m jazyce žáků (tzv. inhibičnı́ interference / *inhibitive interference).* Srov. Hammerlyho (1991) hypotézu inhibujı́cı́ interference a výzkumy Larsen-Freeman*ové* zabývajı́cı́ se akvizicı́ členů nerodilými mluvčı́mi angličtiny.[56]

Základnı́ Ladovu tezi (s. 2)[57], která ve zjednodušené interpretaci tvrdı́, že čı́m je daný jazykový jev odlišnějšı́, tı́m je akvizičně obtı́žnějšı́, značně modifikovali Oller a Ziahosseiny (1970) do podoby tzv. umı́rněné kontrastivnı́ analýzy (viz zde oddı́l 1.5), která tvrdı́, že obtı́že při akvizici cílového jazyka způsobujı́ spı́še podobnosti, než rozdı́ly. Na konci sedmdesátých let dále propracovali problematiku mezijazykového transferu především Kellerman a Eckman. Kellerman (1979) na základě výzkumů interlingválnı́ch chyb holandských studentů angličtiny přicházı́ s teoriı́ prototypičnosti, v jejı́mž rámci vysvětluje principy jazykového transferu, který je podle Kellermana ovlivněn vnı́mánı́m specifičnosti, resp. prototypičnosti konkrétnı́ch jazykových jevů v mateřském jazyce nerodilého mluvčı́ho a zároveň typologickou vzdálenostı́ mezi prvnı́m a druhým jazykem[58]. Eckman (1977) dospěl ve své teorii diferenciálu přı́znakovosti k obdobným závěrům a ve fonologické studii anglických a německých mluvčı́ch shrnuje, že nepřı́znakové formy prvnı́ho jazyka budou negativně (tj. chybně) transferovány do cílového jazyka v přı́padě, že ve druhém jazyce jsou přı́znakové; k transferu, tj. ani ke kontrastivnı́

[56] LARSEN-FREEMAN, D. E. The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly*, 1975, vol. 9, s. 409-419.
[57] LADO, R. *Linguistics Across Cultures.* University of Michigan Press, 1957.
[58] O vzdálenosti mezi prvnı́m a druhým jazykem, jak ji představil Kellerman, hovořı́ v českém kontextu M. Hádková (2011: 105) a nazývá ji vzdálenostı́ mezi výchozı́m a cílovým jazykem (přı́p. zmiňuje také pojem ‘index vzdálenosti mezi jazyky’) .

analýzou predikované interferenci nedojde v případě příznakových forem prvního jazyka. Viz také James (1998: 183).

1.6.3.4.2 Intralingvální chyby

Intralingvální (příp. také vývojové) chyby umožňují pozorovat proces osvojování cílového jazyka a zapojování specifických kognitivních strategií, kterými žák cílový jazyk uchopuje. Tyto chyby je možné považovat za produkt vlastního procesu učení. James (1998: 185n.) prezentuje přehled strategií učení, které jsou zdrojem intralingválních chyb. Rozlišuje
(1) tzv. mylnou analogii (*false analogy*), např. *znát / znám → spát / *spám;* (2) chybnou hypotézu (*misanalysis*); (3) aplikaci nekompletního pravidla (*incomplete rule application*), např. *Petr *je (# ×) byl ve škole*; (4) hyperkorekci *(hypercorrection*) a (5) přílišné zobecnění (*overgeneralization*), např. *To jsou dva stoly → *To jsou pět stoly*. Žák si osvojuje celé paradigma slovesa a rozlišuje užití singulárových a plurálových forem podle jednosti/mnohosti subjektu, např. *To je stůl. To jsou stoly.* Podle distribuce jednosti/mnohosti žáci uplatňují toto pravidlo i na použití číslovek pět a výše, nebo ve spojení s morfologicky plurálovými tvary substantiv typu *To *je* (*#jsou*) *moje hodinky.* Kromě zmíněných typů strategií učení zakládajících intralingvální chyby uvádí James ještě (6) tzv. zneužití redundance (*exploiting redundancy*), např. *Potkal jsem Karla. *Neznáš* (*#ho*)? a (7) přehlížení současně se vyskytujících omezení *(overlooking cooccurrence restrictions),* např. *Je to *nízko* (*#malý*) *kluk.* V daném příkladu se vyskytují současně dvě chyby, tj. chyba ve výběru lexikálního významu (výraz *nízký* nelze použít v daném kontextu) a chyba ve formální podobě zvoleného výrazu (adverbium místo adjektiva). Podle mého názoru je toto rozdělení problematické, protože James nestanoví jednoznačná kritéria, podle kterých by mohly být konkrétní chyby k jednotlivým strategiím přiřazeny. Např. užití chybné podoby préterita v příkladu *ty včera psal dopis?,* můžeme vysvětlit jako problém aplikace nekompletního pravidla, zneužití redundance, nebo přílišného zobecnění (resp. simplifikace systému).
V českém prostředí nabízí poněkud odlišnou klasifikaci Hendrich (1988). Kromě mezijazykového transferu a nedostatků v koncepci výukového materiálu či v práci učitele vyděluje následující příčiny chyb: přílišné zobecnění strukturního pravidla, tj. mylná analogie (viz Jamesovy strategie 1 a 5); vnitrojazyková interference ve smyslu chybné záměny koexistujících jevů (viz bod 7 u Jamese); objektivní obtížnost jazykové struktury, resp. její nezvládnutí (srov. Jamesův bod 2 a 3). Problém hyperkorekce a zneužití redundance Hendrich nezmiňuje.

1.6.3.4.3 Tzv. vynucené chyby

Třetím hlavním zdrojem chyb, na který se poukazuje (Brown 1987), je kontext učení, který může žáka vést k vytvoření chybné hypotézy o užití jazyka. Tento zdroj chyb je u Richardse (1974) nazván mylným konceptem (*false concept*), u Stensonové (1974) je vymezován tzv. vynucenými chybami *(induced errors).* Termíny odkazují k chybám, kterých se žáci dopouštějí z důvodu chybného vysvětlení, chybné prezentace struktury v učebnici, nebo nedostatečné kontextualizace učeného jevu. Tzv. vynucené chyby vznikají tedy v důsledku učebních aktivit samých a díky nim můžeme usuzovat na interferenci učebního a osvojovacího procesu.

1.6.3.4.4 Chyby v rámci tzv. kompenzačních strategií[59]

Dalším zdrojem žákovských chyb jsou tzv. kompenzační strategie, které jsou záležitostí výstupu, neboli jak uvádí Brown (1987: 180), v kontextu SLA se týkají využívání verbálních i nonverbálních mechanismů pro zprostředkování významu v případě, že adekvátní jazykové prostředky a formy jsou z nějakého důvodu nerodilému mluvčímu nedostupné.

Z výše uvedeného zřetelně vyplývá, že vymezení zdrojů žákovských chyb není snadné. Důvodem je skutečnost, že některé zdrojové domény nelze charakterizovat s dostatečnou přesností a zároveň že některé žákovské chyby nemůžeme jednoznačně definovat jako jednozdrojové. Množství chyb může být určeno jako tzv. chyby přechodové *(ambiguous errors),* to znamená, že mohou příslušet k několika zdrojovým doménám (např. syntaktický nedostatek v konstrukci **líbím se Prahu* lze uvažovat jako chybu interferenční, ale také jako chybu intralingvální). Z tohoto faktu plyne i názor, že žákovské chyby je třeba posuzovat jako vícezdrojové, což ovšem může působit problémy při kvantifikaci a statistických přehledech.

[59] Kompenzační strategie jsou podmnožinou tzv. komunikačních strategií, které zahrnují dva druhy procesů vyrovnávajících se s jazykovými problémy v rámci komunikace. Druhým podtypem je tzv. strategie vyhýbání *(avoidance),* kterou v rámci naší práce dále pojednávat nebudeme.
Problematika tzv. komunikačních strategií je v odborné literatuře pojednána značně nesourodě, existuje řada různorodých taxonomií podle nejednotných kritérií. Pro podrobnější informace viz např. Taroneová (1981), která vymezuje soubor komunikačních strategií zaměřený primárně na jazykovou podobu výstupu (vyhýbání, parafráze, výpůjčky, nonverbální prostředky a žádosti o asistenci). Dále pak např. Kellerman, Bongaerts a Poulisseová (Strategy nd system in L2 referential communication. In *Second Language Acquisition in Context.* Ed. R. Ellis. Englewood Cliffs: Prentice Hall, 1987), kteří podrobují kritice předchozí definice komunikačních strategií a svou taxonomii zakládají na identifikaci kognitivních procesů podmiňujících výběr strategie (rozlišují konceptuální *archistrategy* – analytické (opis atd.) a holistické (užití ekvivalentu apod.); jazykové *archistrategy* – morfologická kreativita, transfer). Vymezují také pojem kompenzační strategie. Srov. i BIALYSTOK, E. *Communication Strategies.* Blackwell: Oxford, 1990.

1.6.3.5 Evaluace chyb

Evaluace chyb není ve své podstatě ani tak jedním z kroků chybové analýzy, jako spíš samostatnou (doplňkovou) procedurou s vlastními metodami výzkumu. Jedná se o určení míry závažnosti žákovských chyb *(error gravity)* na základě předem vymezených kritérií a pomocí vybraného evaluačního nástroje. Kritéria pro evaluaci žákovských chyb obvykle reflektují frekvenci konkrétních chyb a frekvenci struktur zasažených chybou, lingvistickou klasifikaci chyb a zhodnocení efektu chyb na komunikaci. James (1998: 218, 221) ještě zmiňuje nápadnost chyb *(noticeability)* a tzv. 'faktor dráždivosti' *(irritation factor)*, který ovšem jako kritérium evaluace žákovských chyb řada odborníků odmítá (např. Albrechtsen*ová* et al.[60]). Procedura evaluace chyb se skládá z několika dílčích kroků. Za prvé jde o výběr chyb, které by měly být hodnoceny. Je možné se zaměřit na všechny chyby identifikované v procesu chybové analýzy, obvykle jsou však vybírány pouze některé typy. Za druhé je třeba zvolit metodologický rámec pro evaluaci, vybrat kritérium, podle kterého budou chyby hodnoceny, a vybrat metodu evaluační procedury (v případě hodnocení komunikačního efektu chyb lze využít např. Likertovu škálu nebo Osgoodův sématický diferenciál apod.). Dále je třeba rozhodnout, kdo bude vybrané chyby hodnotit, tj. zda to budou pouze rodilí mluvčí, zda budou mít hodnotitelé zkušenosti s výukou cizích jazyků apod., a kolik hodnotitelů je pro daný typ analýzy relevantních. Srov. Ellis a Barkhuizen (2009: 67)

Evaluace chyb byla akcentována v sedmdesátých a osmdesátých letech minulého století, avšak obdobně jako sama chybová analýza byla kritizována pro metodologické nedostatky. Především pak pro to, že se jí nepodařilo definovat stabilní, obecně přijatelnou a dostatečně podrobnou škálu pro ohodnocení míry závažnosti chyb. V současnosti je využití evaluace chyb omezeno především na pedagogickou praxi.

1.6.4 Kritika chybové analýzy

Chybová analýza byla na přelomu sedmdesátých a osmdesátých let podrobena značné kritice (Schachter*ová*, 1974; Bell, 1974; Long a Sato*vá*, 1984 atd.). Ellis (1994: 20) shrnuje, že řada studií chybové analýzy ze šedesátých a sedmdesátých let je z důvodů, které uvádím níže, nespolehlivých a těžko aplikovatelných. Jak jsem již zmínila v předcházejících oddílech, byla

[60] ALBRECHTSEN, D., HENRIKSEN, B., FÆRCH, C. Native speaker reactions to learners' spoken interlanguage . *Language Learning*, 1980, vol. 30, s. 365–396.
Na druhou stranu v souladu s Jamesem považuje např. i Johansson 'faktor dráždivosti' vedle srozumitelnosti vyjádření za jedno z hlavních kritérií pro evaluaci žákovských chyb. JOHANSSON, S. The identification and evaluation of errors in foreign languages: a functional approach. In *Errata: Paper in Error Analysis.* Ed. J. Svartvik. Lund: CWK Gleerup, 1973.

chybová analýza kritizována především v souvislosti s metodologickými problémy při identifikaci, popisu i explanaci chyb. Pro dostatečně vypovídající analýzu chyb v projevech nerodilých mluvčích je třeba připravit kvalitní a podrobnou typologii chyb a tuto typologii nelze vytvořit bez rozsáhlého korpusu žákovských dat. Takové korpusy však nebyly v době největšího rozkvětu chybové analýzy zpracovány. Zároveň také, aby mohla chybová analýza přispívat k výzkumům mezijazyka a vývoje jazykové akvizice, je žádoucí pracovat s longitudinálními soubory dat, protože bez nich jsou výstupy značně zkreslené, a mohli bychom říci i spekulativní. Longitudinální databanky pro potřeby chybové analýzy však nebyly (a vlastně ani v současné době nejsou, viz odd. 5.2) k dispozici.

Častou výhradou vůči chybové analýze byl argument, že poskytuje neucelený obrázek žákovského jazyka, protože zkoumá pouze to, v čem nerodilí mluvčí chybují, a nesleduje to, co dělají dobře. V případě, že chybovou analýzu chápeme jako jeden z dílčích přístupů k analýze žákovského jazyka a výstupy chybové analýzy jako základ pro následné výzkumy, není, domnívám se, tato kritika relevantní. Srov. i Ellis (1994: 68) a Hammarberg (34).[61]

Nejpodstatnější kritika chybové analýzy se opírá o známý výzkum Schachter*ové* (1974), který prokázal, že se chybová analýza ve své původní podobě nevypořádala se strategií vyhýbání. Schachter*ová* při porovnávání dvou skupin nerodilých mluvčích majících odlišné mateřské jazyky (čínština / japonština na jedné straně a arabština / perština na straně druhé) zjistila, že žáci, kteří méně chybovali v relativních větách, neovládali daný gramatický jev lépe, než studenti, kteří dělali více chyb, ale že se zkoumané struktury v jejich produkci vyskytovaly jen ve velmi omezeném množství. Schachter*ová* z toho usuzuje, že nerodilí mluvčí se jazykovým jevům, kterými si nejsou ve své produkci jisti, vyhýbají. V souvislosti s tímto závěrem je chybová analýza napadána i pro své úzké zaměření na produkci, která zprostředkovává jen limitovaný přístup k nabývání druhého jazyka.

Výzkumy chybové analýzy byly vždy pedagogicky motivovány a i v současnosti má studium žákovských chyb velký praktický dosah v didaktice cizího jazyka. Od počátku devadesátých let, v návaznosti na vznik prvních, obsáhlých a digitalizovaných žákovských korpusů lze sledovat obrození tohoto metodologického přístupu i ve výzkumech podoby mezijazyka a SLA.

[61] HAMMARBERG, B. The Insufficiency of Error Analysis. In *Errata: Paper in Error Analysis.* Ed. J. Svartvik. Lund: CWK Gleerup, 1973, s. 29-35.

1.7 Komplementární metody pro analýzu žákovského jazyka[62]

V předcházejícím oddíle jsem zmínila, že chybová analýza byla kritizována pro své úzké zaměření na chyby v žákovských projevech. Avšak již Corder (1975: 207) uvádí, že chybová analýza se zaměřuje na „studium chybných výpovědí produkovaných skupinami nerodilých mluvčích" a cílem kompletní analýzy žákovského jazyka je „studium celkové podoby jazykové performance individuálního žáka". Faerch (1978) nazývá tento komplexní přístup k žákovskému jazyku performanční analýzou. Široce vymezená performanční analýza je pak chápána jako analýza celkové produkce nerodilého mluvčího v cílovém jazyce, tj. zahrnuje studium chybných i korektních jevů.

1.7.1 Analýza přirozené posloupnosti akvizice morfémů *('natural oder' of morpheme acquisition)*

Jedním z dalších způsobů výzkumu performance nerodilých mluvčích je tzv. analýza přirozené posloupnosti při akvizici morfémů[63], metoda původně aplikovaná na výzkum osvojování gramatických morfémů prvního jazyka. Pro potřeby výzkumů nabývání cizího jazyka se tohoto přístupu užívá k ověřování, které z jazykových elementů byly skutečně osvojeny. Analýza se zaměřuje na zkoumání přesnosti *(accuracy)* při jejich použití a největší zájem vyvolávala otázka, zda existuje univerzální posloupnost při jejich osvojování. Krashen (1977, 2009: 12) předpokládá, že lze popsat tzv. akviziční hierarchii při osvojování morfémů nerodilými mluvčími a tuto ideu později formuloval ve své vlivné hypotéze přirozené posloupnosti *(natural order hypothesis).* V devadesátých letech minulého století rozvedli ve svých studiích otázky přirozené posloupnosti Zobl a Licera*sová,*[64] kteří se zabývali nesouladem v posloupnostech nabývání morfémů při osvojování prvního a druhého jazyka a pokusili se na jednotném základě Chomského teorie principů a parametrů[65] vysvětlit příčiny těchto rozdílů. Jiným směrem, k

[62] Pro potřeby disertační práce se v tomto oddíle zabývám pouze dalšími dvěma metodami analýzy mezijazyka, a to analýzou přirozené posloupnosti akvizice morfémů a frekvenční analýzou, které chápeme jako komplementární k chybové analýze. Jsem si vědoma, že existuje řada dalších přístupů ke zkoumání žákovského jazyka, např. funkční analýza, přístupy interakční (konverzační analýza, analýza diskurzu aj.), není však úkolem této práce je zde podrobně představit.

[63] Tato metoda jev odborné literatuře někdy zmiňována také jako analýza povinného použití *(obligatory occasion analysis).* Ellis a Barkhuizen (2009: 73).

[64] ZOBL, H., LICERAS, J. M. Functional categories and acquisition orders. *Language Learning,* 1994, vol. 44, s. 159-180.

[65] Podrobněji o teorii např. VESELOVSKÁ, L. Od bariér k minimalismu: Některé aspekty poslední vývojové změny chomskyánského modelu jazyka. *Slovo a slovesnost,* 2001, sv. 62, s. 274-292.

analýze vlivu jednotlivých proměnných, rozvádějí výzkumy přirozené posloupnosti v nabývání cizího jazyka např. Goldschneider*ová* a DeKeyser[66].

Výzkum přirozené posloupnosti osvojování morfémů aplikovaný na angličtinu jako druhý jazyk přinesl množství využitelných poznatků pro pedagogickou praxi, především pro tvorbu kurikulí, volbu adekvátních výukových metod pro danou fázi jazykového vývoje nerodilých mluvčích a pro práci s chybami vyskytujícími se v žákovských projevech. Zároveň je však tento přístup k analýze žákovského jazyka průběžně od osmdesátých let minulého století podrobován kritice, výtky jsou směřovány především k tomu, že výsledky zkoumání nelze generalizovat ani aplikovat na jiný jazyk. Viz např. Larsen-Freeman*ová* a Long (1992), Cook (1993), Kwon*ová* (2005) aj.

1.7.2 Frekvenční analýza *(frequency analysis)*[67]

Doposud zmíněné analýzy žákovského jazyka zaměřující se na výzkum způsobu nabývání znalosti cílového jazyka kladou důraz na odchylky v systému žákova interlanguage od systému cílového jazyka, resp. vyvozují závěry o akvizici jazykových elementů cílového jazyka na základě srovnání jazykového systému mezijazyka s korespondujícími formami cílového jazyka. Toto zaměření může vést k tzv. srovnávacímu omylu *(comparative fallacy),*[68] resp. k neúplnému zobrazení internalizované znalosti systému druhého jazyka, příp. k chybné interpretaci této znalosti. Bley-Vroman (1983: 14) proto tvrdí, že pokud bychom chtěli přesněji charakterizovat jazykovou kompetenci nerodilého mluvčího v případě užívání cizího jazyka, měl by být mezijazyk analyzován sám o sobě a žákovská data (tj. performance) by neměla být porovnávána s (ideální) gramatickou znalostí rodilého mluvčího (tj. kompetencí).

Jednou z analýz žákovského jazyka, které se primárně soustředí na žákovský jazyk jako takový, bez přímého odkazování ke standardu cílového jazyka, je i frekvenční analýza. Frekvenční analýzu, obdobně jako např. funkční analýzu, můžeme charakterizovat jako způsob zkoumání žákovského jazyka v souvislosti s tzv. vnitřními normami mezijazyka, které žák buduje na jednotlivých úrovních (stupních) vývoje jazykové znalosti. Tento typ analýzy se zabývá mapováním variability v performanci nerodilého mluvčího a především je prostředkem pro

[66] GOLDSCHNEIDER, J., DEKEYSER, R. Explaining the 'natural order of L2 morpheme acquisition' in English: a meta-analysis of multiple determinants. *Language Learning,* 2001, vol. 51, no. 1, s. 1-50.

[67] Někdy prezentovaná jako tzv. analýza mezijazyka *(interlanguage analysis).*

[68] Termín poprvé uvedl Bley-Vroman (1983), který tvrdí, že analytické přístupy k výzkumu mezijazyka založené na vztahu k cílovému jazyku neumožňují systematický popis žákovského mezijazyka a mohou vést i k chybným závěrům ohledně povahy interlanguage a principů nabývání cizího jazyka. Srov. např. i LAKSHMANAN, U., SELINKER, L. Analysing interlanguage: how do we know chat learners know? *Second Language Research,* 2001, vol. 17, s. 393-420.

mapování tzv. vývojových sekvencí jazykové akvizice[69] *(sequence of acquisition),* které jsou v projevech nerodilých mluvčích manifestovány tzv. přechodovými konstrukcemi *(transitional constructions),* Dulay*ová* et al. (1982). Komplikací při aplikování frekvenční analýzy je skutečnost, že tento přístup vyžaduje longitudinální výzkum, který je především časově velmi náročný. Z toho důvodu se v praxi uplatňuje tzv. kvazilongitudinální analýza (příp. pseudolongitudinální, jak uvádějí Ellis a Barkhuizen 2009: 97), při které jsou analyzována data skupiny studentů na jedné úrovni znalosti cílového jazyka. Metodologickým problémem frekvenční analýzy je i vymezení pojmu 'stupeň akvizice' *(stage of acquisition).*

1.8 Závěr

Kapitola 1 této práce stručně shrnuje základní přístupy k otázkám nabývání cizího, resp. druhého jazyka a podrobněji představuje proměny teorie chyby v jazyce nerodilých mluvčích. Slouží jako teoretický rámec pro následující výklady o principech výstavby žákovských korpusů a jejich chybových anotacích, především však korpusu češtiny nerodilých mluvčích.

Budování relativně obsáhlých, elektronických databází žákovského jazyka lze nazvat revolucí v oboru zabývajícím se problematikou akvizice druhého jazyka. Žákovský jazyk, který je klíčovým zdrojem zkoumání SLA, protože poskytuje data pro konstruování a testování hypotéz nabývání cizího jazyka, je nyní k dispozici v nebývalém rozsahu a ve formě umožňující poměrně snadné kvantitativní, ale i kvalitativní analýzy. Počátek výstavby žákovských korpusů v devadesátých letech minulého století vedl k renesanci dvou hlavních metodologických přístupů k otázkám SLA, 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[70] v modifikované podobě znovu široce aplikovány jako kontrastivní analýza mezijazyka *(contrastive interlanguage*

[69] Odlišuji tzv. posloupnost jazykové akvizice *(order of acquisition)* a sekvenčnost jazykové akvizice *(sequency of acquisition).* Analýza posloupnosti při osvojování jazykových elementů sleduje pořádek, ve kterém si nerodilí mluvčí osvojují různé jazykové jevy (např. morfémy), analýza sekvenčnosti v jazykové akvizici se zaměřuje na jednotlivé etapy při osvojování konkrétního jazykového jevu či struktury (např. negace, tázací věty). Dále viz Ellis a Barkhuizen (2009: 95).

[70] Standardně se vymezují dva přístupy k využití korpusu jako prostředku pro ověřování, exemplifikaci či budování lingvistické teorie. Pro korpusem řízený přístup, tj. *corpus-driven approach,* 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-Bonelli*ová* (2001: 86).
Pro korpusově založený přístup, tj. *corpus-based research,* 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-Bonelli*ová* (2001: 66); příp. tzv. *knowledge-based metodology* u Atkins*ová*, Clear a Osler (1991).

analysis, CIA) a počítačem podporovaná chybová analýza *(computer aided error analysis, CEA)*[71].

Počítačem podporovaná chybová analýza se ve značné míře vyhýbá nedostatkům původní chybové analýzy, které jsme zmínili v oddíle 1.6.4. 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 jako celek. 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 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. Přepokládáme 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.

Na závěr je nutné připomenout, že pro češtinu jako cizí jazyk nejsou k dispozici žádné obsáhlejší studie týkající se problematiky mezijazyka a procesu osvojování a jen malá pozornost byla věnována dílčím srovnávacím výzkumům. Více článků se věnuje obecným oborovým otázkám, např. vztahu k evropské jazykové politice, působení sociokulturních aspektů apod., a dílčím didaktickým problémům, např. metodám výuky, roli metajazyka ve výuce češtiny pro nerodilé mluvčí, roli obecné češtiny ve výuce nerodilých mluvčích, analýze učebnic, způsobům testování apod. V této souvislosti předpokládáme, že existence chybově značkovaného žákovského korpusu češtiny jako cizího jazyka poskytne badatelům v této oblasti rozsáhlejší datovou základnu pro komplexnější výzkum a možnosti pro aplikaci různých výzkumných metod.

2 ŽÁKOVSKÉ KORPUSY

V diskusi o způsobech využití korpusů v jazykovém, tedy i cizojazyčném vyučování se do relativně nedávné doby uvažovalo pouze v intencích korpusů národních, tj. souborů textů

[71] Viz v této práci oddíl 2.3.

produkovaných rodilými mluvčími.[72] Dvacetiletý dynamický rozvoj korpusové lingvistiky se však projevuje nárůstem rozrůzňování, resp. specializovaností typů korpusů, a jedním z relativně nových přírůstků do korpusové rodiny je i tzv. žákovský korpus, který se rychle stal významným nástrojem té části aplikované lingvistiky, jež se zabývá jazykovým učením i vyučováním. Česká korpusová lingvistika ani platforma zkoumající češtinu jako cizí jazyk se touto specifickou oblastí doposud nezabývaly, proto považujeme za nutné a přínosné věnovat v této kapitole a v kapitolách následujících prostor představení pojmu žákovského korpusu, nebo jak uvádí Geoffrey Leech (1998: xvi), „zdroje užitečného pro každého, kdo chce zkoumat, jak se lidé učí jazyky a jakým způsobem by se je mohli učit lépe."

Shromažďování dat

V kapitolách 2 až 7 podávám souhrnný a ucelený pohled na situaci v oblasti žákovských korpusů, tj. přehled současného stavu oboru. Jednotlivé oddíly se věnují typologii žákovských korpusů a principům jejich výstavby, zároveň jsou zde naznačeny problematické aspekty vztahující se k budování těchto korpusů. Dále jsou u vybraných žákovských korpusů porovnány způsoby značkování, tj. chybové a lingvistické anotace, a některé významné světové žákovské korpusy jsou představeny podrobněji. Pro vytvoření tohoto přehledu jsem prostudovala dostupnou odbornou literaturu týkající se dané problematiky, jejíž souhrnný seznam je uveden v závěru této práce. Na jednotlivé studie a prezentace odkazuji vždy v příslušném oddílu.

Popis standardů pro vytváření korpusů jazyka nerodilých mluvčích opírám o analýzu padesáti sedmi vybraných žákovských korpusů. Nejedná se samozřejmě o všechny existující žákovské korpusy, jde však o významný reprezentativní vzorek. Kritéria výběru pro zařazení do tohoto vzorku byla následující: (1) korpus je hotový nebo v pokročilém stádiu budování, (2) cílem korpusu je shromažďování jazyka nerodilých mluvčí za účelem SLA analýz a pedagogických aplikací, (3) existuje k němu minimálně jedna odborná studie, příp. je veřejně dostupný, (4) není vedlejším subkorpusem některého z paralelních korpusů jazyka rodilých mluvčí (tyto korpusy nejsou obvykle budovány podle přísných kritérií, jsou značně tematicky i obsahově specializované a slouží jako prostředek pro komparaci), (5) nejedná se o součást nějakého většího žákovského korpusu (mezinárodní korpusy pojednáváme jako celek). Všem institucím, které budování vybraných žákovských korpusů řídí, byl zaslán průzkumný dotazník a při

[72] V případě češtiny jako cizího jazyka je však i tento přístup pouze ojedinělý, viz diplomová práce P. Vališové (2009).

shromažďování informací o korpusech byly reflektovány všechny dostupné studie, včetně těch nejnovějších. Shromážděná data, včetně www odkazů, jsou k dispozici v tabulce 3, oddíl 5.

Základní informace o existujících žákovských korpusech je možné načerpat z několika webových stránek, které se dané problematice věnují.[73] Jedná se především o stránku lovaňské univerzity http://www.uclouvain.be/en-cecl-lcWorld, na které jsou pravidelně doplňovány informace o nově vznikajících projektech. Bohužel data uváděná na této stránce nepodléhají dostatečné kontrole, proto je velká část zde uvedených informací neaktuální, zkreslená a zavádějící. Pro potřeby této práce byly informace shromažďovány v první řadě pomocí dotazníkové metody. Strukturovaný průzkumný dotazník obsahoval deset otázek, z nichž byla část dále zpřesňována výběrem z možností (viz příloha 1). Zjišťovala jsem jméno korpusu, současný rozsah korpusu, typ anotace a dosah této anotace, objem anotovaných dat, typ chybové taxonomie, počet chybových značek, způsob anotace, úroveň znalosti cílového jazyka u zařazovaných vzorků a velikost vzorku zařazovaného do korpusu. Dotazník byl zaslán vybraným padesáti sedmi institucím, návratnost dotazníku byla 53 %, tj. odpověď jsem obdržela ze třiceti oslovených pracovišť (viz i příloha 2).

Další informace ohledně jednotlivých žákovských korpusů jsem čerpala z dostupných studií, které reflektují současný stav jednotlivých žákovských korpusů.[74] Srov. například Zinsmeister*ová* a Breckle*ová* (2010) o žákovském korpusu ALeSKO[75]; Feldman*ová* et al. (2008) o arabském korpusu nerodilých mluvčích ARIDA; Tenfjord*ová* et al. (2006) o norském korpusu ASK; Hammarberg (2010) o švédském žákovském korpusu ASU; Nicholls*ová* (2003) o žákovském korpusu CLC; Tagnin*ová* (2003) o multilingválním brazilském korpusu; Randall a Groom (2009) o arabském žákovském korpusu BUiD; Boyd*ová* (2010) o chybově anotovaném žákovském korpusu němčiny; Jantunen (2010) o finském korpusu ICFLI; Fitzpatrick*ová* a Seegmiller (2004) o americkém korpusu MELD; Stritar*ová* (2009) o pilotním slovinském korpusu PiKUST; Muehleisen*ová* (2006) o žákovském korpusu SILS; Kwon (2007) o korpusu SKELC, Shih*ová* (2000) o budování žákovského korpusu TLCE; Granger*ová* (2003a) o francouzském žákovském korpusu FRIDA; Lüdeling*ová* et al. (2008) o německém korpusu nerodilých mluvčích FALKO; Milton a Chowdhury (1994) o korpusu HKUST; Izumi et al. (2004) o korpusu japonských mluvčích angličtiny NICT JLE; Meurers a Wunsch (2010) o pokusech aplikovat lingvistickou anotaci na žákovský korpus NOCE; Han at al. (2010) o využití

[73] Např. http://www.staff.amu.edu.pl/~przemka/CLClinks.html a http://jones.ling.indiana.edu/wiki/LearnerCorpora.
[74] Podrobný přehled odborné literatury s tematikou žákovských korpusů je k dispozici i na webových stránkách lovaňské univerzity http://www.uclouvain.be/en-cecl-lcBiblio.html.
[75] Seznam korpusů včetně zkratek viz tabulka 3 v oddíle 5.2.

žákovského korpusu CHUNGDAHM pro automatickou detekci chyb; Granger*ová* (1998, 2002) o vlivném lovaňském žákovském korpusu ICLE; Atwell et al. (2003) o mluveném žákovském korpusu ISLE; Axelsson (1999) o žákovském korpusu USE; Belz*ová* et al. (2005) o multimodálním korpusu německých anglických mluvčích TELKORP; Seidlhofer*ová* (2010) o korpusu VOICE; Gilquin et al. (2007) o korpusu pro specifické účely VESPA; Van Rooy a Schäfer (2003) o aplikaci taggerů na žákovský korpus TLEC; Brand*ová* a Kämmerer*ová* (2006) o mluveném korpusu angličtiny LINDSEI; Bartning (2009) představující žákovský korpus InterFra; Hasselgren*ová* (1997) o korpusu jazyka norských studentů angličtiny EVA; Tono et al. (2001) o korpusu angličtiny japonských mluvčích JEFLL a řada dalších. Všechny studie týkající se jednotlivých korpusů, ze kterých jsem pro následující přehledy čerpala, uvádím v seznamu literatury.

V souvislosti s krátkou historií a rychlým vývojem odvětví zabývajícího se korpusy nerodilých mluvčích jsou dílčí odborné studie značně rozptýlené a tematicky roztříštěné.[76] K základním přehledovým pracím mapujícím komplexně problematiku korpusů projevů nerodilých mluvčích se řadí studie Tono*vé* (2000), Pravec*ové* (2002) a Nesselhauf*ové* (2004). Tono*vá* (ibid.) pojednává obecně o způsobech budování žákovských korpusů, stručně o možnostech jejich pedagogického využití a podrobněji se zaměřuje na žákovské korpusy angličtiny jako druhého jazyka, devět evropských (tj. ICLE, LINDSEI, LLC, PELCRA, UAM, ISLE, JPU, CLC a IBLC) a deset asijských (tj. JEFLL, CEJL, JETC, SSTcorpus, TELEC, POLY U, NTOU, MET, HKUST a Parallel corpus of Japenese learners of English)[77]. Sumarizující text Pravec*ové* (ibid.) předkládá stručný přehled deseti žákovských korpusů angličtiny (CLC, HKUST, ICLE, JEFLL, JPU, LLC, MELD, PELCRA, TSLC, USE), zmiňuje jejich dostupnost pro výzkumy a dále se dotýká problematiky jejich značkování, technického zpracování databází a částečně i problematiky nástrojů pro vyhledávání. Nesselhauf*ová* (ibid.) řeší otázky pedagogických aplikací žákovských korpusů, představuje tzv. *data-driven* přístup k jazykovému vyučování a zmiňuje možnosti využití žákovských korpusů pro vytváření didaktických materiálů. Všechny zmíněné přehledy dané problematiky však nejsou aktuální, protože byly sestavovány v první fázi budování těchto korpusů, a jsou zároveň také limitované v tom smyslu, že se zaměřují na malou skupinu vybraných žákovských korpusů, nebo jen na některé dílčí problémy spojené s jejich

[76] V této části práce uvádíme pouze texty zabývající se budováním žákovských korpusů a jejich anotací. Stručný přehled textů reflektujících možné využití žákovských korpusů (ať již lingvistické či pedagogické) uvádíme v příloze 3.

[77] IBLC = Indianapolis Business Learner Corpus, CEJL = Corpus of English by Japanese Learners, JETC = Japanese / English Translation corpus, MET = Chinese middle school students of English corpus.

budováním a využitím. Neposkytují proto přehled o dané problematice v dostatečné šíři. Je tedy možné tvrdit, že ani ve světovém kontextu nejsou k dispozici tzv. kanonické texty, které by shrnovaly a zprostředkovávaly aktuální stav poznání v této oblasti. Za výjimku lze v jistém smyslu považovat sborníky *Learner English on Computer* z roku 1998 a *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* z roku 2002, které ovšem také vycházejí z poznatků necelé první dekády existence žákovských korpusů. K novějším přehledovým studiím se řadí podrobná kapitola o žákovských korpusech v rukověti korpusové lingvistiky z roku 2008, kterou zpracovala S. Granger*ová* (2008), a jejím funkčním doplňkem je část stati R. Xiaoa tamtéž (Xiao, 2008), jehož text kategorizuje typy jazykových korpusů včetně korpusů žákovských a představuje nejznámější z nich. Není třeba připomínat, že česká odborná veřejnost se otázkami, které se pojí s tematikou elektronických žákovských korpusů, začala zabývat teprve nedávno. Z toho důvodu je pro české uživatele k dispozici pouze několik dílčích textů.[78]

2.1 Definice žákovského korpusu

První žákovské korpusy se začínají objevovat na počátku 90. let minulého století[79] a ideově navazují na dílčí anglické nedigitalizované korpusy žákovského jazyka z 60. – 70. let, které byly sbírány pro potřeby chybové analýzy.

Žákovský korpus charakterizuji v souladu s všeobecně přijímanou definicí jako elektronizovaný soubor autentických jazykových projevů produkovaných studenty daného jazyka jako jazyka cizího nebo druhého, který je shromažďován na základě jasně stanovených výstavbových kritérií pro specifické účely analýz v oblastech nabývání druhého/cizího jazyka a cizojazyčného vyučování (Granger*ová*, 2003: 465).

Uvedenou definici je třeba upřesnit vzhledem k chápání autenticity v souvislosti s daty žákovského jazyka. Sinclair (1996: 7) v kontextu budování národních korpusů staví do protikladu materiál získávaný z autentických, tj. neelicitovaných komunikačních situacích a neautentická data získávaná v experimentálních, řízených podmínkách. Autentičnost chápe jako implicitní hodnotu charakterizující kvalitu korpusu. Aplikovat toto chápání autentičnosti na žákovská data je obtížné. V cílovém jazyce, a obzvlášť v případě studia cílového jazyka mimo jeho historické území, se student zřídka dostává do autentických situací v sinclairovském smyslu, tj. získávání jazykových projevů z takových situací je maximálně obtížné. Produkci

[78] Štindlová (2011), Štindlová a Škodová (2011), Škodová (2009), Šebesta (2010, 2011a, 2011b), dále viz kapitola 8.

[79] Prvním projektem byl International Corpus of Learner English (ICLE) v belgické Lovani (1990).

studentů cizího jazyka, která je v zásadě téměř vždy více či méně elicitována, chápe např. Grangerová jako autentickou v tom smyslu, že vyplývá z autentického výukového procesu a učebních aktivit. Takto vymezená kategorie autentičnosti žákovských projevů představuje škálu zahrnující jak volnou produkci (např. tematicky neřízené eseje, neformální interview, chatovací diskuse v rámci telekolaborativních projektů apod.), tak i výrazně řízené jazykové projevy (např. čtení nahlas, dokončování vět apod.).[80] Odlišný názor na autenticitu žákovských dat prezentuje Šebesta (2010), který nepovažuje řízené jazykové projevy za autentické. Protože žákovské korpusy nejsou v souladu s atributy kvality a kvantity standardních jazykových korpusů, řadí se k tzv. korpusům speciálním, stejně jako korpusy dětského jazyka, nářeční korpusy, ale také i korpusy orální. Srov. Sinclair (1996: 7) a Granger*ová* (2002: 8), v českém prostředí Čermák a Schmiedtová (2004), Šebesta (2010).

Žákovský korpus shromažďuje projevy nerodilých mluvčích v rozsahu, který nebyl doposud pro badatele v oblasti nabývání cizího jazyka k dispozici. Stává se tak významným zdrojem i nástrojem pro analýzy cizojazyčné akvizice, umožňuje podrobnější popis žákovského mezijazyka[81] a zprostředkovává pochopení faktorů, které jej ovlivňují. Existence žákovského korpusu umožňuje např. testování existujících hypotéz o nabývání cizího jazyka na rozsáhlé a systemizované databázi žákovských jazykových projevů. Díky žákovským korpusům je možné ověřovat, zda lze závěry výzkumů o nabývání cizího jazyka generalizovat; je možné prezentovat kvantifikační analýzy (např. frekvenční statistiky) atd. Žákovský korpus se zároveň nabízí i jako funkční prostředek pro přípravu učitelů, východisko pro úpravy výukového procesu a podstatný činitel při vyvíjení nových pedagogických nástrojů a metod.

Z metodologického hlediska žákovský korpus aplikuje rámce přebírané (1) z korpusové lingvistiky, tj. kvantifikační přístup k datům a mechanismy analýzy, (2) z výzkumů nabývání druhého/cizího jazyka, především kontrastivní a chybovou analýzu,[82] a podle mého názoru i (3) z výzkumů osvojování prvního jazyka, především principy budování korpusů zaměřených na

[80] Takto je také autenticita chápána v rámci definice žákovského korpusu, kterou uvádíme výše. Míra autentičnosti jazykových projevů, které jednotlivé databanky akceptují, je jedním z rozlišujících parametrů žákovských korpusů.

[81] V textu užíváme i termín 'interlanguage'. Termín *interjazyk,* který se objevuje např. v překladu dokumentu Rady Evropy *Společný evropský referenční rámec pro jazyky* (Council of Europe, 2001), nepovažujeme za přijatelný. *Mezijazyk,* tj. *interlanguage* je termín používaný pro jazyk nerodilých mluvčích, který má výrazně individuální a dynamickou povahu. Někdy je chápán jako specifická jazyková varieta (srov. Selinker, 1972; Corder, 1981). Je charakterizován permanentním vývojem směřujícím od využívání struktur mateřského jazyka žáka k využívání struktur jazyka cílového v souvislosti s rozvojem jazykových schopností jedince. Pojem mezijazyk není jednoznačně přijímán všemi odborníky v oblasti zkoumání cizích jazyků a jejich osvojování (viz zde oddíl 1.4). V české literatuře srov. např. Hrdlička (2010: 144). Termín *žákovský jazyk* je ve svém rozsahu širší a obecnější, neboť se nevymezuje vzhledem k mateřskému jazyku žáka, ani k dosažené úrovni v jazyce cílovém.

[82] Dále k tomu Granger*ová* (2008: 259).

jazykový vývoj dětí, např. zaznamenávání metadat, metodiku sběru materiálu a jeho vytěžování (srov. i Šebesta, 2011a).

2.2 Motivace budování žákovského korpusu

K hlavním důvodům budování žákovských korpusů patří snaha shromáždit objektivní data, na jejichž základě lze popsat žákovský jazyk. Korpusy poskytují informace o specifikách jazyka určité skupiny (nerodilých) mluvčích, o jeho neshodách se standardem cílového jazyka, který je vymezován (dospělými) rodilými mluvčími,[83] a umožňují tak popis těchto odchylek od standardu.

2.3 Metody analýzy žákovských korpusů

Žákovské korpusy jsou v současnosti využívány zejména dvěma způsoby. První možností je kontrastivní srovnávání jazyka rodilých a nerodilých mluvčích, resp. srovnávání žákovských mezijazyků (tzv. *contrastive interlanguage analysis, CIA).*[84] Na rozdíl od tradiční kontrastivní analýzy se nejedná o komparaci dvou různých jazyků, ale o „komparaci toho, co nerodilí a rodilí mluvčí jazyka dělají ve srovnatelných situacích" (Péry-Woodley*ová,* 1990: 143). V zásadě jde o možnost studia žákovského korpusu na pozadí korpusu národního, což umožňuje sledovat v jazyce nerodilých mluvčích různé odchylky od standardního užití, např. frekvenční vzorce užívání prvků jednotlivých jazykových rovin v podobě nadužívání, nebo naopak nedostatečného užívání, pokud srovnáváme s rodilými mluvčími.

V rámci tohoto přístupu k žákovským korpusům jsou dále porovnávány shody a rozdíly v žákovských jazycích, tj. jde o komparaci jednoho typu interlanguage s jiným interlanguage. Subkorpusy konkrétních mezijazyků se mohou různit ve smyslu jazykového pozadí respondentů, tedy podle jejich prvního jazyka, úrovně znalosti cílového jazyka, věku, typu textů atd. Komparace mezijazyků nabízí možnost zhodnocení vlivu intralingválních a interlingválních vlivů na utváření mezijazyka, lze např. analyzovat působení jednotlivých proměnných na podobu žákovského jazyka apod.

Druhým způsobem využití žákovských korpusů je tzv. počítačem podporovaná chybová analýza *(computer aided error analysis, CEA),* která má kořeny v metodologii chybové analýzy

[83] Standardem národního jazyka rozumíme soubor jazykových prostředků, které jsou rodilými mluvčími považovány za správné. Srov. Encyklopedický slovník češtiny (s. 438) nebo Daneš et al. (1997: 13–14).
[84] Termín navrhla v polovině 90. let minulého století S. Granger*ová.* Podrobněji viz např. Granger*ová* (1996: 43n.).

šedesátých a sedmdesátých let minulého století a zabývá se studiem žákovských chyb.[85] Existence rozsáhlé elektronické databáze žákovského jazyka, je-li navíc vybavena určitým typem anotace (chybovou a lingvistickou, tj. slovnědruhovou, morfologickou, příp. syntaktickou), překračuje kritiku směřovanou k tradiční chybové analýze, která zdůrazňovala nesystematičnost a roztříštěnost soudobých kolekcí dat, problematickou generalizaci výsledků výzkumů a jejich obtížnou verifikovatelnost a také neadekvátnost prezentovaných chybových kategorizací, především jejich subjektivitu a arbitrárnost.[86] Výzkum žákovských chyb na základě korpusu nerodilých mluvčích můžeme chápat jako objektivnější aktualizaci tradiční chybové analýzy.

Dané oblasti výzkumu jsou významně ovlivněny komputační technologií, která umožňuje automatické analýzy v podobě různých typů statistického vyhodnocení (ne)užívání vybraného jevu. Pro kontrastivní zkoumání žákovského jazyka není bezprostředně nutná chybová či jiná anotace a lze pracovat s čistým (tj. surovým) žákovským korpusem. Mnoho světových žákovských korpusů proto také anotováno není,[87] příp. je částečně značkováno slovnědruhově či morfologicky (např. korpusy JPU, SILS, TLCE, ASU aj.). Studie vycházející z tohoto typu žákovských korpusů a aplikující metodologické rámce CIA jsou pak převážně analýzou lexika, kolokací, frazémů atd. a používají aplikace vyvinuté pro potřeby národních korpusů, jako jsou WordSmith Tools, KWIC, MonoConc apod.[88]

Pro počítačem podporovanou chybovou analýzu má zcela zásadní důležitost tzv. chybové značkování (tj. tagování). Chybové značkování znamená přiřazení značek se striktně definovaným významem jednotlivým chybným výrazům; je prostředkem umožňujícím následné vyhledávání v korpusu. Toto značkování je u každého korpusu, pokud jej používá, založeno na chybové typologii, jejíž vymezení je v mnoha teoretických aspektech problematické, avšak pro řešení řady výzkumných otázek nesporně užitečné, neboť poskytuje cenné kvantitativní i kvalitativní vhledy do žákovských jazykových problémů (Granger*ová,* 2003a).

2.4 Typologie žákovských korpusů

Pro budování a zpracování žákovských korpusů jsou zásadní následující klíčové oblasti: (1) metodologie, koncepce a účel korpusu; (2) sběr dat a jejich povaha; v případě sběru

[85] Viz např. Díaz-Negrillo*vá* – Fernández-Domínguez (2006: 84n.) .
[86] Viz také v této práci oddíl 1.6.4. Dále i Nesselhauf*ová* (2005: 40n.)*;* Ellis (1994: 49n.).
[87] Viz zde oddíl 5.2.
[88] SCOTT, M. *Wordsmith Tools version 5*. Liverpool: Lexical Analysis Software Ltd., 2008.
BARLOW, G. M. *MonoConc Pro 2.2* (MP2.2). Athelstan, 2002.
Pro KWIC viz např. MANNING, C. D., SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999, s. 35.

nedigitalizovaných dat i problematika jejich převodu do elektronické podoby, včetně technických aspektů; (3) anotace lingvistická (tj. připojení tagů) i nelingvistická (strukturace dat a metadat; technické zpracování). V tomto oddíle zmíníme problematiku tvorby korpusů jazyka nerodilých mluvčích z obecného hlediska, v následujících kapitolách se pak budeme věnovat jednotlivým oblastem podrobněji.

Korpusy lze klasifikovat podle různých kritérií (srov. Granger*ová*, 2008: 260n.), která v následujícím textu stručně představíme. Zároveň uvedeme vybrané světové korpusy reprezentující dané kritérium. Veškerá zobecňující tvrzení, která zde uvádím, jsou založena na konkrétních datech vyplývajících z analýzy vybrané sady korpusů, která jsou prezentována ve shrnující tabulce 3 v oddíle 5.2.

Tabulka 1: Kritéria výstavby žákovského korpusu

KRITÉRIUM		
cílový jazyk	angličtina	jiný (tj. ne angličtina)
původ	akademický	komerční
sběr dat	průřezový	longitudinální
rozsah	velký	malý
médium	psaný	mluvený
chybová anotace	ano	ne
první jazyk	jeden	více

2.4.1 Cílový jazyk

V souvislosti s cílovým jazykem lze žákovské korpusy dělit na anglické, tj. zaměřující se na angličtinu jako cizí jazyk, a ostatní, tzn. neanglické. V celkovém počtu korpusů výrazně převažují žákovské korpusy anglické (např. ICLE, LLC, HKUST, JEFLL, USE, PELCRA aj.), v současnosti se však zvyšuje i počet korpusů pro jiné jazyky, existuje např. korpus francouzský

(FRIDA), německý (FALKO), norský (ARS), italský (VALICO), čínský (LCIC), finský (ICLFI), španělský (CEDEL2) aj.[89]

2.4.2 Původ

Dalším možným klasifikačním kritériem je prostředí, ve kterém daný žákovský korpus vzniká. Hovoříme o komerčních a nekomerčních (resp. akademických) žákovských korpusech. Nekomerčních žákovských korpusů je víc a primárně se zaměřují na jednu cílovou skupinu, resp. respondenti[90] mají shodný první jazyk a příp. i stejnou úroveň znalosti cizího jazyka. Jendou z výjimek je mezinárodní akademický korpus ICLE. Komerční žákovské korpusy jsou obvykle větší a nabízejí data od respondentů s různým jazykovým pozadím a na různých úrovních znalosti cílového jazyka (např. LLC a CLC).

2.4.3 Sběr dat

Žákovské korpusy se odlišují i ve způsobu sběru dat. Většina soudobých žákovských korpusů je průřezových, tj. sbírá data od různorodých studentů v jednom časovém období. Longitudinálnímu shromažďování materiálů se systematicky věnují pouze korpusy ASU a LONGDALE.[91] V souvislosti s bádáním v oblasti vývoje interlanguage jsou však longitudinální korpusy nezastupitelné. Některé projekty se proto cíleně zaměřují na budování tzv. kvazilongitudinálních korpusů[92], pro něž jsou paralelně sbírána data od studentů se stejným prvním jazykem a na různé úrovni znalosti cílového jazyka.

2.4.4 Rozsah

Existující žákovské korpusy se významně různí i svým rozsahem. Srov. odd. 5.1.4. Největšími jsou komerční korpusy CLC (35 mil. slov), LLC (10 mil. slov.) a hongkongský korpus HKUST (25 mil. slov), k nejmenším patří slovinský PiKUST (35 tis. slov), americký MELD (100 tis. slov.) a švédský CEFLE (100 tis. slov). Velikost korpusu je u národních korpusů chápána jako

[89] Specifickým projektem je v souvislosti s cílovým jazykem Multilingvální žákovský korpus (MLC) vznikající na University of São Paulo v Brazílii, který se zaměřuje na studenty s jediným mateřským jazykem (brazilská portugalština) učící se různé cílové jazyky (angličtina, němčina, španělština), srov. Tagnin*ová* (2006).
[90] Termín *respondent* užíváme pro žáky / studenty, jejichž jazykové projevy jsou zahrnuty do žákovského korpusu.
[91] Jako longitudinální jsou prezentovány také (1) součást žákovského korpusu Falko, tzv. Georgetown University subkorpus; (2) španělská varianta mezinárodního korpusu ICLE shromažďovaná na univerzitě v Jaén. Podle našeho názoru sporně je jako longitudinální vymezován i lundský korpus CEFLE, jehož data byla shromážděna v průběhu jednoho akademického roku (2003/2004).
[92] Viz např. i Granger*ová* (2004: s. 131).

jeden z důležitých atributů reprezentativnosti, který umožňuje, resp. usnadňuje generalizaci výstupů jazykových analýz.[93] U speciálních typů korpusů, tedy i u korpusu žákovského jazyka, nelze jednoznačně tvrdit, do jaké míry je velikost relevantní. Jak uvádí de Haan (1992: 3), optimální velikost korpusu, resp. velikost vzorků je podmíněna typem prováděného výzkumu.

Získávání materiálu pro žákovské korpusy není snadné a nelze přepokládat, že by žákovské korpusy mohly dosáhnout takového objemu dat, jakým disponují korpusy národní. Zcela zásadní jsou proto pro žákovské korpusy jasná kritéria výstavby. V tomto případě jde především o proporciálnost vzhledem k rozsahu jednotlivých vzorků zařazovaných do korpusu a také vzhledem k proměnným parametrům týkajícím se respondentů (první jazyk, úroveň znalosti cílového jazyka apod.).[94]

2.4.5 Médium

Naprostá většina existujících žákovských korpusů (tj. 74 %) se soustředí pouze na psané projevy, a to především na jazyk pro akademické účely. Srov. graf 7 v oddílu 5.1.4. Do těchto korpusů jsou zahrnovány texty vysokoškolských studentů a obvykle se jedná o delší, rozsahově řízené, psané výstupy z příslušných jazykových zkoušek. Od tohoto schématu se odlišuje např. žákovský korpus TELEKORP, bilingvální korpus zahrnující neformální chatovou komunikaci mezi studenty němčiny v USA a angličtiny v Německu. Jiným typem korpusu je i vznikající český CzeSL (viz zde kapitola 8), který bude zahrnovat psané i mluvené projevy respondentů na všech úrovních znalosti cílového jazyka.

Sběr a zpracování mluvených projevů jsou v kontextu žákovského korpusu mnohonásobně obtížnější, podobně jako u korpusů národních. Mluvené žákovské korpusy, jichž není mnoho, jsou také často velmi malé (např. LEAP – 73 tis. slov, ARIDA – 8 tis. slov, ISLE – 18 hodin nahrávek, LINDSEI – 800 tis. slov). Výjimkou jsou mluvené korpusy FLLOC, který obsahuje dva miliony slov, NICT JLE také s dvěma miliony slov a MICASE zahrnující 1,8 milionů slov. Srov. oddíl 5.2, tab. 3, sloupec 5.

2.4.6 Anotace

Žákovské korpusy se liší také v tom, zda jsou či nejsou chybově anotovány, příp. v rozsahu anotace a zaměření na zpracování chyby. V případě chybově anotovaných korpusů, nebo přesněji řečeno částečně anotovaných korpusů, se uplatňují různé anotační přístupy i rozdílné

[93] Viz např. Sinclair (1996: 6): „The whole point of assembling a corpus is to gather data in quantity."
[94] Viz zde dále oddíl 3.1.

anotační formáty. Srov. oddíl 5.2, tab. 3, sloupec 6. Jedním ze specifických přístupů k anotaci je neřízená emendace bez chybové klasifikace, kterou používá americký korpus MELD. Srov. Fitzpatrick*ová* a Seegmiller (2004). Jiné korpusy se soustředí pouze na značkování vybraných jazykových jevů, např. korpus TLEC na chyby ortografické (viz i Van Rooy a Schäfer, 2003), korpus ISLE na chyby ve výslovnosti (viz Atwell et al., 2003), korpus ALeSKO na chyby syntaktické (viz Zinsmeister*ová* a Breckle*ová*, 2010) apod. Jen malá část žákovských korpusů se pokouší o systematickou, ucelenou chybovou anotaci na základě jasně vymezené chybové taxonomie, např. CLC, ICLE, FALKO. (Srov. Nicholls*ová,* 2003; Granger*ová,* 1998; Lüdeling*ová* et al., 2008.) Soudobé anotované žákovské korpusy používají v zásadě dva základní typy anotačních formátů. Majoritně se v nich uplatňuje tzv. lineární anotační schéma, odlišným anotačním formátem je tzv. několikaúrovňová distanční anotace, kterou používá ve světovém kontextu žákovských korpusů pouze německý korpus FALKO.[95]

3 ZÁSADY VÝSTAVBY ŽÁKOVSKÉHO KORPUSU

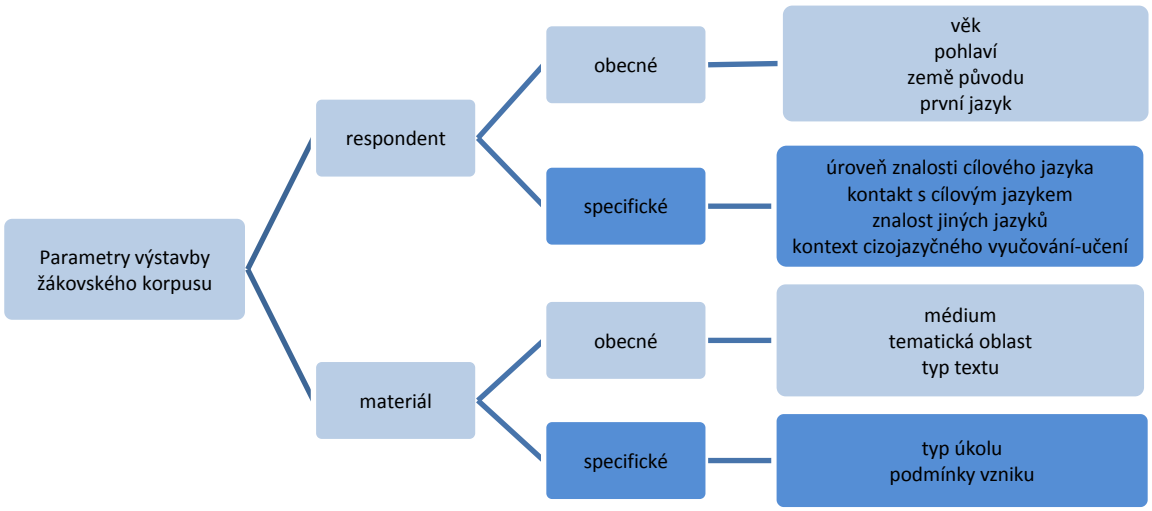
Vznik chybových textů, které jsou shromažďovány v žákovských korpusech, a tedy i korpusy samé, ovlivňuje velké množství lingvistických, sociolingvistických a situačních faktorů. Vzhledem k povaze žákovského korpusu, tj. systematické, validní databáze textů nerodilých mluvčích, je pro zachování kontroly nad těmito faktory nutné stanovit striktní designová kritéria.[96] Ztráta kontroly nad systematickou registrací těchto faktorů znamená omezení spolehlivosti výzkumu a také komplikace pro možnost sdílet korpus s jinými výzkumníky

Při budování žákovského korpusu je třeba zohledňovat především tři skupiny parametrů. Jedná se o (1) kritéria jazyková, např. klasifikace textů podle média přenosu na mluvené a psané, dle stylového vymezení či tematického zařazení textu; (2) kritéria týkající se sbíraného materiálu, např. jsou-li do korpusu řazeny texty jednoho autora za účelem longitudinálního sběru dat, informace o případné elicitaci projevu a způsobu elicitace apod.; (3) kritéria reflektující žáka, tj. sociolingvistické proměnné jako např. věk, první jazyk atd.[97] Pokud srovnáme zásady výstavby žákovských korpusů se standardními kritérii budování národních korpusů (Granger*ová,* 2008: 264), lze vyčlenit výstavbové parametry, které jsou pro oba případy shodné, ale také ty, které jsou specifické pro žákovské korpusy (viz tabulka 2). Vymezení těchto specifických parametrů není však vždy jednoznačné a v jednotlivých žákovských korpusech se různí. Srov. zde kap. 7.

[95] O anotaci žákovských korpusů viz např. Díaz-Negrillo*vá* a Fernández-Domínguez (2009) nebo Hana, Rosen, Škodová a Štindlová (2010) a dále v této práci kapitola 6.
[96] K tomu i Sinclair (1991: 9): „The results are only as good as the corpus" a Granger*ová* (1998: 7) „the quality of the investigation is directly related to the quality of the data."
[97] Viz i Tono*vá* (2003: 800).

Vedle žákovských korpusů, které poskytují podrobná metadata (např. ICLE uvádí 25 proměnných, MELD 21 apod.), existují korpusy zacílené na problematiku vyučování cizímu jazyku, jejichž externí informace jsou minimální (např. CLC uvádí pouze 6 sledovaných parametrů).

Tabulka 2: Parametry výstavby žákovského korpusu[98]



3.1 Parametry: respondent

Kromě standardního sociologického značkování[99] vztahujícího se k respondentovi (věk, pohlaví atd.) jsou pro žákovské korpusy určující parametry týkající se jeho prvního jazyka, úrovně znalosti studovaného jazyka, příp. znalost dalších cizích jazyků.

3.1.1 Úroveň znalosti cílového jazyka

Pro analýzy žákovského jazyka, resp. zkoumání vývoje mezijazyka a pro pedagogické využití žákovského korpusu je informace o úrovni dosažené znalosti v cílovém jazyce klíčová.

[98] Adaptováno podle Grangerové (2008: 264). Podstatně však rozšiřuji a upravuji obsah původních parametrů *learning context*, *L2 exposure* a *conditions*. Termín 'cizojazyčné vyučování-učení' přebírám od R. Choděry (2006: 9).
[99] Standardním sociologickým značkováním mám na mysli běžnou evidenci sociologických dat, jak je uváděna v národních (mluvených) korpusech. Viz např. ORAL2008.

V souvislosti s existencí Společného evropského referenčního rámce pro jazyky (SERR[100]) by bylo vhodné klasifikovat míru znalosti jazyka v návaznosti na standardizované popisy jednotlivých úrovní tohoto dokumentu.[101] Zjednodušila by se tím dosavadní praxe běžná pro většinu žákovských korpusů, které s hodnocením znalosti jazyka dle SERR převážně nepracují a pro popis znalosti cizího jazyka neaplikují obvykle ani standardy jiné klasifikace, jako je např. IRL, ALTE, ACTFL, STANAG 6001 apod. Existující žákovské korpusy využívají tradiční, často poměrně vágní a obtížně přenositelnou klasifikaci znalosti typu začátečník, (středně/mírně) pokročilý a pokročilý, která je případně úzce navázána na typ institucionalizované výuky a testování. Srov. graf 2 v oddílu 5.1.2. Nebo úroveň znalosti cílového jazyka vůbec nediagnostikují a nechávají na uživatelích korpusu, aby v kontextu svého výzkumného záměru stanovili úroveň dosažené znalosti podle prezentovaných externích parametrů (typ studia, délka studia, počet hodin studia apod.).

Viz např. standardy korpusu ICLE, do něhož jsou zařazovány texty respondentů, jejichž úroveň znalosti angličtiny je vymezena následovně: „Advanced students can, for the purpose of the project, be broadly defined as university students of English in their 3rd or 4th year of study. In cases where the comparability of the level is in doubt, sample pieces of writing should be submitted beforehand.“ (cit. dle http://www.uclouvain.be/en-317607.html). Korpus USE shromažďuje texty produkované studenty angličtiny „at three different levels, the majority in their first term of full-time studies“ (cit. dle http://icame.uib.no/ij24/use.pdf). Oproti tomu žákovský korpus CLC klasifikuje úroveň znalosti cílového jazyka podle jazykových zkoušek ESOL *(English for Speakers of Other Languages)* a tyto zkoušky jsou v souladu s úrovněmi SERR (např. zkouška PET odpovídá úrovni B1, zkouška FCE úrovni B2 atd.).

Nestejnorodými klasifikačními kritérii při stanovení úrovně znalosti cílového jazyka se ztěžuje validní komparace dat napříč korpusy. Resp. i analytické výsledky zkoumání založené na datech jednoho korpusu by byly obtížně porovnatelné s výsledky stejného zkoumání založeného na datech jiného korpusu.

3.1.2 Kontakt s cílovým jazykem

Tato proměnná obvykle v žákovských korpusech vypovídá o délce institucionalizovaného, primárně akademického studia a zároveň informuje o tom, zda respondent pobývá, příp. pobýval

[100] SERR, tj. Společný evropský referenční rámec pro učení se a vyučování jazykům a pro hodnocení v jazycích; resp. CEFRL, tj. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Viz např. http://www.msmt.cz/mezinarodni-vztahy/spolecny-evropsky-referencni-ramec-pro-jazyky?lang=1 .

[101] Tuto klasifikaci však lze prozatím obtížně použít pro žákovské korpusy neevropských jazyků (např. čínština, japonština), ačkoli pokusy o aplikaci dané stupnice existují.

na historickém území cílového jazyka, či nikoli. Prostý údaj o délce studia však není sám o sobě relevantní, je třeba ho doplnit především o informaci o časové dotaci jazykové výuky, způsobu studia apod.

Informace o kontaktu s cílovým jazykem umožňuje porovnávat vliv různých typů institucionalizované výuky na podobu žákovského jazyka. Možnost srovnání jazyka nerodilých mluvčích, kteří si cílový jazyk osvojují na jeho historickém území, a těch, kteří nikoli, by také mohlo usnadnit i výzkum odlišností a podobností v případě nabývání druhého a cizího jazyka (v úzkém slova smyslu).

3.1.3 Znalost jiných jazyků

Informace o znalosti jiných jazyků, než jsou jazyky první a cílový, je podstatná pro výzkum tzv. jazykového transferu a pro kontrastivní analýzu.[102] Zhodnocení úrovně však nepatří mezi objektivně měřitelné údaje, proto není ve světových žákovských korpusech akcentováno a je obvykle ponecháno na samotném respondentovi, příp. na osobě, která materiál sbírá. I v tomto případě by ale bylo možné uvažovat o využití klasifikační standardů (SERR ALTE, příp. jiné), žádný z analyzovaných korpusů však toto řešení neaplikuje. Velmi specifickým problémem v této kategorii je navíc otázka bilingvismu, resp. jeho vymezení.

3.1.4 Kontext cizojazyčného vyučování-učení

Jedná se o parametry mapující typ vzdělávací instituce, didaktické materiály, výukové metody, uplatňovaní zprostředkovacího jazyka apod. Předpokládáme, že díky těmto informacím by bylo možné porovnat míru vlivu výukových materiálů na podobu žákovského jazyka, zajímavé by bylo i srovnání produkce nerodilých mluvčích nabývajících cílový jazyk v prostředí se zprostředkovacím jazykem a bez něho.

[102] Existuje několik studií zabývajících se problematikou nabývání dalšího cizího jazyka (L3) a vlivem jednoho cizího jazyka na další studovaný cizí jazyk (tj. působení L2 na L3), resp. studující otázky multilingvalismu. Z novějších např. Bardel*ová* aFalk*ová* (The role of the second language in third language acquisition: the case of Germanic syntax. *Second Language Research,* vol. 23, no. 4, 2007, s. 459-484), Flynn*ová,* Foley*ová* a Vinnitskaya (The Cumulative-Enhancement Model for Language Acquisition: Comparing Adults' and Children's Patterns of Development in L1, L2 and L3 Acquisition of Relative Clauses. *The International Journal of Multilingualism*, r. 1, no. 1, 2004, s. 3-16) nebo Foote*ová* (Transfer in L3 Acquisition: the Role of Typology. In *Third Language Acquisition and UG*. Ed. I. Y. Leung. 2009, s. 89-114).

3.2 Parametry: materiál

Jednotlivé projevy zahrnované do žákovských korpusů jsou parametrizovány také vzhledem k povaze textu a k podmínkám jeho vzniku. V případě detailnějšího externího značkování se sleduje, zda je text součástí nějaké jazykové zkoušky, zda byl zadán povinný rozsah, byla-li tvorba textu časově limitována, příp. měl-li respondent k dispozici referenční pomůcky, jako je překladový slovník, monolingvální výkladový slovník ap. V případě, že žákovský korpus shromažďuje jazykové projevy jednoho typu, například žánrově či dokonce tematicky řízené eseje, příp. nahrávky ze standardizovaných jazykových zkoušek (jak je tomu u korpusů CLC, NICT JLE i jiných), jsou informace o materiálu přiřazovány snadno a souhrnně. Problematičtější je situace, kdy jsou do žákovského korpusu začleňovány texty různorodého původu, účelu a rozsahu.

4 CÍLE BUDOVÁNÍ ŽÁKOVSKÝCH KORPUSŮ

Žákovské korpusy se liší počtem sledovaných parametrů, i jejich obsahem a vymezením. Tato různorodost úzce souvisí mimo jiné s odlišným zaměřením a účelem jednotlivých korpusů. Na základě provedených výzkumů (dotazník a analýza sekundární literatury, viz kapitola 2) však lze v zásadě klasifikovat čtyři základní cíle, pro které jsou žákovské korpusy sestavovány.

Za prvé jde o analýzu mezijazyka a výzkumy nabývání cizího/druhého jazyka, které jako hlavní nebo vedlejší cíl postuluje většina světových žákovských korpusů (např. ICLE, JEFLL, NICT JLE, TLCE, SULEC, ASU, EAGLE, ICLFI aj.). Druhým důvodem pro budování žákovského korpusu jsou možnosti zprostředkované a případně i přímé pedagogické aplikace. Takto se prezentují především korpusy CLC, LLC, FLLOC, HKUST, LCIC. Tyto žákovské korpusy se uplatňují při tvorbě didaktických materiálů, sylabů, CALL[103] programů, které mohou být díky analýzám databáze textů nerodilých mluvčích lépe zacíleny na konkrétní typ cizojazyčných studentů. Některé z těchto žákovských korpusů (HKUST) jsou také využívány přímo ve výuce. Třetím účelem výstavby žákovských korpusů, který uvádějí např. korpusy USE, JPU, CLEC, PELCRA, BALC, SKELC, je zkoumání chyb v projevech nerodilých mluvčích, jejich diagnostika a klasifikace. Posledním z hlavních cílů budování žákovských korpusů je možnost srovnávat projevy rodilých a nerodilých mluvčích, příp. projevy různých typů nerodilých mluvčích (CEDEL2, VALICO, LINDSEI atd.).

[103] CALL, tj. computer-assisted language learning.

Vedle výše uvedených základních výzkumných cílů, ovlivňujících podobu žákovských korpusů, se některé korpusy profilují jako úzce specializované. Takový je např. žákovský korpus ALeSKO, který je sestavován pro výzkum koherence v žákovských textech, nebo korpus ISLE, který byl primárně vytvářen za účelem testování nově vyvíjeného systému diagnostiky výslovnostních chyb.

5 SOUČASNÉ ŽÁKOVSKÉ KORPUSY

Cílem tohoto oddílu je představit současné žákovské korpusy a prezentovat jejich jednotlivé parametry. Vzhledem k nejednotnosti a malé dostupnosti dílčích přehledových studií a proto, že množství dat z těchto starších prací zastaralo, považujeme tento přehled za důležitý a přínosný.
V přehledu představuji padesát sedm žákovských korpusů. Jsem si samozřejmě vědoma, že nezahrnuji všechny existující korpusy jazyka nerodilých mluvčích. V souhrnu nejsou uvedeny korpusy, které jsou v počáteční fázi výstavby (např. CzeSL), korpusy velmi malé, úzce specializované, nebo nedostupné, tj. ty, ke kterým nelze dohledat relevantní informace (např. korpusy MET, NICE, IBLC, UWI, DsA aj.).[104] Do přehledu nejsou zahrnuty ani bilingvální korpusy shromažďované v databázi mezinárodního interdisciplinárního projektu TalkBank.[105] Zároveň se samostatně nezabývám dílčími subkorpusy, které jsou součástí větších projektů, jako jsou například subkorpusy mezinárodního korpusu ICLE (tj. PICLE, GICLE, MACLE atd.). Do přehledu jsem však začlenila vybrané multilingvální korpusy, ovšem pouze ty, jejichž základním cílem je mapování žákovského jazyka.[106] Ke způsobu shromažďování dat viz dále zde kapitola 2. Zjištěné informace jsou dle jednotlivých parametrů, jichž je celkem šest (první jazyk, cílový jazyk, úroveň znalosti cílového jazyka, médium, rozsah, anotace), zařazeny do přehledné tabulky, viz oddíl 5.2. Každý parametr je dále zpracován samostatně do grafu a podrobně okomentován, viz oddíl 5.1.

Z výsledků analýzy je zcela zřejmé, že principy budování jednotlivých žákovských korpusů, přístupy ke značkování, ani nároky na programové nástroje pro zpracování korpusů nejsou ve

[104] CzeSL, tj. Czech as a Second Language (psaný i mluvený, cílový jazyk čeština). Viz zde kapitola 8.
MET, tj. čínský subkorpus projektu Corpus of English Education (psaný, cílový jazyk angličtina); NICE, tj. Nagoya Interlanguage Corpus of English (psaný, cílový jazyk angličtina); IBLC, tj. Indianopolis Business Learner Corpus (psaný, cílový jazyk angličtina); UWI, tj. University of West Indies Learner Corpus (mluvený, cílový jazyk francouzština).; a DsA, tj. SamtaleBank Dansk som Andetsprog Corpus (mluvený, cílový jazyk dánština).
[105] Projekt, jehož koordinátorem byl Brian MacWhinney z Carnegie Mellon University, probíhal v letech 1999–2004. Zahrnuje databáze: CHILDES, AphasiaBank, BilingBank, CABank, DementiaBank, PhonBank, TBIBank. http://www.talkbank.org/
[106] K dílčím informacím viz i http://www.uclouvain.be/en-cecl-lcWorld.html

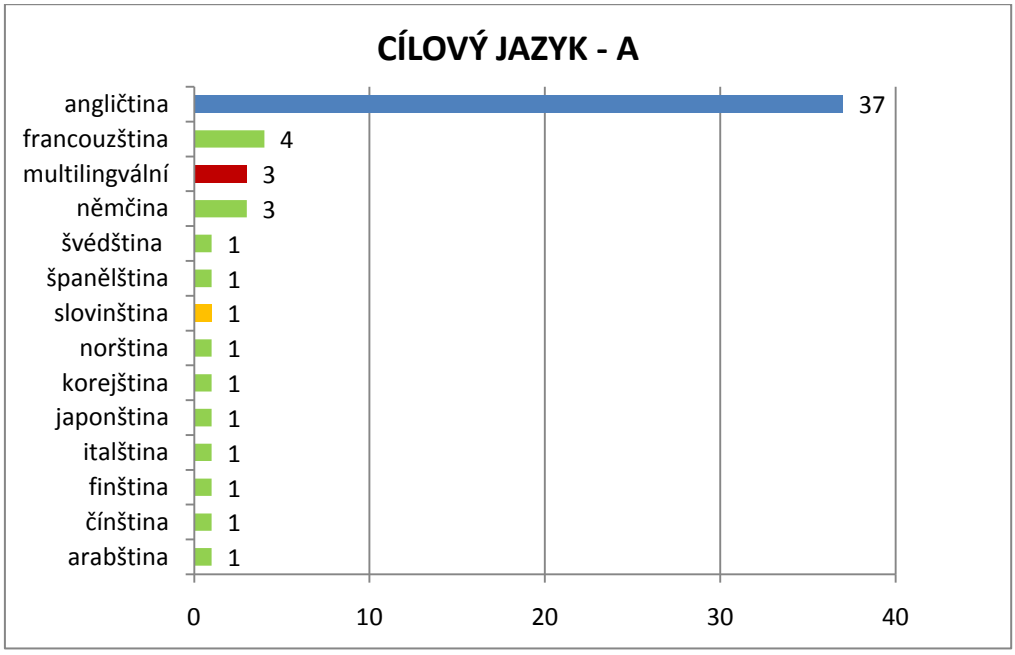
světovém kontextu jednotné a konkrétní podoba žákovského korpusu primárně odráží účel, pro který je vytvářen. Je však třeba si uvědomit, že ke korpusu přistupují jiní badatelé, učitelé i sami žáci s vlastními, novými hypotézami a analýzami. Z toho důvodu by měl být žákovský korpus, jehož cílem je zkoumání produkce nerodilých mluvčích, výzkum nabývání druhého jazyka a zároveň i pedagogická aplikace, dostatečně flexibilní a jeho zpracování uživatelsky přínosné.

5.1 Parametry současných žákovských korpusů

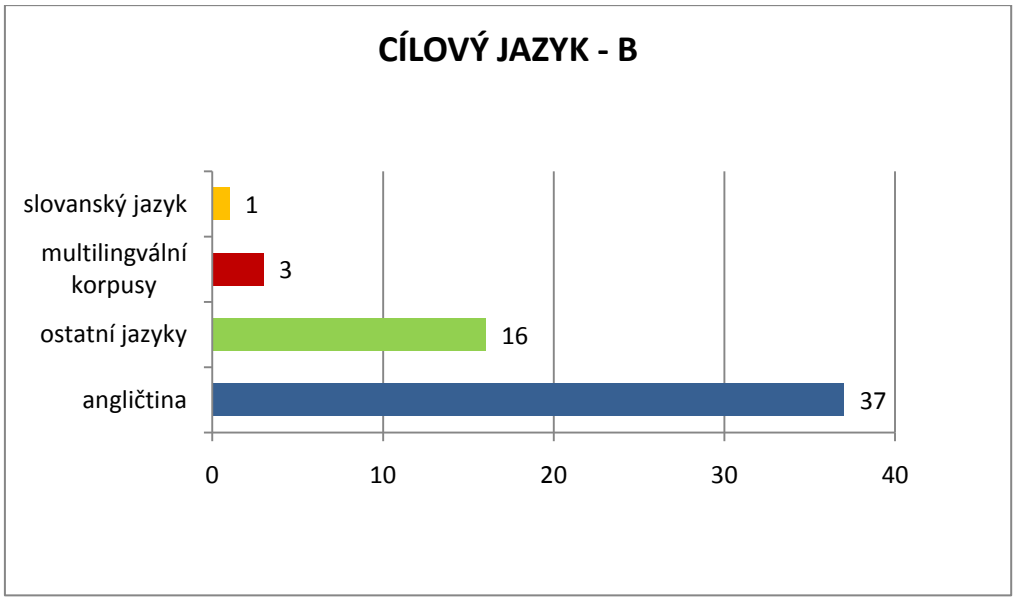
5.1.1 Cílový jazyk

Graf 1: Přehled žákovských korpusů podle zaměření na cílový jazyk (srov. odd. 5.2, tab. 3, sloupec 1)

A. Graf poskytuje přehled žákovských korpusů podle zaměření na cílový jazyk



B. Graf ukazuje počet korpusů zaměřených na angličtinu, korpusů neanglických a multilingválních).



Majoritní množství žákovských korpusů je zaměřeno na angličtinu jako cílový jazyk. Tento fakt vyplývá přirozeně z dlouhodobé tradice výzkumů angličtiny jako cizího jazyka a bohatých analýz nabývání angličtiny jako druhého jazyka. Korpusy vznikající v prostředích, kde existuje nativizovaná varianta anglického jazyka jako oficiální komunikační prostředek, se mezi žákovské neřadí (např. korpusy ICE a Kolhapur Corpus[107]). Nejednotně jsou v dostupných studiích nazírány korpusy zahrnující texty pokročilých nerodilých uživatelů anglického jazyka (od úrovně B2) v případě, kdy angličtina funguje jako společný komunikační prostředek mluvčích s různými mateřskými jazyky, tedy jako lingua franca. Ve své práci zahrnuji korpus VOICE, který mapuje mluvené interakce v angličtině mající roli lingua franca, do žákovských korpusů. Vedle anglických korpusů nerodilých mluvčích se ale významně rozšiřuje i paleta korpusů pro jiné druhé jazyky (viz graf 1A). Pokud je mi známo, existuje prozatím pouze jediný korpus pro slovanský jazyk.[108] Slovinský žákovský korpus PiKUST je velmi subtilním zkušebním projektem Mojci Stritar*ové* (2009) z Ljubljanské univerzity, který obsahuje 35 tisíc slov a adaptuje koncepty norského korpusu ASK.

Zvláštní kategorií jsou multilingvální žákovské korpusy, jež jsou všechny svým ustrojením longitudinální, vzájemně se však významně odlišují. MLC korpus shromažďuje materiály studentů s jedním mateřským jazykem (brazilská portugalština) a zaměřuje se na tři cílové jazyky (angličtina, němčina, španělština). Cílem korpusu je možnost identifikace a rozboru

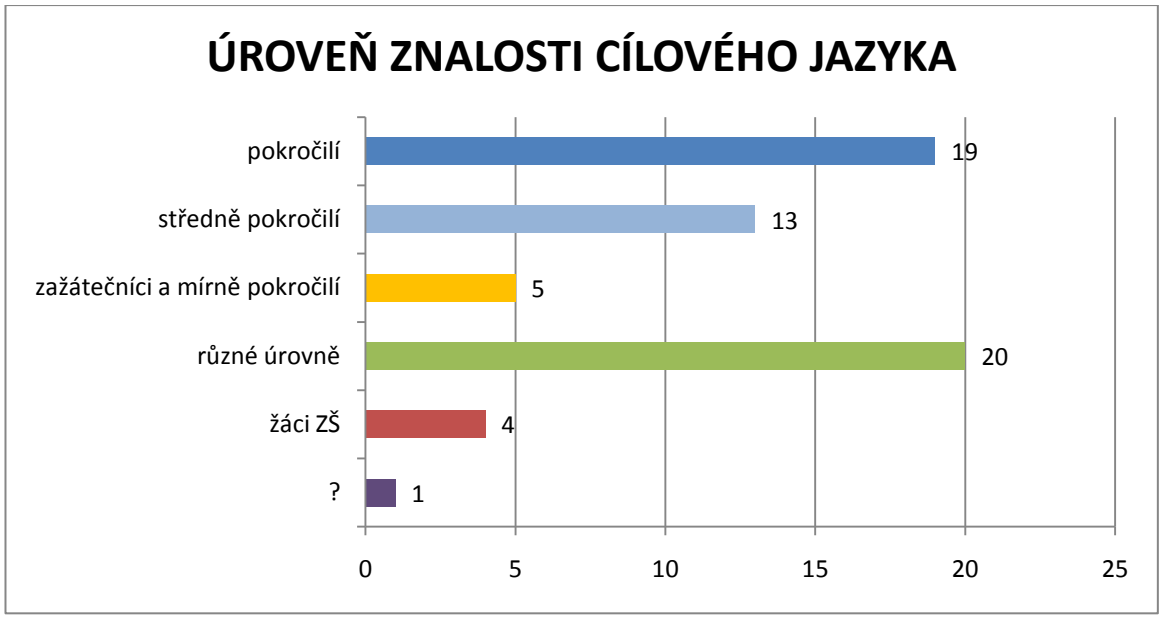
[107] ICE, tj. International Corpus of English (http://ice-corpora.net/ice); Kolhapur Corpus, tj. Kolhapur Corpus of Indian English (http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM)

[108] Ruský tzv. překladový žákovský korpus (Russian Translation Learner Corpus) nepovažujeme za žákovský korpus v námi vymezeném smyslu. V tomto případě jde o paralelní korpus, který se skládá z původních anglických textů a jejich ruských ekvivalentů, jež jsou překládány ruskými univerzitními studenty. Dále viz Sosnina, E. P. Russian Translation Learner Corpus: The First Insights. In *The proceedings of the 6 international scientific conference 'Interactive systems: problems of human-computer interaction'*. Ulyanovsk: ULSTU, 2005.

společných problémů, které mají brazilští studenti při akvizici cizích jazyků, Tagnin*ová* (2003). Multilingvální korpus TELEKORP je příkladem bilingválního kontrastivního žákovského korpusu, který obsahuje projevy rodilých i nerodilých mluvčích anglického a německého jazyka a nabízí se tak jako významný zdroj pro kontrastivní analýzu (CIA), Belz*ová* et al. (2005). Multiligvální korpus ESF Database mapuje spontánní nabývání cizího jazyka u dospělých migrantů s různými mateřskými jazyky (celkem 6) v obdobném komunikačním kontextu v pěti vybraných evropských zemích.[109]

5.1.2 Úroveň znalosti cílového jazyka

Graf 2: Přehled žákovských korpusů podle úrovně znalosti cílového jazyka 1 (srov. odd. 5.2, tab. 3, sloupec 3.



Pro klasifikaci úrovní znalosti cílového jazyka neužívám v grafu ani přehledu žákovských korpusů žádnou standardizovanou škálu (např. dle SERR, ALTE apod.), ale přidržuji se tradičního, i když ne zcela explicitního a jednoznačného členění do kategorií začátečník / mírně a středně pokročilý / pokročilý. Činím tak proto, jak jsem již zmiňovala v oddíle 3.1.1, že většina současných korpusů tento parametr vymezuje právě takto, a protože aplikace jednotného

[109] Viz http://www.mpi.nl/world/tg/lapp/esf/esf.html

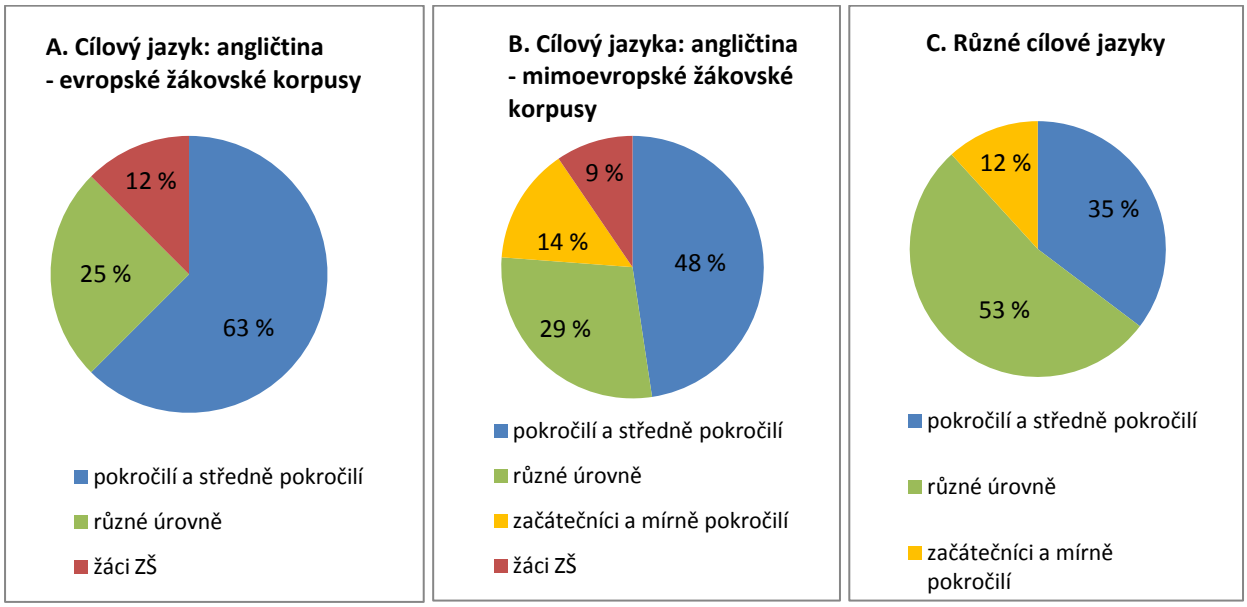
klasifikujícího přístupu na různorodé žákovské korpusy by vyžadovala rozsáhlejší analýzu, která není předmětem této práce. Kategorii žáků nižšího vzdělávacího stupně, obvykle vymezovanou věkem do čtrnácti let, uvádím zvlášť vzhledem k tomu, že znalosti cílového jazyka se v žákovských korpusech pro toto věkové rozmezí stupňovitě nehodnotí.

Evropské žákovské korpusy angličtiny jako cizího jazyka se zřetelně zaměřují na projevy studentů s pokročilou znalostí (viz graf 3A). Domníváme se, že tento fakt je dán několika skutečnostmi. Především jde o snahu kompenzovat dlouhodobé zacílení výzkumů nabývání cizího jazyka na nižší úrovně znalosti, které v praxi znamenalo i nedostatek didaktických materiálů zacílených na pokročilé studenty (Granger*ová,* 2003b) a zároveň jde jednoznačně i o působení vlivného lovaňského korpusu ICLE, který v evropském prostředí (a nejen zde) slouží často jako vzor pro budování nových žákovských korpusů a který se cíleně soustředí na jazyk pokročilých nerodilých mluvčích angličtiny (viz výše odd. 3.1.1). Vedle těchto důvodů mají vliv i technické aspekty budování žákovského korpusu: získávání dat od pokročilých studentů je jednodušší, je možné požadovat větší rozsah jednotlivých vzorků a aplikace existujících softwarových nástrojů je snazší.

Evropské anglické žákovské korpusy se tak zřetelně odlišují od korpusů mimoevropských a korpusů s jiným cílovým jazykem (graf 3A) preferujících sběr materiálu, který pochází od nerodilých mluvčích na různých úrovních znalosti studovaného jazyka. Primárním důvodem sběru takového materiálu je snaha vytvořit kvazilongitudinální korpusy pro potřeby studia interlanguage, především pro vývojové analýzy. Data od studentů na všech úrovních znalosti cílového jazyka shromažďuje také většina korpusů, jejichž výstavba je časově limitována. Podstatným vnějším faktorem u neanglických korpusů je i skutečnost, že některé menší jazyky obvykle nedisponují dostatečným počtem nerodilých mluvčích s pokročilou znalostí jazyka, jejichž projevy by bylo možné využít k vytvoření středního, či středně velkého korpusu.[110]

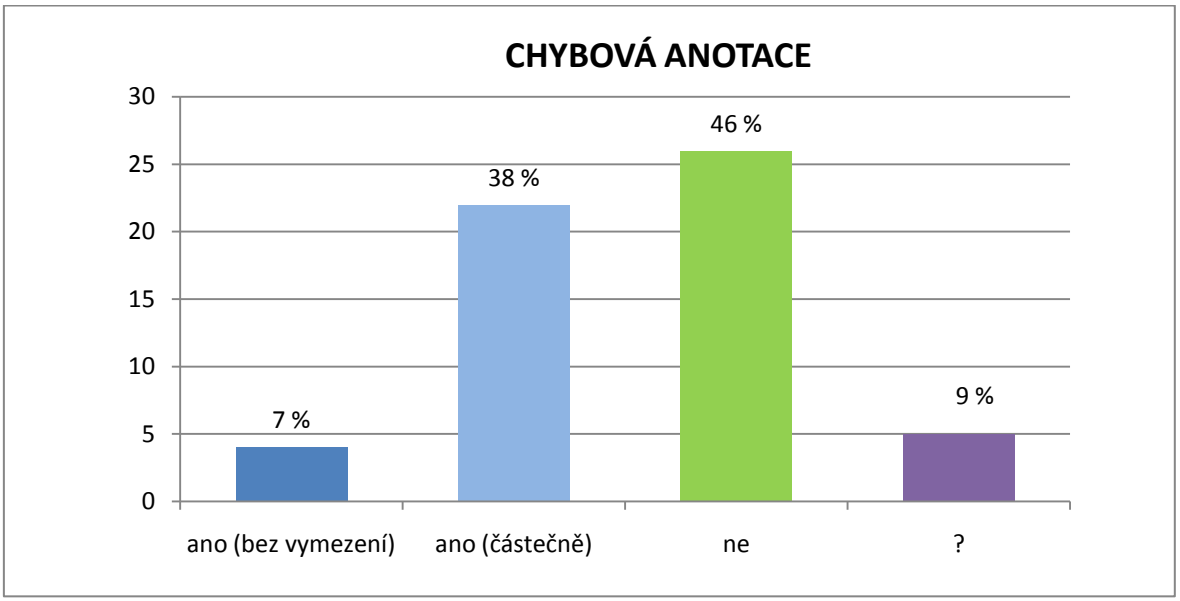
[110] Kvantitativní typologie korpusů viz např. CHIARI, I. *Introduzione alla linguistica computazionale*. Bari: Laterza. 2007, s. 45.

Graf 3: Přehled žákovských korpusů podle úrovně znalosti cílového jazyka 2



5.1.3 Chybová anotace[111]

Graf 4: Chybová anotace aplikovaná v žákovských korpusech 1 (srov. odd. 5.2, tab. 3, sloupec 6)

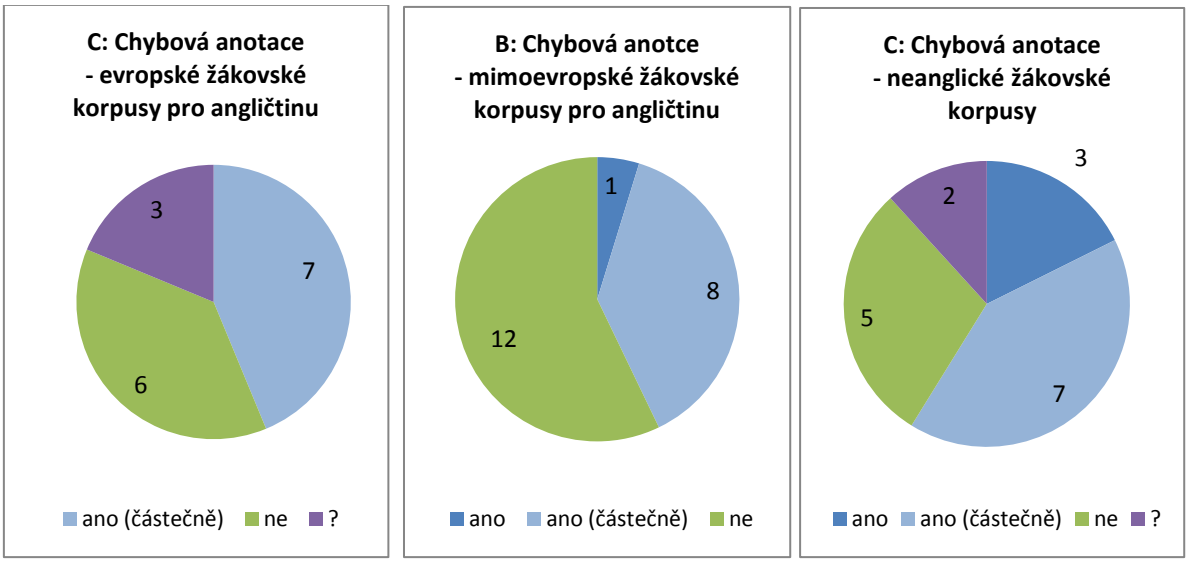


[111] Viz také dále v této práci, kapitola 6.

Pokud hovořím o chybové anotaci žákovského korpusu, mám tím na mysli značkování chyb[112] v žákovských projevech podle navržené chybové taxonomie. Toto značkování je do žákovských korpusů standardně implementováno manuálně. Ačkoli pokusy o automatizaci chybové anotace existují (srov. Izumi et al. 2005, Meurers 2009), nejsou v současnosti v žádném žákovském korpusu aplikovány, resp. případné aplikace nejsou prozatím veřejně dostupné. Chybová taxonomie,[113] která je základem chybového značkování žákovského korpusu, deskriptivně vymezuje třídy chybové klasifikace, odrážejíc lingvistické kategorie a charakter tzv. povrchové modifikace *(target modification, superficial alternations).*

V souvislosti s chybovou anotací existujících žákovských korpusů lze říci, že je vyrovnaný počet těch korpusů, které se o nějakých typ chybového značkování pokoušejí, a těch, které nikoli. Zajímavé je zjištění, že více než padesát procent mimoevropských korpusů zaměřených na angličtinu jako cílový jazyk chybově anotováno není. Téměř polovina evropských žákovských korpusů s angličtinou jako cílovým jazykem a skoro šedesát procent korpusů zacílených na jiný, než anglický jazyk chybovou anotaci aplikují. Ačkoli se však velká část žákovských korpusů (46 %) prezentuje jako chybově anotovaná, je třeba říci, že se ve většině případů (39 %) jedná o omezenou, částečnou chybovou anotaci, tj. úzce zaměřenou na konkrétní jazykovou rovinu (např. lexikální), specifický jazykový jev (výslovnost, ortografii, referenci apod.), nebo limitovanou na určitý rozsah vzorku (např. z třímilionového korpusu ICLE je anotována jedna čtvrtina).

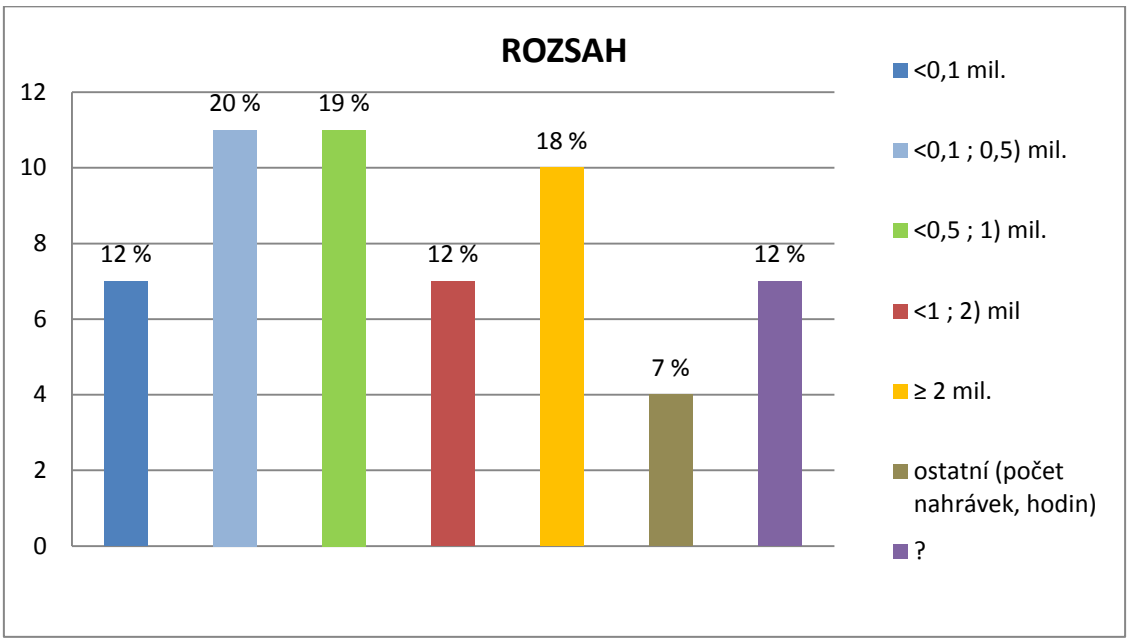
Graf 5: Chybová anotace aplikovaná v žákovských korpusech 2



[112] Žákovskou chybu chápeme jako odchylku od standardu projevu rodilých mluvčích (viz odd. 1.3 v této práci).
[113] Viz dále v této práci odd. 6.3.

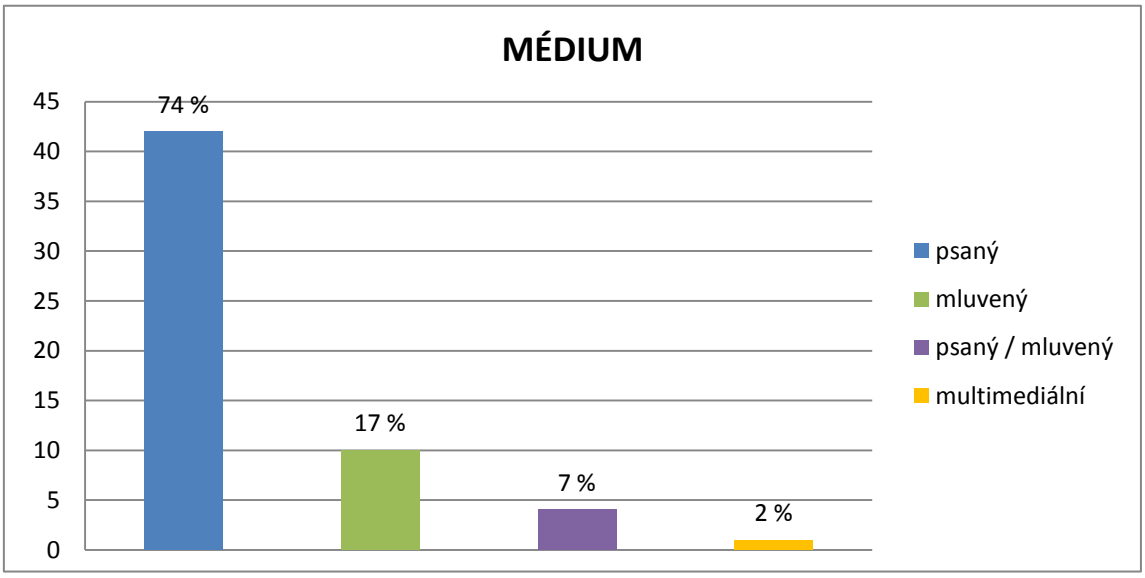
5.1.4 Rozsah a médium

Graf 6: Rozsah žákovských korpusů (srov. odd. 5.2, tab. 3, sloupec 5)



Žákovské korpusy jsou nejednotné co do rozsahu. Žádný existující žákovský korpus nedosahuje kvantitativního standardu národních korpusů. Dvanáct procent žákovských korpusů lze klasifikovat jako velmi malé, tj. v rozsahu do 100 tis. slov. Čtyřicet procent žákovských korpusů patří ke středním (do 1 mil. slov) a sedmnáct korpusů lze zařadit ke středně velkým (do 50 mil. slov), z toho ovšem pouze čtyři korpusy mají víc než čtyři miliony slov (Chungdahm Corpus: 131 mi., CLC: 40 mil., HKUST: 25 mil. a LLC: 10 mil. slov).

Graf 7: Médium v žákovském korpusu (srov. odd. 5.2, tab. 3, sloupec 3)



Jak již jsem zmiňovala v oddílu 2.4.5, na náročný sběr mluvených žákovských projevů se soustředí pouze menší procento (17 %) korpusů žákovského jazyka (celkem deset). Čtyři žákovské korpusy sbírají jak psané, tak mluvené projevy (EVA, LONGDALE, SULEC a ASU), a momentálně existuje jediný multimediální žákovský korpus MAELC.

5.2 Přehled současných žákovských korpusů

Přehled žákovských korpusů členíme do tří částí:

v části (A) uvádíme přehled žákovských korpusů, které se zaměřují na angličtinu jako cílový jazyk. V tomto oddíle dále dělíme korpusy podle toho, kde vznikají, tj. na evropské a neevropské;

v části (B) prezentujeme žákovské korpusy, které mapují jiné žákovské jazyky;

doplňková část (C) představuje multilingvální typy žákovských korpusů.

Žákovské korpusy, které jsme do přehledu zahrnuli, jsou parametrizovány podle:

cílového jazyka, tj. jazyka, na který je korpus zaměřen;

prvního jazyka, tj. výchozího / mateřského jazyka studentů. V případě, že je korpus mezinárodní, resp. když materiál pochází od respondentů s různým jazykovým pozadím, uvádíme v závorce počet výchozích jazyků (např. *různé (12));*

média (typu textu), tj. informace, zda jsou do korpusu zahrnuty psané, či mluvené projevy, příp. oboje;

úrovně znalosti, tj. informace o dosažené znalosti cílového jazyka u respondentů, jejichž projevy korpus zahrnuje. V případě, že korpus není v této oblasti vymezen a neomezuje se na konkrétní cílovou skupinu, uvádíme *všechny* úrovně znalosti;

rozsahu, tj. velikosti korpusu, obvykle vyjádřené v počtech slov. Pokud není v některých případech počet slov uveden, uvádíme sekundární informaci o počtu vět, nahrávek či nahraných hodin apod.;

chybové anotace, tj. informace o způsobu zpracování databanky, resp. má-li korpus chybové značkování. Pokud je tato informace dostupná, uvádíme také, jaký typ chybové anotace daný korpus využívá a je-li anotován celý, či jen částečně;

instituce, tj. pracoviště, které korpus buduje.

V případě, že některá z informací není k dispozici, uvádíme znak *xxx.*

Tabulka 3: Přehled současných žákovských korpusů

A) Cílový jazyk **angličtina**

	2	3	4	5	6	7
EVROPSKÉ KORPUSY	**PRVNÍ JAZYK**	**MEDIUM**	**ÚROVEŇ ZNALOSTI**	**ROZSAH**	**CHYBOVÁ ANOTACE**	**INSTITUCE**
CLC Cambridge Learners Corpus	různé (130)	psaný	různé	35 mil.	ano (částečně)	Cambridge University Press
http://www.cambridge.org/elt/corpus/learner_corpus.htm						
CYLIL Corpus of Young Learner Interlanguage	různé (4)	mluvený	různé	0,5 mil.	ne	Vrije Universiteit Brussel
xxx						
EVA Evaluation of English in Schools	norština	mluvený psaný	žáci ZŠ (14–1 5 let)	0,08 mil.	ne	Norwegian Association for Applied Linguistics
http://www.hf.ntnu.no/anla/EVAdescription.htm						
CHALC Cologne-Hanover Advanced Learner Corpus	němčina	psaný	pokročilí	0,2 mil.	ne	Leibniz Universität Hannover Universität zu Köln
xxx						
ICLE International Corpus of Learner English	různé (18+8)	psaný	pokročilí	3 mil.	ano (částečně, 1/4)	Université catholique de Louvain
http://www.uclouvain.be/en-cecl-icle.html						
ISLE ISLE Corpus of non-native spoken English	němčina italština	mluvený	středně pokročilí	cca 18 hodin nahrávek	ano (výslovnost)	University of Leeds Università di Milano-Bicocca Universität Hamburg
http://nats-www.informatik.uni-hamburg.de/~isle/index.html						
JPU Janus Pannonius University Corpus	maďarština	psaný	pokročilí	0,3 mil.	ne	József Horváth
http://joeandco.blogspot.com						
LeaP LeaP corpus: Learning Prosody in a Foreign Language	němčina	mluvený	středně pokročilí	0,073 mil.	ne	Universität Augsburg
http://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/applied/Research/leap						
LINDSEI Louvain International Database of Spoken English	různé (11)	mluvený	pokročilí	0,8 mil.	ano (částečně)	Université catholique de Louvain

Interlanguage						
http://www.uclouvain.be/en-cecl-lindsei.html						
LLC Longman Learners´ Corpus	různé (160)	psaný	různé	10 mil.	ano (částečně)	Pearson Longman
http://www.pearsonlongman.com/dictionaries/corpus/learners.html						
LONGDALE Longitudinal Database of Learner English	různé	psaný mluvený	pokročilí (studenti VŠ, longitudinální)	xxx	xxx	Université catholique de Louvain
http://www.uclouvain.be/en-cecl-longdale.html						
NOCE NOn-native Corpus of English	španělština	psaný	různé (studenti VŠ, 18–19 let)	0,3 mil.	ano (částečně, 1/4)	Universidad de Granada /Jaén
xxx						
PELCRA Polish Learner English Corpus	polština	psaný	různé	0,5 mil.	ano (částečně)	Uniwersytet Łódzki
http://pelcra.ia.uni.lodz.pl						
SULEC Santiago University Learner of English Corpus	španělština	psaný mluvený	různé	0,5 mil.	ne	Universidade de Santiago de Compostela
http://www.sulec.es						
USE Uppsala Student English Corpus	švédština	psaný	pokročilí	1,2 mil.	ne	Uppsala universitet
http://www.engelska.uu.se/use.html#anchor4						
VESPA Varieties of English for Specific Purposes dAtabase	různé	psaný	pokročilí	xxx	ne	Université catholique de Louvain
http://www.uclouvain.be/en-258647.html						
NEEVROPSKÉ KORPUSY	**PRVNÍ JAZYK**	**MEDIUM**	**ÚROVEŇ ZNALOSTI**	**ROZSAH**	**CHYBOVÁ ANOTACE**	**INSTITUCE**
BALC BUiD Arab Learner Corpus	arabština	psaný	začátečníci středně pokročilí	0,29 mil.	ne	British University in Dubai University of Birmingham
http://ucrel.lancs.ac.uk/publications/cl2009/54_FullPaper.doc						
CALES Corpus Archive of Learner English in Sabah/Sarawak	malajština	psaný	středně pokročilí	0,48 mil	ano (částečně – 9%)	Universiti Teknologi MARA Sarawak
http://www.melta.org.my/modules/tinycontent/Dos/botley_09012008.pdf						
CLEC Chinese Learner English Corpus	čínština	psaný	různé (5 úrovní)	1 mil.	ano	Shanghai Jiaotong Univeristy Guangdong University of Foreign

						Studies
http://langbank.engl.polyu.edu.hk/corpus/clec.html						
EMAS English of Malaysian School Students	malajština	psaný	žáci ZŠ	0,5 mil.	ne	Universiti Putra Malaysia
xxx						
EnglishTLC English of Taiwan Learner Corpus	čínština	psaný	různé	2 mil.	ano (částečně)	National Central University
http://lrn.ncu.edu.tw/Teacher%20Web/David%20Wible/The%20English%20TLC.htm						
HKUST HKUST Corpus of Learner English	čínština	psaný	pokročilí	25 mil.	ano (částečně, 200 tis.)	Hong Kong University of Science and Technology
http://repository.ust.hk/dspace/handle/1783.1/1087						
CHUNGDAHM Chungdahm English Learner Corpus	korejština	psaný	různé (10–16 let)	131 mil.	ano (částečně, 6,6 mil.)	Chungdahm Institute
http://www.lrec-conf.org/proceedings/lrec2010/pdf/821_Paper.pdf						
ICCI International Corpus of Crosslinguistic Interlanguage	různé (8)	psaný	žáci ZŠ	524 textů	ne	Tokyo University of Foreign Studies
http://cblle.tufs.ac.jp/llc/icci						
ICNALE[114] International Corpus Network of Asian Learners of English	japonština čínština	psaný	pokročilí středně pokročilí	1 mil.	ne	Kobe University
http://language.sakura.ne.jp/s/ceeause.html						
ILIAD Indiana Linguistically-Informed Annotated Data	různé	psaný	různé	0,028	ne	Indiana University
xxx						
JEFLL JEFLL Corpus Project	japonština	psaný	začátečníci	0,7 mil.	ano (částečně)	Tokyo University of Foreign Studies
http://jefll.corpuscobo.net						
MAELC Multimedia Adult English Learner Corpus	různé	multimediální	začátečníci mírně pokročilí	2300 hodin nahrávek	ne	Portland State University
http://www.labschool.pdx.edu/research/methods/maelc/intro.html						
MELD Montclair Electronic Language Learners' Database	různé (16)	psaný	pokročilí	0,1 mil.	ano (1/2)	Montclair State University
http://www.chss.montclair.edu/linguistics/MELD						
MICASE	různé	mluvený	pokročilí	1,8 mil.	ne	University of Michigan

[114] Korpus obsahuje subkorpus anglických textů od rodilých mluvčích a japonských textů od rodilých mluvčích.

Michigan Corpus of Academic Spoken English						
http://micase.elicorpora.info/						
NICT JLE *NICT* Japanese Learner English (dříve Standard Speaking Test Corpus)	japonština	mluvený	všechny (9 úrovní)	2 mil.	ano (částečně)	Kobe University
http://www.ijcim.th.org/v12n2/pdf/p119-125-Emi%20IZUMI-emi-paper_nict.pdf						
Quebec learner corpus	různé	psaný	středně pokročilí pokročilí	0,25 mil.	ne	Université du Québec à Montréal
http://www.lextutor.ca/cv/pdf/learner_corpus.pdf						
SILS School of International Liberal Studies Learner Corpus	japonština (převážně)	psaný	různé (studenti VŠ)	---	ne	Waseda University
http://www.f.waseda.jp/vicky/learner/index.html						
SKELC Seoul National University Korean-speaking English Learner Corpus	korejština	psaný	středně pokročilí	1,5 mil.	ne	Seoul National University
http://english.daejin.ac.kr/~elsok/xe/?document_srl=1692						
TLCE Taiwanese Learner Corpus of English	čínština	psaný	pokročilí	2 mil.	ne	National Sun Yat-sen University
http://rocling.iis.sinica.edu.tw/CLCLP/Vol5-2/paper4.pdf						
TLEC Tswana Learner English Corpus	setswana	psaný	středně pokročilí pokročilí	0,21 mil.	ano (ortografie)	North-West University
http://ctext.nwu.ac.za/ProductsCorporaTLEC.html						
VOICE Vienna-Oxford International Corpus of English	různé (50)	mluvený	pokročilí	1 mil.	ne	Universität Wien Oxford University Press
http://www.univie.ac.at/voice						

B) Různé cílové jazyky

	1	2	3	4	5	6	7
	CÍLOVÝ JAZYK	**PRVNÍ JAZYK**	**MEDIUM**	**ÚROVEŇ ZNALOSTI**	**ROZSAH**	**CHYBOVÁ ANOTACE**	**KONTAKT**
ALeSKo An annotated learner corpus	němčina	čínština	psaný	různé	0,05 mil.	ano (syntaktická, diskursivní)	Universität Konstanz Vilniaus pedagoginis universiteta
http://ling.uni-konstanz.de/pages/home/zinsmeister/alesko.html							
ARIDA Arabic Interlanguage Database	arabština	angličtina	mluvený	pokročilí středně pokročilí	0,008 mil.	ano (částečně)	Montclair State University
http://chss.montclair.edu/~feldmana/publications/flairs-2008.pdf							
ASK Norsk andrespråkskorpus	norština		psaný	pokročilí	2000 textů	ano	University of Bergen
http://ask.uib.no							
ASU ASU Corpus	švédština		mluvený psaný	různé (longitudinální)	0,5 mil.	ne	Stockholm University
http://www.ling.su.se/staff/ham/projects.html							
CEDEL2 Corpus Escrito del Español como L2	španělština		psaný	různé	0,75 mil.	ano, částečně (syntaktická, diskursivní)	Universidad Autónoma de Madrid Universidad de Granada
http://www.uam.es/proyectosinv/woslac/cedel2.htm							
CEFLE Lund CEFLE Corpus	francouzština	švédština	psaný	různé	0,1 mil.	xxx	Lund University
http://projekt.ht.lu.se/cefle/information							
CIC (LCIC) Chinese Interlanguage Corpus	čínština	různé (96)	psaný	středně pokročilí	3,5 mil. znaků	ano (lexikální)	Beijing Language and Culture University
xxx							
EAGLE Error-Annotated Corpus of Beginning Learner German	němčina	angličtina	psaný	začátečníci	xxx	ano (částečně)	The Ohio State University
http://www.ling.ohio-state.edu/~adriane/boyd-lrec-2010.pdf							
FALKO Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache	němčina	různé (5)	psaný	pokročilí	0,264 mil.	ano	Humboldt-Universität zu Berlin

http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en							
FLLOC French Learner Language Oral Corpora	francouzština	angličtina	mluvený	různé	2 mil.	ne	Newcastle University University of Southampton
http://www.flloc.soton.ac.uk							
FRIDA French Interlanguage Database	francouzština		psaný	středně pokročilí	0,2 mil.	ano (2/3)	Université catholique de Louvain
http://www.uclouvain.be/en-cecl-frida.html							
ICLFI International Corpus of Learner Finnish	finština	různé (22)	psaný	různé (3 úrovně)	0,6 mil. znaků	ne	University of Oulu
http://www.oulu.fi/hutk/sutvi/oppijankieli/en/index.html							
InterFra Interlangue française	francouzština	švédština	psaný	různé (část longitudinální)	xxx	ne	University of Oulu
http://www.fraita.su.se/interfra							
Korean Learner Corpora	korejština		psaný	začátečníci středně pokročilí	0,01 mil.	ano (ortografie)	Yonsei University
http://www.fas.harvard.edu/~herpro/files/Harvard_2008.ppt							
PiKUST Poskusni korpus usvajanja slovenščine kot tujega jezika	slovinština	různé (18)	psaný	pokročilí	0,035 mil.	ano	Univerza v Ljubljani
http://kuscholarworks.ku.edu/dspace/bitstream/1808/5274/1/8Stritar.pdf							
VALICO Online Corpus of the Learning Varieties of the Italian Language	italština	různé	psaný	různé	0,56 mil.	ne	Università degli Studi di Torino
http://www.bmanuel.org/projects/br-HOME.html							
TUFS TUFS Learners' Corpus: Japanese	japonština	různé	psaný	různé	0,6 mil. znaků	xxx *(plánováno)*	Tokyo University of Foreign Studies
http://cblle.tufs.ac.jp/llc/ja							

C) **Multilingvální** žákovské korpusy

	1	2	3	4	5	6	7
	CÍLOVÝ JAZYK	**PRVNÍ JAZYK**	**MEDIUM**	**ÚROVEŇ ZNALOSTI**	**ROZSAH**	**CHYBOVÁ ANOTACE**	**KONTAKT**
ESF database European Science Foundation Second Language Database	holandština angličtina francouzština němčina švédština	punjabi italština turečtina arabština španělština finština	mluvený	různé (dospělí migranti)	xxx	ne	Max Planck Institute for Psychlinguistics
http://www.mpi.nl/world/tg/lapp/esf/esf.html							
MLC Multilingual Learner Corpus	angličtina němčina španělština	brazilská portugalština	psaný	xxx	xxx	ne	University of São Paulo
http://www.fflch.usp.br/dlm/comet/artigos/A%20multilingual%20learner%20corpus%20in%20Brazil.pdf							
TELEKORP Telecollaborative Learner Corpus of English and German	angličtina němčina	angličtina němčina	psaný	středně pokročilí (longitudinální)	1,5 mil.	ne	Pennsylvania State University Pädagogische Hochschule Heidelberg
http://nflrc.hawaii.edu/networks/nw44/belz.htm							

5.3 Využití žákovských korpusů[115]

Většina dosavadních výzkumů žákovského jazyka se opírala pouze o množstevně omezená experimentální data, která jsou jen nespolehlivým zdrojem pro analýzu autentické produkce studenta cizího jazyka. Žákovské korpusy však umožňují zkoumání jazyka nerodilých mluvčích v takovém rozsahu, v jakém to ještě před méně než dvaceti lety nebylo myslitelné. Badatelé mají při sledování produkce nerodilých mluvčích přístup nejen k chybám, jak tomu bylo v období předkorpusovém, ale k celému žákovskému mezijazyku (Granger*ová* 1998: 6). Data jsou digitalizována a výzkumníci mohou při analýzách využívat množství softwarových nástrojů, primárně určených pro národní korpusy a adaptovaných pro potřeby výzkumů žákovského jazyka, příp. i aplikací vyvinutých speciálně pro korpusy nerodilých mluvčích. Elektronický žákovský korpus lze v ideálním případě snadno distribuovat a díky tomu mohou být výsledky analýz žákovského jazyka ověřovány a srovnávány v širším odborném plénu. Žákovské korpusy poskytují reálný a obsáhlý materiál, ke kterému je možné přistupovat s různorodými hypotézami, zaměřujícími se na jednotlivé jazykové jevy, ale i na otázky pragmatické, příp. diskursivní. V případě, že je korpus sestavován systematicky, lze postulované hypotézy ověřovat v návaznosti na jednotlivá kritéria, resp. v souvislosti se sociologickým a didaktickým značkováním (úroveň znalosti, pohlaví, věk, první jazyk studenta, délka studia apod.). Žákovský korpus tak umožňuje srovnávání mezi jednotlivými subkorpusy, ať již jsou vymezeny dle úrovně znalosti, jazykového pozadí studentů, či jinak. Žákovský korpus může sloužit jako nástroj pro poznávání a popis mezijazyka, příp. jej lze využít ke srovnávání jazyka nerodilých a rodilých mluvčích.

Lze tedy shrnout, že využití žákovského korpusu směřuje primárně do tří oblastí:

(1) odborníkům v oblasti SLA a FLT můžou žákovské korpusy sloužit jako základ výzkumu problematiky nabývání cizího jazyka a cizojazyčného vyučování. Zároveň se ale domníváme, že by korpus jazyka nerodilých mluvčích mohl přispívat i k poznávání cílového jazyka, obdobně jako je tomu např. u dat pocházejících od uživatelů jazyka s poruchami řeči.

(2) pro učitele je žákovský korpus zdrojem identifikace chyb, zdrojem pro evaluaci výukových metod a materiálů a prostředkem pro přímou aplikaci ve výuce (např. čerpání příkladů apod.). Předpokládá se také, že na základě analýzy dat z žákovského korpusu bude možné vytipovat jazykové jevy, které by měly být v explanaci zvláště zdůrazňovány, a v souvislosti s výzkumy vývoje jazykové akvizice bude možné do jisté míry determinovat pořadí, v jakém by měly být

[115] Oddíl 5.3 naší práce byl v upravené podobě zahrnut do příspěvku ‘Žákovský korpus. Budoucnost pro poznávání akvizice cizího jazyka’ přednesného na mezinárodní konferenci Minulost, přítomnost a budoucnost v jazyce a v literatuře (Pedagogická fakulta UJEP v Ústí nad Labem, 1. 9. – 3. 9. 2010), viz Štindlová (2011).

dané jevy vyučovány. Významnou roli může žákovský korpus hrát v přípravě budoucích pedagogů, kteří mají možnost setkat se s reálným žákovským jazykem mimo samotný výukový proces.
(3) v neposlední řadě může žákovský korpus sloužit (většinou zprostředkovaně) žákům v přímé výuce či při studijní přípravě.

Žákovský korpus jako nový nástroj pro výzkum nabývání cizího jazyka má pochopitelně i svá omezení. Vzhledem k povaze materiálů zahrnovaných do těchto korpusů je problematické jejich využití pro výzkum receptivních schopností studentů. Na základě žákovského korpusu lze zároveň jen obtížně analyzovat některé didaktické aspekty, např. roli výukové metody, vliv jazykového vstupu a interakce mezi učitelem a žákem. V případě, že se daný jazykový prvek či struktura v korpusových textech nevyskytují, nelze na základě korpusu ověřit, zda student uplatňuje strategii vyhýbání, prvek tedy zná, či nikoli.

V přibližně patnáctileté tradici žákovských korpusů byla doposud hlavní badatelská pozornost věnována především principům samotného vytváření žákovského korpusu, jeho struktuře, chybové anotaci, a také charakteru jeho vztahu ke korpusům národním. V současné době ale významně narůstá i počet konkrétních lingvistických studií jazyka nerodilých mluvčích založených právě na analýzách žákovských korpusů. Nejvíce textů se zabývá otázkami lexikálními, nejmenší zájem vyvolávají problémy ortografické. Protože jen minimum korpusových dat je longitudinálních či mluvených, jsou žákovské korpusy prozatím jen zřídka využívány k analýzám principů akvizice cizího jazyka a výzkumům fonologickým. Příklady současných studií vycházejících ze žákovského korpusu viz Příloha 3.

6 CHYBOVÁ ANOTACE VE SVĚTOVÝCH ŽÁKOVSKÝCH KORPUSECH

V následujících dvou kapitolách budou souhrnně představeny různé přístupy k anotaci, které se uplatňují ve světových žákovských korpusech. Při prezentaci vycházím z podrobného rozboru vybraných korpusů nerodilých mluvčích, který byl představen v předcházející kapitole, a z několika dílčích studií, které se dané problematiky dotýkají. Jde především o příspěvek Díaz-Negrillo*vé* a Fernández-Domíngueze (2006), kteří podrobně rozebírají otázky počítačem podporované chybové analýzy (CEA) a dávají do kontrastu čtyři rozdílné značkovací systémy, tj. anotační nástroje korpusů CLC, NICT JLE a ICLE (tzv. Louvain systém), a také značkovací

nástroj FreeText vyvinutý v souvislosti s CALL programy také na lovaňské univerzitě. Na základě těchto anotačních nástrojů porovnávají odlišné chybové taxonomie. Díaz-Negrillo*vá* a Fernández-Domínguez vycházejí ve své studii z textu Nicholsové (2003), která uvádí podrobnou analýzu značkovacího schématu žákovského korpusu CLC, a také z textu Izumiho et al. (2005), komplexně představujícího anotační schéma žákovského korpusu NICT JLE. Značkovací systém FreeText rozebírá ve své stati Granger*ová* (2003a), která prezentuje i značkovací schéma korpusu ICLE (Granger, 2003b). Další odborné texty reflektující problematiku chybové anotace uvádím vždy v příslušném oddílu této práce.

Zmapování problematiky anotace v žákovském korpusu nám poslouží jako metodologický rámec pro evaluaci anotačního konceptu navrženého pro žákovský korpus češtiny jako druhého jazyka (CzeSL). V kapitole 6 shrnu, jakým způsobem pracují s chybovou anotací světové korpusy zaměřené na nerodilé mluvčí. Vycházím z analýzy reprezentativního vzorku žákovských korpusů, kterou jsem uvedla v kapitole 5 této práce. Dále se zaměřím na charakteristiku anotačních schémat a chybových taxonomií, které se vyskytují v chybově značkovaných korpusech, představím možnosti chybového značkování a architektury chybové anotace. V následující kapitole 7 se podrobněji zaměřím na vybrané žákovské korpusy.

Geoffrey Leech (1997: 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.[116] 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 jako celek, včetně nekorektních struktur.

Anotaci v žákovských korpusech lze rozdělit do dvou na sobě nezávislých typů. Za prvé se jedná o možnost lingvistického značkování (tím máme na mysli značkování slovních druhů,

[116] Řada odborníků má k různým typům anotací elektronických jazykových korpusů výhrady. Dlouhodobě diskutována je např. otázka adekvátnosti využití surového, resp. anotovaného korpusu pro lingvistické analýzy, která se odráží ve dvou víceméně protichůdných metodologických přístupech, tzv. *data-driven linguistics* a *data-based linguistics.* Srov. např. Sinclair (2004a: 192) „As long as we rely on tags we are forcing the attention (and the resources) on pre-corpus models of language which require only small corpora anyway. Tagged corpora will not meet the requirements of the information society because they are not sensitive enough.“. Příp. i Tognini-Bonelli, E. *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins, 2001.

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

Druhou rovinou značkování žákovských korpusů je tzv. chybová anotace (viz např. Díaz-Negrillo*vá* a Fernández-Domínguez (2006) a také zde, odd. 6.2 a 6.3), která je také jedním z centrálních témat této práce. Chybové anotování korpusu žákovského jazyka znamená manuální přiřazení odpovídající značky (neboli tagu) konkrétní chybě vyskytující se v žákovském projevu. V současné době se některá výzkumná pracoviště soustředí na vývoj automatizace chybové anotace, prozatím však jejich výsledky nebyly evaluovány, srov. Tono*vá* (2000), 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, Milton a Chowdhury (1994).

6.1 Anotace v žákovských korpusech

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 atd., viz zde kapitola 3).

[117] Výjimkou jsou, pokud vím, následující studie:
Van Rooy - Schäfer (2003); DE HAAN, P. Tagging non-native English with the TOSCA-ICLE tagger. In *Corpus Linguistics and Linguistic Theory*. Eds. C. Mair, M. Hundt. Amsterdam: Rodopi, 2000, s. 69–79; DE MÖNNINK, I. Parsing a learner corpus. In *Corpus Linguistics and Linguistic Theory*. Eds. C. Mair, M. Hundt. Amsterdam: Rodopi, 2000, s. 81–90.

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. Přesto nelze tvrdit, že chybová anotace je pro žákovské korpusy standardní výbavou. V souvislosti s analýzou z oddílu 5.1.3 konstatuji, že dvacet šest (tj. 45 %) žákovských korpusů chybově anotováno není, z dvaceti šesti (tj. 45 %) korpusů, které aplikují chybovou anotaci, se čtyři z nich (pouhých 7 %) pokouší o široce pojatou chybovou anotaci s komplexní taxonomií chyb. Dvacet dva žákovských korpusů značkuje chyby v souvislosti se zaměřením korpusu a výzkumnou hypotézou Např. korpus CIC se zaměřuje na značkování lexikálních problémů, korpus ISLE na výslovnostní chyby apod. 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 přibližně 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 (srov. Rastelli, 2009; Meurers, 2005; Díaz-Negrillo*vá* a Fernández-Domínguez, 2006 atd.), 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. Fitzpatrick*ová* a Seegmiller, 2004).

6.2 Anotační modely

6.2.1 Lineární anotační model[118]

Současné anotované žákovské korpusy používají v zásadě dva typy anotačních schémat. Majoritně se v žákovských korpusech stejně jako ve většině značkovaných referenčních korpusů[119] 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.[120] Zároveň jsou

[118] Tzv. *flat token-tag architecture*, příp. *single-layer annotation; inline corpus architecture.*
[119] Referenční korpus chápeme v souladu s Čermákem a Blatnou (1995: 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."
[120] Srov. např. Weinberger*ová,* 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 …). Viz také LÜDELING, A. *Vortrag & Workshop über Lernerkorpora / Korpusbasierte Studien zum Erwerb von komplexen*

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 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. Nedostatky lineárního, jednorovinného zachycování chyb v textech nerodilých mluvčích shrnují např. Lüdeling*ová* et al. (2005), Zeldes et al. (2009) a Fitzpatrick*ová* – Seegmiller (2003).

Nejjednodušší strukturou jednorovinné anotace je tzv. tabulární model (k termínu viz Lüdeling*ová* 2006), kdy je materiál značkován na úrovni tokenů,[121] přesněji řečeno chybová značka je spojena vždy s jednotkou psaného textu určitého rozsahu, příp. s časovými segmenty verbálního, nebo vizuálního signálu (tzv. vertikální formát), srov. s Carletta*ová* et al. 2002: 3. V rámci žákovských korpusů je variací tohoto modelu značkování korpusu C-LEG (Weinberger*ová* 2002). V příkladu 2 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.

(2) *<LxPhCh>Es gibt eine veränderte Gesellschaft und ...*
Lx- lexikální doména, Ph – exponent chyby (fráze), Ch –chybný výběr (specifikace chyby)
Hypotéza*: die Gesellschaft hat sich verändert*
(Weinberger*ová*, 2002: 25; podtrženo BŠ)

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 (3) z žákovského korpusu NICT JLE je

Verben bei Lernern des Deutschen als Fremdsprache am SFB Mehrsprachigkeit, Universität Hamburg, Juni 2006. Příp. také Granger (2003a: 467n.).

[121] 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). Viz např. Pala (1996).

rekonstrukční hypotéza umístěna před chybový výraz, který je zároveň oboustranně vymezen příslušným chybovým tagem. Rekonstrukční hypotéza je v tomto případě explicitní.

(3) *I belong to two baseball <n_num crr=„teams“>team</n_num>*
n- substantivum, num – číslo, crr – korekce
(Izumi, Uchimoto a Isahara, 2005: 75; podtrženo BŠ)

Příklady (4a,b) a (5a,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.

(4) a. Celý den ***myslela o*** *kamarádka.*[122]
b. Celý den myslela ***o kamarádka.***

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 (termín viz Ellis 1994: 5). V následujícím příkladu je možná dvojí hypotéza: buď je chybně užito sloveso *(zapamatovat* místo *vzpomenout),* nebo je chybně užita prepozice *(na).*

(5) 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 (tematicky zaměřená sonda viz Lüdeling, 2008). Zachycení různých rekonstrukčních variant umožňuje víceúrovňová architektura.

[122] České příklady vyexcerpovala autorka z databanky textů nerodilých mluvčí, která je shromažďována pro korpus CzeSL (korpus češtiny jako druhého jazyka).

6.2.2 Víceúrovňová distanční anotace

Odlišným anotačním schématem je tzv. víceúrovňová distanční anotace,[123] kterou používá ve světovém kontextu žákovských korpusů pouze německý korpus FALKO. Tento anotační systém navazuje na modely, které byly v posledním desetiletí vyvinuty pro mluvené a multimodální korpusy. Srov. např. Bird a Liberman (1999), kteří konstruovali několikaúrovňový anotační model pro část korpusu BU Radio News.[124]

Multidimezionální distanční anotace aplikovaná na žákovský korpus se metodologicky opírá o předpoklad Lüdeling*ové* 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. (Anotační formát pro žákovský korpus FALKO byl navržen na základě EXMARaLDA Partitur-Editor.[125]) 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 4).

Tabulka 4: Příklad anotace překrývajících se řetězců[126]

	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						

[123] Tj. *multi-layer stand-off annotation,* příp. *multi-level model* nebo *flexible annotation model,* v češtině pak někdy jako *několikaúrovňová anotace.*

[124] Příp. i BIRD, S. An Integrated Framework for Treebanks and Multilayer Annotations. *Proceedings of the Third International Conference on Language Resources and Evaluation,* Paris: European Language Resources Association, 2002.
Multidimenzionální anotace mluvených i psaných korpusů je v současnosti plně etablovaným nástrojem lingvistické analýzy. Srov. např. formáty NITE (Carletta*ová* et al., 2003), EXMARaLDA (Schmidt, 2004), PAULA (Dipper, 2005), SGF (Stührenberg – Goecke, 2008) atd.

[125] K systému EXMARaLDA viz např. http://www.exmaralda.org/partitureditor.html a Schmidt, T. EXMARaLDA - ein System zur computergestützten Diskurstranskription. In *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*. Eds. A. Mehler, H. Lobin. Wiesbaden: Verlag für Sozialwissenschaften, 2004, s. 20 –218.

[126] Obrázek převzat z Lüdeling*ové* et al. (2005).

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 mám 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.

6.3 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 zmínili v úvodu této kapitoly, prozatím odbornou veřejností neevaluováno, a na počítačem podporovanou chybovou analýzu žákovských projevů. Srov. dále i Tono*vá,* 2003: 804.

Chybovou analýzu[127] lze standardně charakterizovat v pěti krocích: sběr dat, resp. výběr vzorku pro analýzu, identifikace chyby, popis chyby, její vysvětlení a zhodnocení (srov. Ellis 1994: 48).[128] 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

[127] Podrobněji o problematice chybové analýzy viz v této práci odd. 1.6.

[128] Corder (1974) uvádí kromě těchto čtyř stupňů chybové analýzy ještě krok pátý: evaluaci chyb. Ta však bývá v pozdějších studiích oddělována jako samostatný problém s vlastními metodami výzkumu, resp. jako varianta chybové analýzy ze 70. – 80. let, jež se zabývá otázkou míry závažnosti chyb v projevech nerodilých mluvčích a jejich působení na komunikačního partnera. Srov. např. LUDWIG, J. Native-speaker judgments of second-language learners' efforts at communication: A review. *Modern Language Journal,* 1982, r. 66, s. 274–283.
HUGHES, A., LASCARATOU, C. Competing criteria for error gravity. *ELT Journal,* 1982, vol. 36, no. 3, s. 175–182. Alternativní klasifikace jednotlivých korků chybové analýzy: kolekce, identifikace, klasifikace, kvantifikace, analýza zdroje chyby, náprava. Viz Gass*ová* a Selinker (2008).

přístup k relevantnímu materiálu pro následný lingvistický výzkum, nikoli tento výzkum provádět.[129] Viz i Stritar*ová* (2009: 139).

Existují dva základní přístupy k zachycování chyb v žákovských projevech. Za prvé jde o rekonstrukci, kdy je chyba v textu detekována a nahrazena korektní formou (viz Seegmiller a Fitzpatrick*ová,* 2003), za druhé se jedná 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).[130] K cílové hypotéze a mezianotátorské shodě viz dále v této práci, kapitola 9.

Fitzpatricková a Seegmiller (2003) uvádějí, že výhodou rekonstrukčního přístupu je primárně absence klasifikačního schématu: 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. Domnívám se však, že samotná rekonstrukce bez kategorizace chyb může být 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.

6.3.1 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 tagy, užívanými v anotačním procesu k zařazení konkrétních žákovských chyb, které

[129] „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: 467)

[130] Standardně při anotování žákovských projevů značkují jeden text nezávisle vždy minimálně dva anotátoři.

se v žákovském korpusu vyskytují, do příslušných kategorií dle zvolené taxonomie. Přehled vybraných chybových taxonomií viz Díaz-Negrillo*vá* a Fernández-Domínguez (2006: 92).

Na základě provedeného výzkumu lze stanovit, že v chybově anotovaných žákovských projevech se v současnosti uplatňují dva základní typy konstrukčních přístupů k zachycení a popisu chyb. Pro další popis je definujeme jako taxonomii parciální a taxonomii komplexní.[131] Chybovou anotaci ve smyslu prosté emendace, tzn. čisté opravy textu, bez chybové klasifikace, tedy bez chybové taxonomie a bez tagů, používá pouze americký korpus MELD.

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).[132] 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). Vyžaduje podrobnou rozpracovanost anotačního manuálu a klade značné nároky na kvality anotátora. Je však významnou podporou pro analýzu nabývání i výuky cílového jazyka.[133]

6.3.2 Struktura chybových taxonomií

James (1998: 104–113), a shodně i Granger*ová* (2003a: 467), Tono*vá* (2003: 804), Díaz-Negrillo*vá* a Fernandéz-Domínguez (2006: 92) uvádějí, že každá deskriptivní chybová taxonomie by měla reflektovat dvě svébytné oblasti: lingvistickou kategorizaci a formální popis chyby (tj. informaci o tom, že jev chybí, přebývá, je chybně umístěn, je chybně použit).[134] Obě hlediska pro klasifikaci chyb jsou v existujících chybových taxonomiích žákovských korpusů strukturována buď bidimenzionálně, resp. propojeně, obvykle v podobě slovnědruhové značky a značky pro typ povrchové realizace (příklad (6) z korpusu CLC), nebo separátně, tj. tagy pro značkování chybějících/nadbytečných jevů, chyb ve slovosledu atd. nejsou spojovány

[131] Oba termíny zavádí do kontextu problematiky chybově značkovaných žákovských korpusů autorka této práce.
[132] Srov. Nicholls*ová*, 2003.
[133] Tono*vá* (2002: 801): „A generic error tagset, however, stills seems to be a very useful goal to work towards…"
[134] Podrobnější informace o chybové analýze a přístupech ke klasifikaci chyb viz v této práci odd. 1.6.3.3.

s lingvistickou kategorizací (příklad (7) z korpusu ICLE).[135] Některé žákovské korpusy povrchový charakter chyby nereflektují (např. NICT JLE).

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

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

Značkování chyb je v existujících chybových taxonomiích různorodě strukturováno. 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, viz Díaz-Negrillo*vá* a Fernandéz-Domínguez, 2006) a FRIDA (tzv. systém FreeText, viz Granger*ová*, 2003a). 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.[136]

7 ANALÝZA VYBRANÝCH ŽÁKOVSKÝCH KORPUSŮ

V následujícím oddíle podrobněji představím šest vybraných žákovských korpusů a zaměřím se především na způsob zachycování chyb v textech nerodilých mluvčích, který dané korpusy používají.

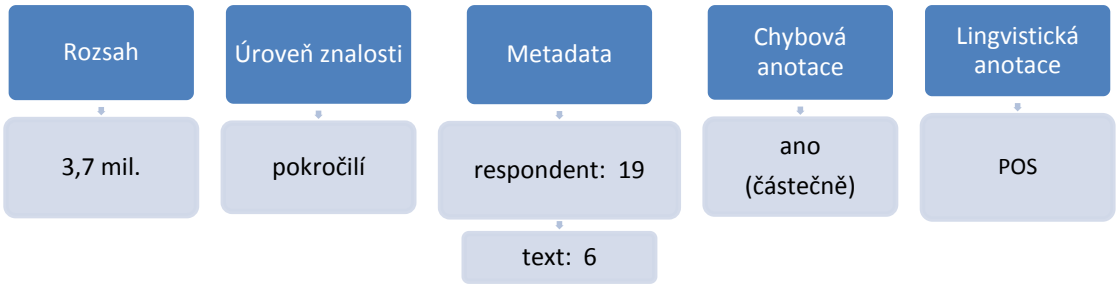
[135] Příklady převzaty z Nicholls*ová,* 2003: 573; Dagneaux*ová* et al., 1998: 166.
[136] Srov. Izumi et al. (2005).

Lovaňský korpus ICLE je reprezentantem dlouhodobě budovaného žákovského korpusu a lze říci, že je vzorovým korpusem pro většinu evropských, ale i světových korpusů projevů jinojazyčných mluvčích. Projekt, který v současnosti zahrnuje dvacet jedna subkorpusů, členěných podle prvního jazyka respondentů, od nichž byla data získávána, řídí Sylviane Granger*ová* 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 Fitzpatrick*ové* 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 Cambridge Learner Corpus (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 Lüdeling*ové* 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 začleňuji i tzv. pilotní korpus PiKUST, který je prozatím jediným doloženým žákovským korpusem zaměřeným na slovanský jazyk. Vznikal jako součást disertační práce Mojci Stritar*ové* na Filozofické fakultě Univerzity v Ljubljani a v současné době je nedostupný. Na jeho základě je v počáteční fázi budování rozsáhlejší korpus slovinštiny jako cizího jazyka.

Vybrané korpusy analyzuji vzhledem k parametrům, podle kterých byly budovány, a podle typu aplikované chybové anotace. Navržené chybové taxonomie jsou prezentovány pro lepší ilustraci také na příkladech, které pocházejí z databanky projevů nerodilých mluvčích, shromažďované pro žákovský korpus CzeSL.[137]

[137] Pro autentické příklady z jednotlivých korpusů viz příslušnou sekundární literaturu.

7.1 ICLE – International Corpus of Learner English[138]



7.1.1 Korpus

Žákovský korpus ICLE je v současnosti pravděpodobně nejvlivnějším projektem v této oblasti výzkumu. Jeden z prvních nekomerčních žákovských korpusů začal vznikat na počátku devadesátých let minulého století na bázi mezinárodní univerzitní spolupráce. Jeho současný rozsah je tři miliony slov a na jeho výstavbě se podílí osmnáct, resp. dvacet šest[139] 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 (podle Grangerové, 1998). 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;[140] za druhé možnost komparace s jazykem rodilých mluvčích.[141]

7.1.2 Metadata

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

[138] Kontakt: Sylviane Granger, http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm
[139] Osmnáct subkorpusů je v současné době hotových, osm subkorpusů je ve výstavbě.
[140] A při chybové analýze rozhodnout, zda jde o chyby univerzální nebo jazykově specifické.
[141] Tj. s korpusem rodilých mluvčí angličtiny LOCNESS (Louvain Corpus of Native English Essays).

apod.), dalších šest parametrů se týká samotného textu (časová limitovanost, referenční pomůcky apod.). Dále viz Granger*ová,* 1998: 9–10.

7.1.3 Chybová anotace

Data žákovského korpusu ICLE jsou automaticky slovnědruhově anotována pomocí nástroje TOSCA-ICLE, který je aplikován na bezchybné úseky textů. Chybové značkování je založeno na porovnání originálního textu a jeho korigované varianty, opravené rodilým mluvčím. Systém klasifikace chyb korpusu ICLE[142] 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[143], 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“.[144] 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 (8 a,b,c). Viz i Díaz-Negrillo*vá* et al., 2006: 94. Pro anotování chyb v korpusu ICLE se využívá pro manuální anotaci specifický editační nástroj vyvinutý v Lovani, tzv. Error Editor.

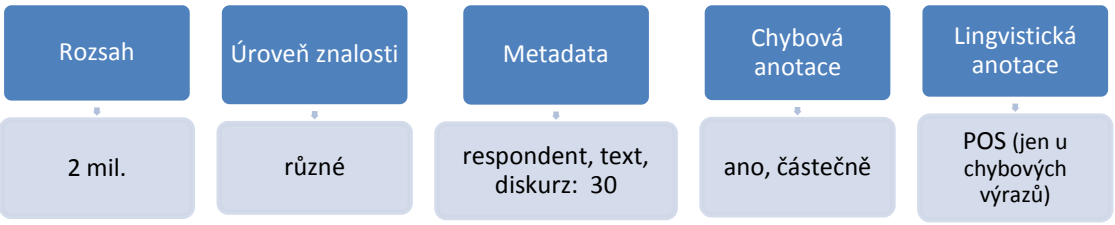
(8) a. *(WR) Jsem 0 studuju na univerzitě.*
W – slovo, R – nadbytečné

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

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

[142] Někdy nazývaná jako‘ lovaňský systém’ (Díaz-Negrillo*vá* et al., 2006 : 93).
[143] Termín *registr* používáme v souladu se stylisticko-sociolingvistickou teorií M. A. K. Hallidaye. V českém odborném diskurzu se termín objevuje např. u Hoffmannové (1997:163).
[144] Tj. missuse / replacement.

7.2 NICT JLE – National Institute of Information and Communications Technology Japanese Learner English Corpus[145]



7.2.1 Korpus

Mezi žákovskými korpusy se jen několik málo z nich zaměřuje na mluvený jazyk. Mezi takové patří japonský korpus NICT JLE. Srov. Izumi et al. (2004b, 2005). V současné době obsahuje dva miliony slov, resp. 1281 patnáctiminutových interview nahraných při mluvních zkouškách z angličtiny.[146] Jedná se o tzv. kvazilongitudinální 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.

7.2.2 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.

7.2.3 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

[145] Kontakt: Emi Izumi, Kiyotaka Uchimoto, Hitoshi Isahara; www.nict.go.jp
[146] An English oral proficiency interview test ACTFL-ALC Standard Speaking Test (SST).

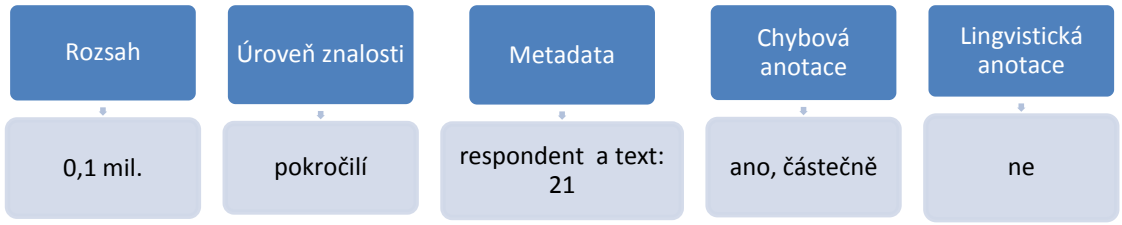
lexému', kategorie 'relativa' zahrnuje subkategorie 'chyba v pádu, lexému' atp. Srov. příklad (9) a (10).

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

(10) *<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.[147] Ačkoli 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.

7.3 MELD – Montclair Electronic Language Database[148]



7.3.1 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

[147] Evaluace výzkumu Dulay*ové* a Burt*ové* z roku 1974.
[148] Kontakt: Eileen Fitzpatrick a Milton Steve Seegmiller, http://chss.montclair.edu/linguistics/MELD

obsahuje přibližně 100 000 slov, polovina z nich je chybově emendována.[149] Databáze MELD je specifická v tom, že se zaměřuje pouze na shromažďování dat reprezentujících angličtinu jako druhý, nikoli cizí jazyk[150] a jejím cílem je přispět k revizi výsledků analýz v oblasti nabývání druhého jazyka. Autoři přijímají hledisko Gerharta Nickela (1989: 298), že nedostatečné oddělování cizího a druhého jazyka je částečně příčinou protichůdných výsledků výzkumů nabývání druhého jazyka.

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

7.3.3 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. Viz Fitzpatrick*ová* a Seegmiller (2004: 4). 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 (11 a,b). Stylistické problémy se opravují jako [chyba/rekonstrukce].

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

[149] 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*ová* a Seegmiller, 2004: 3)
Termín emendace používám ve shodě s praxí v korpusu CzeSL ve smyslu prosté opravy, resp. rekonstrukce textu dle standardů cílového jazyka.
[150] K rozdílům viz terminologický slovníček v příloze této práce.

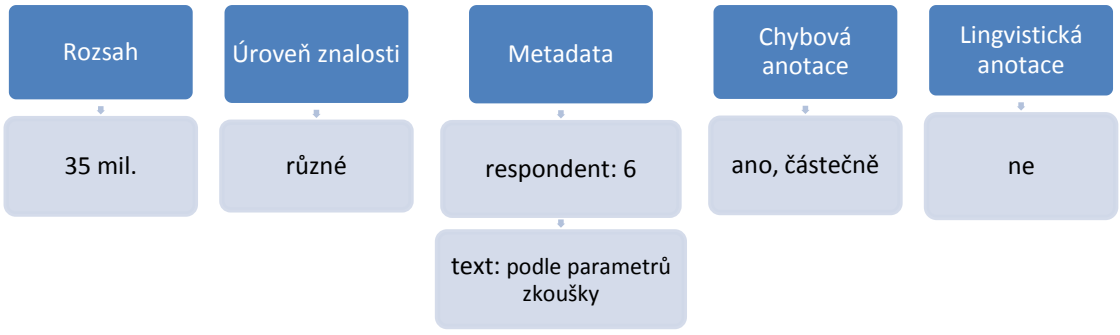
Problémem takové rozvolněné anotace chyb je častá mezianotátorská neshoda v tom, co považovat v daném kontextu za chybu a do jaké domény řešený problém spadá. Srov. např. příklad (12) a (13).

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

(13) a. ? *Mám {hodne/hodné} 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.
Žákovský korpus MELD není lingvisticky značkován, proběhly pouze dílčí pokusy o uplatnění slovnědruhové anotace pomocí Brill taggeru (Brill, 2005).

7.4 CLC – Cambridge Learner Corpus[151]



[151] Kontakt: http://cambridge.org/cz/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus

7.4.1 Korpus

Žákovský korpus CLC, který je součástí Cambridge International Corpus (CIC),[152] připravuje v Cambridge University Press ve spolupráci s Cambridge ESOL.[153] 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.[154] 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.

7.4.2 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.

7.4.3 Chybová anotace

Chybové značkování žákovského korpusu CLC si neklade za úkol vytvoření systematické taxonomie žákovských chyb, jeho cílem je okamžitá praktická využitelnost. Veškerá data CLC jsou manuálně značkována pouze dvěma anotátory, aby byla zaručena co největší konzistentnost tagování. Viz Nicholls*ová*, 2003: 572. Chybová taxonomie žákovského korpusu CLC vychází z kombinace povrchové a lingvistické deskripce žákovských chyb, ačkoli tento dvojdimenzionální přístup není aplikován důsledně (Nicholls*ová*, 2003: 574, 576). 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.[155] Tato základní kategorizace je doplněna o

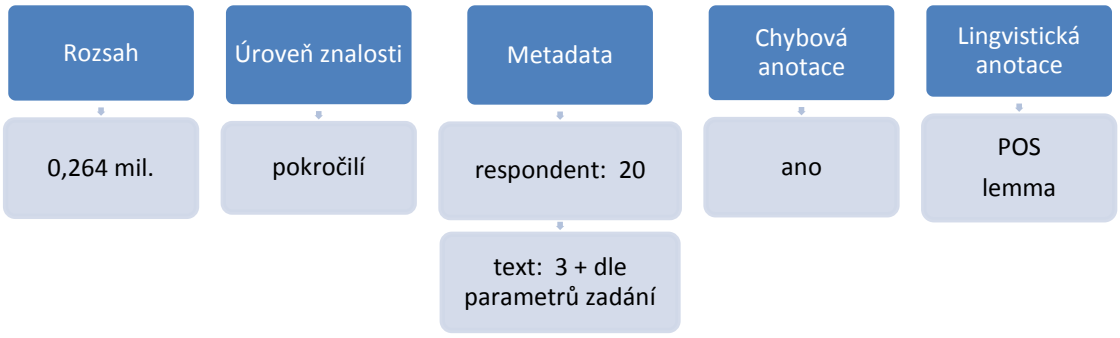
[152] CIC v současnosti zahrnuje více než miliardu slov a člení se do dílčích subkorpusů, např. Cambridge and Nottingham Corpus of Discourse in English (CANCODE), Cambridge Corpus of Business English, Cambridge Corpus of Financial English, Cambridge Corpus of Academic English atd.
[153] ESOL, tj. English for Speakers of Other Languages. UCLES, tj. University of Cambridge Local Examination Syndicate.
[154] Zkoušky KET, PET, FCE, CAE, CELS apod.
[155] Pronoun (A), conjunction (C), determiner (D), adjective (J), noun (N), quantifier (Q), preposition (T), verb (include modals (V), adverb (Y).

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 (14) a (15). 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.

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

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

7.5 FALKO - Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache[156]



7.5.1 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); (2) obdobné texty od rodilých mluvčích, které vznikaly za víceméně stejně řízených podmínek jako texty nerodilých mluvčích;

[156] Kontakt: Anke Lüdeling, http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

(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.

7.5.2 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.).

7.5.3 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. Srov. Lüdeling*ová* et al. (2008). 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 cílové 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,[157] a podrovinu explanační (viz tab. 5).

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

Tabulka 5: Prezentace cílové hypotézy v žákovském korpusu FALKO
(převzato z Lüdeling*ové* et al., 2005: 7)

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“,[158] 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*ová,* 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 (16 b,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). Dále viz Reznicek et al. (2010).

(16) a. *Praha, bude líbit jí se.*

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

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

[158] 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.

b. Převzato z Reznicek et al. (2010a: 35)[159]

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

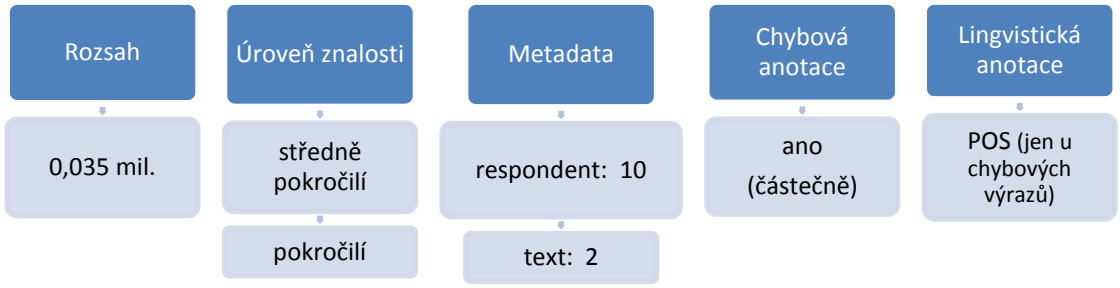
SPLIT – rozděleno

c. Převzato z Reznicek et al. (2010b)

LT	Das	ist	vielleicht	das	Grund	,	,	aus	dem		mir		das	Wort	und	die	Ideologie	,	wenn	man	,	so	sagen	kann	,	,	die	sie	sich		hineinsteckt	,	,	nicht	echt	im	Begriff	habe	.
ZH2a	Das	ist	vielleicht	der	Grund	dafür	,			dass	ich		dieses	Wort	und	die	Ideologie	,	wenn	man		so	sagen	kann	,		die			darin	enthalten	ist	,	nicht	wirklich		begriffen	habe	.
ZH2b	Das	ist	vielleicht	der	Grund		,			weshalb	ich	für	das	Wort	und	die	Ideologie	,	wenn	man		so	sagen	kann	,	was				darin	steckt		,	keinen			Begriff	habe	.
ZH2c	Das	ist	vielleicht	der	Grund		,			warum	ich		das	Wort	und	die	Ideologie	,	wenn	man	es	so	nennen	kann	,		die			darin	steckt		,	nicht	wirklich		begriffen	habe	.
ZH2d	Das	ist	vielleicht	der	Grund		,	aus	dem		ich		das	Wort	und	die	Ideologie	,	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.).

7.6 PiKUST (Poskusni korpus usvajanja slovenščine kot tujega jezika)[160]



7.6.1 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

[159] Reznicek et al. (2010a: 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.
[160] Kontakt: Mojca Stritar, http://kuscholarworks.ku.edu/dspace/bitstream/1808/5274/1/8Stritar.pdf

anotace, které by reflektovaly problémy slovanského, flektivního jazyka. Zpracovávání tohoto korpusu se úzce opírá o model norského žákovského korpusu ASK.[161] 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.[162] 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.).

7.6.2 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.

7.6.3 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

[161] ASK, tj. Norsk andrespråkskorpus.
[162] Korpus neobsahuje žádné texty respondentů na nízké úrovni znalosti slovinštiny s výjimkou tří textů od začátečníků Slovanů. Srov. Stritarová (2009: 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“.

chyby ve vidu, prefixaci, chyby slovnědruhové apod. Srov. např. příklad (17).[163] 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.

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

7.7 Závěr

Žákovské korpusy se 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. Srov. Díaz-Negrillo*vá* a Fernández-Domínguez (2006: 86). 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,[164] 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

[163] Bohužel nemáme k dispozici autentický záznam grafické anotace z korpusu PiKUST, rekonstruovali jsme tedy ukázku na základě korpusu ICLE.
[164] Podrobněji o klasifikaci chyb v textech nerodilých mluvčích viz např. James (1998), Dulay*ová* et al. (1984), a v této práci s. 1.6.3.3.

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[165] jsou kategorizovány 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. Dulay*ová* et al. (1982: 150n.) původně vymezují čtyři základní typy cílových modifikací (chybějící element, přebývající element, chybná forma/chybný výběr a chybný slovosled) a pokoušejí se tak uvést čistě deskriptivní taxonomii chyb, zaměřující se pouze na pozorovatelné, povrchové rysy chyb, která nepropojuje popis chyby s její explanací. Dulay*ová* et al. chápou takovou klasifikaci žákovských chyb jako reflexi kognitivních procesů odehrávajících se při studentově rekonstruování cílového jazyka. Tato myšlenka vyvolává řadu metodologických diskusí, protože předpokládá operace s povrchovými strukturami cílového jazyka spíš než budování interlanguage. (Ellis, 2003: 56) Minimálně se v chybové anotaci žákovských korpusů odráží jiné typologie chyb, např. Corderova (1974, s. 123n.) klasifikace na chyby formální utvářenosti výrazu[166] 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.[167]

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. Viz Fitzparick*ová* a Seegmiller (2003), Rastelli (2009) aj. Odmítnutí počítačem podporované chybové analýzy jako metodologického konceptu výzkumů jazyka nerodilých mluvčích na základě chybově anotovaného žákovského korpusu má kořeny v původní kritice

[165] Termín *deviace* používám ve smyslu odchylky od standardní podoby cílového jazyka.
[166] Toto termín používám v širším slova smyslu, tj. nikoli pouze ve významu české formální morfologie.
[167] Srov. např. komentář Granger*ové* (2003a: 467) k chybové anotaci žákovského korpusu FRIDA, která je „descriptive rather than interpretative".

tradiční chybové analýzy zaměřené především proti nedostatkům v její metodologii, tj. obtíže při identifikaci jednotného zdroje chyby[168] 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. Srov. Schachter*ová* a Celce-Murcia*ová* (1977), Long a Sato*vá* (1984), Van Els et al. (1984) aj.

Ačkoli vzhledem k uvedeným výhradám byla chybová analýza některými badateli odmítána jako nespolehlivá, je studium žákovských chyb kontinuálně využíváno jako validní součást performanční analýzy.[169] Chyby jsou neoddělitelnou součástí žákovského jazyka a jako takové by měly být podrobovány výzkumu stejně jako jiné aspekty mezijazyka. Obdobně jako Ringbom (1987: 69), Taylor (1986), Lennon (1991), Ellis (1994: 20), Granger*ová* (2003a: 466) aj. zastávám názor, že chybová analýza má své uplatnění při analýze nabývání cizího jazyka a velký dosah pro pedagogickou praxi.[170] 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). Dále i Milton a Chowdhury (1994: 130), Granger*ová* (2002: 14), Izumi et al. (2004: 35), Díaz-Negrillo*vá* a Fernández-Domínguez (2006: 86).

Ellis (2003), ve shodě s Corderem (1974) tvrdí, 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 něj především vymezení hranice chyby, detekce zdroje chyby

[168] Srov. tzv. „ambiguous goofs“, Dulay*ová* a Burt*ová* (1974b).
[169] K termínu viz např. Larsen-Freeman*ová* a Long (1992).
[170] „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: 69)

v mentálním lexikonu žáka a chápání povrchových modifikací jako akvizičních faktů. Zároveň zpochybňuje konzistentnost a spolehlivost chybového značkování v souvislosti s jeho interpretační povahou a subjektivitou. Ve shodě s nimi je Rastelli (2009), který při anotaci mezijazyka rezignuje na chybové značkování a navrhuje tzv. SLA-značkování (srov. také Rastelli 2007, Rastelli a Frontini*ová* 2008). 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, 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.

8 KORPUS ČEŠTINY JAKO DRUHÉHO JAZYKA

Problematika žákovského korpusu češtiny byla zpracována již v několika dílčích studiích. Štindlová (2011) představuje žákovské korpusy jako nový fenomén v bádání o akvizici cizího jazyka a popisuje možnosti odborného i didaktického využití korpusu jazyka nerodilých mluvčích. Šebesta (2010) zasazuje koncepci akvizičních korpusů do kontextu výzkumu osvojování jazyka i jeho domácích tradic a představuje jednotlivé projekty akvizičních korpusů češtiny AKCES, včetně korpusu žákovského. Šebesta (2011a) a Šebesta a Škodová (2011) zmiňují význam akvizičních korpusů včetně korpusů žákovských pro výzkumy osvojování jazyka rodilými i nerodilými mluvčími. Šebesta (2011b), Štindlová a Škodová (2011), Štindlová a Lábus (2009) podrobněji charakterizují korpus češtiny jako druhého jazyka a poskytují informace o základních parametrech jeho budování. Bedřichová, Šormová a Šebesta (2011) prezentují podrobněji připravovaný korpus jazyka romských žáků. Další studie se zaměřují na problematiku anotace korpusu nerodilých mluvčích češtiny. Škodová (2009) se obecně zamýšlí nad možnostmi zachycování chyb v žákovských korpusech, Hana et al. (2010) a Štindlová et al. (2011) poskytují informace o konceptu chybové anotace a návrhu anotačního formátu, Jelínek (2010) hovoří o problematice automatického zpracování manuálně anotovaného žákovského korpusu.

Následující kapitola shrnuje závěry všech dostupných studií o žákovském korpusu češtiny a doplňuje poznatky, které prozatím v odborném plénu prezentovány nebyly. Stručně představuje

vznikající korpus češtiny jako druhého jazyka a zmiňuje metadatové pozadí žákovských projevů. Dále se podrobněji soustředí na pravidla pro přepis rukou psaných textů Nečechů (odd. 8.3), která byla nově vyvinuta pro potřeby žákovského korpusu češtiny a která mohou sloužit jako inspirace i pro jiné nově vznikající korpusy, jež hodlají shromažďovat rukopisy nerodilých mluvčích. Pravidla pro přepis zde zevrubně zmiňuji z toho důvodu, že zcela zásadně ovlivňují podobu dat, která jsou dále postoupena k chybové anotaci a k aplikaci automatických anotačních nástrojů, a mohou tak zprostředkovaně ovlivňovat i celkovou podobu žákovského korpusu, resp. výstupních analýz na něm založených. Anotační model vytvořený pro potřeby značkování chyb v češtině jako cizím jazyce je představen v oddíle 8.4.

8.1 Korpus CzeSL

Žákovský korpus češtiny jako druhého jazyka (CzeSL)[171] je budován v rámci projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk*[172] ve spolupráci Technické univerzity v Liberci a Univerzity Karlovy v Praze.[173] Bude sloužit jako reprezentativní a spolehlivý zdroj poznání češtiny jako druhého jazyka, který nebyl až do této chvíle k dispozici. Koncepčně lze korpus češtiny nerodilých mluvčích zařadit do skupiny akvizičních korpusů AKCES (Akviziční korpusy češtiny), které vznikají na FF UK od roku 2005. Korpus CzeSL je vytvářen tak, aby byl v maximální míře kongruentní s ostatními korpusy této skupiny, tj. především s korpusy SCHOLA a SKRIPT.[174] Zamýšlená sourodost akvizičních korpusů mnohonásobně znásobí možnosti jejich využití. Viz Šebesta (2010, 2011a).

Korpus CzeSL, který by měl obsáhnout až dva miliony slov, se primárně soustředí na (1) jazykové projevy mluvčích s prvním jazykem slovanským, tedy příbuzným (většina textů pochází od mluvčích ruštiny a polštiny); (2) jazykové projevy mluvčích ze vzdálených

[171] Zkratka CzeSL je prozatím pracovní variantou, v některých textech lze dohledat alternativní podobu C2J.

[172] Projekt (CZ.1.07/2.2.00/07.0259) se realizuje v rámci Operačního programu Vzdělávání pro konkurenceschopnost a je financován ze zdrojů Strukturálních fondů EU (ESF) a státního rozpočtu České republiky. Příjemcem dotace je Technická univerzita v Liberci, na řešení se jako partneři podílejí Univerzita Karlova v Praze a Asociace učitelů češtiny jako cizího jazyka.

[173] Na budování korpusu se podílejí katedra českého jazyka a literatury FP TUL, na UK Ústav jazykové a odborné přípravy (sběr jazykových dat a metadat) a několik ústavů Filozofické fakulty: sběr jazykových dat a metadat, jejich přepisy a další zpracování zajišťují primárně pracovníci Ústavu bohemistických studií a Ústavu českého jazyka a teorie komunikace s dalšími spolupracovníky, přípravu anotací a jejich realizaci primárně pracovníci Ústavu teoretické a komputační lingvistiky s dalšími spolupracovníky, včetně autorky této práce. Kromě toho přispívá k budování korpusu řada dalších spolupracujících škol středních a základních, občanská sdružení a velký počet individuálních spolupracovníků, především, ale ne výlučně, studentů doktorského, magisterského i bakalářského studia obou univerzit.

[174] Korpus vyučovacích hodin SCHOLA byl budován v letech 2005 – 2008. Korpus SKRIPT je v současné chvíli ve výstavbě.

jazykových rodin (především Vietnamců); (3) jazykové projevy mluvčích jiných (indoevropských) jazyků. Specifickou skupinou, jejíž projevy budou do korpusu nerodilých mluvčích taktéž zahrnuty, tvoří romští žáci. Tento subkorpus je však částečně vytvářen odlišně, na základě některých specifických kritérií (podrobnější informace viz Bedřichová, Šormová, Šebesta, 2011). Psané i mluvené projevy, začleněné do korpusu, budou sociologicky i didakticky značkovány. Specifickým rysem českého žákovského korpusu je jeho značkování chybové, které jej ve světovém kontextu řadí k nemnohým chybově anotovaným korpusům nerodilých mluvčích. Žákovský korpus češtiny se na rozdíl od většiny světových žákovských korpusů (jako např. ICLE, ISLE, JPU, LeaP apod.) nezaměřuje na sběr materiálů jedné úrovně znalosti cílového jazyka, ale zahrnuje texty od mluvčích na všech úrovních znalosti češtiny. Tyto úrovně jsou v metadatové informaci definovány podle popisu SERR.

Žákovský korpus češtiny bude využíván jako didaktický nástroj pro přípravu budoucích pedagogů, bude sloužit jako pramen poznání mezijazyka a principů nabývání druhého jazyka, a v neposlední řadě jej bude možné využít jako inspirující zdroj dat pro tvorbu učebních materiálů a optimalizaci výukového procesu. Srov. i Štindlová (2011).

8.2 Metadata[175]

Pokud srovnáváme podrobnost externí anotace ve světových žákovských korpusech, můžeme český korpus žákovského jazyka přiřadit k těm detailněji značkovaným. CzeSL eviduje celkem osmnáct metadatových parametrů (srov. např. korpus ICLE, který jich uvádí 25, nebo korpus MELD s 21 parametry), dvanáct z nich se týká respondenta a dalších šest proměnných upřesňuje charakter materiálu, především podmínky jeho vzniku. Srov. analýzu žákovských korpusů v kapitole 7 této práce.

8.2.1 Metadata: respondent

Všechny texty nerodilých mluvčích, které budou zařazeny do žákovského korpusu češtiny, jsou opatřeny základními sociologickými charakteristikami autorů textů, konkrétně informacemi o jejich pohlaví, věku a prvním jazyku. Další obligatorní proměnné vypovídají podrobněji (1) o míře znalosti českého jazyka, která je vymezena úrovněmi A1 – C2 dle popisu SERR; (2) o způsobu osvojování cílového jazyka, tj. kde, jak dlouho a za jakých podmínek se žák učí češtinu (instituce, časový rozsah, výuka na historickém území České republiky, či nikoli); (3) o

[175] K tématu viz i článek Štindlová a Škodová (2011).

výukovém materiálu, resp. o učebnici, kterou respondent při studiu češtiny používá. Fakultativně mohou být doplněny informace o respondentově znalosti dalšího (nemateřského) jazyka, bilingvismu, délce jeho pobytu na území ČR, a také jeho rodinná jazyková anamnéza, tj. informace o tom, zda se v respondentově rodině vyskytují mluvčí českého jazyka. V rámci shromažďování metadat se nesledují detailně informace o dosaženém vzdělání respondenta, a to z několika důvodů: projekt se zaměřuje na sběr dat převážně od mladší populace s neuzavřeným vzděláním, informace o vzdělání jsou u části nerodilých mluvčích nedostupné a problémem je i divergence respondentů co do možnosti srovnání absolvovaných školských systémů.

8.2.2 Metadata: materiál

Vedle klasických textových značek, které poskytují informace o časové limitovanosti projevu nerodilého mluvčího, o požadovaném rozsahu, o práci s referenčními pomůckami a o testovém charakteru projevu, sleduje CzeSL u jednotlivých textů míru a způsob jejich elicitace.

Protože jazykové projevy pro korpus CzeSL jsou sbírány i v semestrálních, ročních a příp. delších jazykových kurzech, jsou značkovány tak, aby bylo možné dohledat všechny texty jednoho autora a analyzovat je vzhledem k časově podmíněným proměnám mezijazyka.

Tzv. anamnestický dotazník o autorovi textu a průvodka k nasbíranému materiálu jsou uvedeny v příloze 4.[176]

8.3 Přepis materiálů pro žákovský korpus češtiny jako druhého jazyka

První fází při zpracování dat sbíraných pro korpus žákovského jazyka nerodilých mluvčích je jejich převod do elektronické podoby. Pro některé ze světových žákovských korpusů není tato etapa relevantní, protože se zaměřují na sběr materiálů v elektronické podobě (např. ICLE, CLC, LLC aj.), v pravém slova smyslu tedy nasbírané texty nepřepisují, pouze standardizují podobu sociolingvistického značkování a formátování dokumentu. Řada dalších žákovských korpusů však akceptuje i rukopisy a některé jsou na tento druh materiálu dokonce výhradně zacíleny. Tyto korpusy se musí vyrovnávat s otázkami spojenými s digitalizací textů a zabývat se problematikou rukopisných anomálií. Řešení přepisu se však v jednotlivých korpusech značně odlišují. Problémy převodu rukou psaných textů do elektronické podoby jsou zřetelné: proces elektronizace materiálů je časově velmi náročný a klade poměrně značné požadavky na přepisovače, kteří musí činit v nejednoznačných případech klíčová rozhodnutí, jež mohou

[176] Autorkou obou materiálů je B. Štindlová.

ovlivnit podobu dat zařazovaných do vznikajícího korpusu a modifikovat charakter následných analýz.

V této kapitole představím návrh přepisovacích pravidel pro rukou psané texty nerodilých mluvčích českého jazyka a zdůvodním jednotlivé postupy v rámci tohoto návrhu. V české korpusové tradici máme v této oblasti jen minimální podporu, jediným korpusem sbírajícím rukopisy je Korpus soukromé korespondence, jehož stručná pravidla pro přepis zde taktéž uvedeme a budeme je konfrontovat s potřebami přepisu textů nerodilých mluvčích. Zaměříme se zároveň na doporučení pro přepis rukopisů, která uvádí Text Encoding Initiative (TEI).[177] Tato doporučení jsou často výchozím bodem pro přepis rukou psaných textů v mnoha světových žákovských korpusech (viz např. ASK, ICLE, EARS aj.), ačkoli samozřejmě dochází k jejich výrazné úpravě a rozšíření. V případě tvorby přepisovacích pravidel pro žákovský korpus češtiny bylo třeba také zohlednit jeho pozici mezi skupinou akvizičních korpusů AKCES a dbát na kompatibilitu s budovaným korpusem slohových prací žáků SKRIPT.

8.3.1 TEI – doporučená pravidla pro přepis rukopisů

V kontextu současné korpusové lingvistiky se pravidla pro přepis rukou psaných textů vztahují především k přepisu starých rukopisů. Hlavním cílem zpracování historických textů pro korpus je především jejich snadné a všestranné prohledávání, zachování vysoké míry lingvisticky relevantních informací a zpřístupnění zpracovaného korpusu současným badatelům. Z toho důvodu se řada těchto korpusů zaměřuje na transkribování originálního textu a na rekonstrukci (tzv. normalizaci) v případě, že ke grafické podobě nelze jednoznačně přiřadit odpovídající transkripci.[178]

Autoři korpusů historických manuskriptů, stejně jako autoři korpusů zahrnujících jiné typy rukopisů obvykle při vymezování přepisovacích pravidel vycházejí z doporučení TEI, které rozšiřují podle potřeb konkrétního zacílení jednotlivých korpusů.

V následující části budeme prezentovat koncepci TEI pro kódování rukou psaných materiálů na základě verze P5 (2007)[179]. V sekci *Simple Editorial Changes* (kapitola 3.4) jsou uvedeny tři základní typy intervencí do původního textu (tzv. *editorial interventions).* Jedná se o opravy tzv.

[177] *The Text Encoding Initiative* (TEI) je odborné společenství, které kolektivně vytváří a prosazuje standardy pro reprezentaci textů v digitální podobě. Jeho hlavním úkolem je sestavovat metodické pokyny, které specifikují kódovací metody pro strojově čitelné texty, a to především v oblasti humanitních věd, včetně lingvistiky. Viz dále na http://www.tei-c.org/index.xml

[178] Srov. např. principy kódování diachronního korpusu DIAKORP, http://ucnk.ff.cuni.cz/diakorp.php#transkripce

[179] Viz http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COED, kapitoly 3.4 a 11.5.1.

zjevných chyb, regularizaci variant, resp. nestandardních forem, a zásahy ve smyslu doplnění, odstranění apod. U zjevných chyb nabízí TEI dvojí způsob kódování: označení provedené emendace <corr>, nebo jednoznačnou identifikaci chybového místa bez přímé korekce <sic>. Pokud byla provedena oprava, uvádí TEI možnost označení tzv. odpovědnosti za korekci a stupně spolehlivosti korekce. V případě široce uplatňovaných variantních forem, nebo nestandardního pravopisu navrhuje TEI provést tzv. regularizaci, to znamená stanovit standardní, resp. normalizovaný ekvivalent nestandardních forem <reg>. Za předpokladu, že nechceme text korigovat, lze využít označení signalizující původní, tj. neopravenou variantu <orig>. Pravidla TEI umožňují sledovat, zda byl úsek textu přepisovačem vynechán <gap>, obvykle pro nečitelnost, v případě mluvených korpusů neslyšitelnosti, zda byla část textu autorem doplněna <add>, nebo vymazána <del>. V rámci pravidel TEI je také umožněno zaznamenávat přítomnost různorodých grafických informací, např. obrázků, diagramů, ať již pevně pozičně ukotvených <graphic>, či nikoli <figure>. Přepisovací pravidla TEI nabízejí základní způsob kódování některých problematických jevů, kterými se rukopisy vyznačují. Nezachycují však všechny specifické rysy rukou psaných materiálů nerodilých mluvčích, jako jsou např. škrty, přesuny v textu, autorské alternativy atd., které chápeme jako relevantní pro analýzy nabývání druhého/cizího jazyka. Proto je třeba koncept kódování TEI modifikovat podle konkrétních potřeb žákovského, resp. obecněji akvizičního korpusu.

8.3.2 Přepis textů v Korpusu soukromé korespondence[180]

Vedle korpusů skupiny AKCES[181] se v českém prostředí v současnosti pouze jediný korpus zabývá digitalizací a následnou anotací rukou psaných materiálů. Jedná se o Korpus soukromé korespondence (KSK), který zachycuje „možná v posledním existenčním stadiu tradiční ručně psanou korespondenci“[182]. KSK byl sestaven v Ústavu českého jazyka na Masarykově univerzitě v Brně pod vedením Z. Hladké s cílem vytvořit veřejně přístupný referenční zdroj pro studium současné soukromé korespondence a naznačit možnosti lingvistického využití epistolárních textů. Korpus zahrnuje přibližně dva tisíce dopisů z let 1990 – 2004. E-maily a SMS zprávy, které jsou součástí databanky, o niž se KSK opírá, prozatím do korpusu vtěleny nejsou. Viz Hladká (2005).

[180] Za podrobnější poznámky k přepisům textů pro korpus KSK děkuji doc. PhDr. Zdeňce Hladké, Dr.
[181] Zaměření na rukopis cizinců vyvolal v rámci korpusů AKCES potřebu podrobnějších a důkladnějších pravidel pro přepis rukou psaných textů, ta byla zpracována tak, aby vyhovovala posléze všem typům jazykových dat v korpusech AKCESu (např. odstavcové členění nebo opravy učitele jsou pro jazyk cizinců vcelku málo relevantní nebo irelevantní, pro české žáky relevanci mají).
[182] http://ucnk.ff.cuni.cz/dopisy.php

Autoři KSK byli nuceni vyrovnat se s některými specifiky ručně psaných textů a navrhli postup přepisu pro tento typ materiálu. V zásadě se snaží podat co nejpřesnější obraz originálu, aby bylo možné jednoznačně rekonstruovat původní podobu textu. V přepisu jsou zachovány všechny chyby, včetně pravopisných, z důvodu snazšího korpusového vyhledávání jsou upravovány pouze případné chybné delimitace. Tento přístup ne zcela koresponduje s doporučeními TEI.

Zvláštním znakem jsou kódovány identifikační údaje, stejně tak jsou zachycovány také informace o grafické úpravě dopisů. Specificky jsou značkovány citátové pasáže, které lze v korpusu zviditelnit, nejsou však zahrnuty do vyhledávacích procedur a statistických analýz.

Autoři KSK jsou si vědomi problematičnosti při zachycování pravopisných chyb v textech, protože tyto chyby, ačkoli jejich zaznamenání je jistě přínosné (např. pro srovnání dodržování pravopisných norem, hledání současných pravopisných tendencí apod.), znesnadňují aplikaci automatického morfologického značkování a zároveň problematizují vyhledávání výskytů slov a tvarů např. při frekvenčních analýzách. Autoři KSK zvažovali alternativu paralelního zaznamenávání chybné i správné varianty již při přepisu, nakonec je však tato problematika řešena až ve fázi morfologického značkování při tzv. disambiguaci[183] pomocí uvedení pravopisně správného lemmatu a poznámkového aparátu.

Přepisovací pravidla pro KSK se pokouší vyrovnat i s chybami v diakritice, které jsou charakteristické především pro elektronickou korespondenci. V textech, které jsou zcela bez diakritiky, je diakritika doplňována, texty s částečným užitím diakritiky jsou ponechány v reálné podobě. Je zřejmé, že jde o problematické disparátní řešení, které bylo navrženo především pro snadnou aplikaci automatické procedury.

Pro přepisy, škrty a opravování v textu se KSK řídí principem „poslední ruky“ a následné komentáře jsou vtělovány do poznámek. Nečitelné části textů se při přepisu korespondence nerekonstruují, vpisky jsou řazeny na místo vsunutí a dodatečné informace jsou uvedeny v komentářových poznámkách. Veškeré vysvětlivky a doprovodné komentáře jsou tedy v přepisovaném textu uvedeny jako poznámky v hranatých závorkách, přičemž jsou opakované typy informací standardizovány.

Gramatické značkování a lemmatizace jsou v korpusu KSK aplikovány pouze na polovinu textů, protože se jedná o ruční a náročnou proceduru. Obecně se KSK zaměřuje především na detailnější sociolingvistickou anotaci, která umožňuje práci s vybranými subkorpusy.

[183] Disambiguace, tj. volba správné morfologické a slovnědruhové interpretace na základě kontextu.

Koncept přepisu soukromé korespondence je zpracovaný s ohledem na povahu korpusu KSK. Podle našeho názoru pracuje s přetíženým poznámkovým aparátem, který není plně standardizován, není tedy možné jej snadno využít k vyhledávání. Některá řešení jsou pro potřeby žákovského korpusu inspirativní, např. kódování grafických úprav a přístup k pravopisným chybám, některá řešení se v souvislosti s texty nerodilých mluvčích jeví jako nedostatečná, např. zaznamenávání přepisů, škrtů, příp. autorských alternativ.

8.3.3 Přepis textů v žákovských korpusech

Žádný ze současných žákovských korpusů se nezaměřuje na výzkum písma. Obdobně je v odborné literatuře, která se zabývá analýzami jazyka nerodilých mluvčích na základě žákovských korpusů, opomíjena problematika rukou psaných textů, ačkoli přinejmenším v případě, že první jazyk studenta užívá jiný ortografický systém než jazyk cílový, může být z hlediska SLA, ale i FLT zajímavou výzkumnou oblastí.[184] Přestože však není problematika rukopisů v kontextu výzkumů žákovského jazyka atraktivní, všechny žákovské korpusy, které akceptují rukou psané texty, musí vymezit svůj přístup k jejich digitalizaci a navrhnout přepisovací pravidla. Tato pravidla obvykle nejsou veřejně dostupná, ale z dílčích komentářů ve studiích lze odvodit, že jednotlivé korpusy řeší problematiku přepisu rukou psaných textů velmi různorodě, ačkoli obecně navazují na doporučení TEI.

Jednou z otázek souvisejících se snahou zachovat co největší množství informací o původním textu i v jeho elektronizované podobě je to, do jaké míry v přepisech textů nerodilých mluvčích zaznamenávat jejich grafické, resp. ideografické charakteristiky, jako je např. značení odstavců, hranic věty, citací, podtržení, symbolů apod. Ačkoli žákovské korpusy mají tendenci k tomu, obsahovat množství specifických grafických prostředků, měl by být stupeň jejich evidence podle Grangerové (1998: 12) nízký, přestože by zaznamenávání těchto jevů nemělo být zcela vyloučeno. Domnívám se, že takový minimalistický přístup k reflektování grafických atributů v psaných projevech nerodilých mluvčích není nejvhodnější, protože typickým znakem psaných cizojazyčných textů je, že komplexně prezentují odchylky od standardního užití jazyka, a pokud by tyto odchylky nebyly v případě grafických prostředků zachyceny, byla by částečně omezena i využitelnost žákovského korpusu (srov. Somers, 2005: 151).

[184] Výjimkou je sborník COOK, V., BASSETTI, B. (eds.) *Second Language Writing Systems.* Clevedon/Buffalo: Multilingual Matters Ltd, 2005. Srov. také SASSON, R. *The Acquisition of a Second Writing System.* Oxford: Intellect, 1995.

Část žákovských korpusů se přiklání k ortografickému přepisu, příp. hybridnímu typu přepisu, kdy jsou v souvislosti se zaměřením korpusu některé grafické, příp. pravopisné nedostatky normalizovány, některé jsou ponechány v originální, tj. nestandardní podobě k dalšímu zpracování. Tyto korpusy se liší přístupem ke značkování těchto narovnání: buď jsou při přepisu zaznamenávána formalizací (BUiD), nebo nikoli (PiKUST, MELD).

8.3.4 Pravidla pro přepis textů nerodilých mluvčích českého jazyka[185]

Pro účely vznikajícího korpusu češtiny nerodilých mluvčích se sbírají i texty v elektronické podobě, ale primárním zdrojem dat jsou rukopisy. Důvody ke sběru rukopisných textů jsou pragmatické, protože naprostá většina materiálu, který je v této oblasti k dispozici, jsou rukou psané texty, např. eseje, domácí úkoly, apod. Výhodou sběru rukopisů je fakt, že je možné se vyhnout automatickým opravám pravopisu i gramatiky, které textové editory běžně poskytují. Cílem žákovského korpusu češtiny je poskytnout celkový obraz žákovského jazyka, který by byl v případě automatické kontroly značně deformován.[186]

Pokud jsou tedy pro potřeby jazykového korpusu sbírány rukou psané materiály, není možné aplikovat metody adaptace elektronických dat ani skenování dokumentů s využitím optického rozpoznávání znaků (OCR).[187] Jediným způsobem, jak lze data získat, je jejich přepis. Při přepisu však není možné vyhnout se přepisovacím chybám (srov. Dagneaux*ová* et al., 1998; Granger*ová*, 2002; Somers, 2005 aj.), a proto je zcela zásadním aktem supervize přepisu. V souvislosti s digitalizací nestandardních textů nerodilých mluvčích je kontrola přepsaných textů ještě podstatnější, protože supervizor musí odlišit chyby, které vznikly v průběhu přepisu, od chyb, které jsou součástí původního textu a které musí zůstat zachovány. Granger*ová* (1998: 11) k tomu poznamenává, že „supervizor se musí ujistit, že vypustil všechny chyby, které byly do textu zaneseny při přepisu, … ale že ponechal chyby, které byly v původním žákovském textu uvedeny; *[jde o]* ošidný a časově náročný úkol.“ (Doplnila BŠ.) Z toho důvodu jsou všechny přepsané texty pro korpus češtiny jako druhého jazyka revidovány a pro případnou pozdější verifikaci přepisu bude korpus provázán s databází skenů originálů textů, které však nebudou veřejně přístupné.[188]

---

[185] Ve své práci se zabývám pouze problematikou psaných projevů. Pro korpus češtiny jako druhého jazyka jsou však sbírány i projevy mluvené, na něž se ovšem v našem textu nezaměřujeme.

[186] Vzhledem k tomu, že sbírané texty nejsou homogenní, resp. nepocházejí z centrálně řízených situací, nemůže být v případě elektronických textů zaručeno, že nebudou zneužívány automatické opravy.

[187] OCR, tj. *Optical Character Recognition*, digitalizace tištěných textů.
K metodám získávání korpusových dat viz více např. Sinclair, J. M . Corpus, Concordance, Collocation . Oxford: Oxford University Press, 1991.

[188] Pro další informace k technickým aspektům přepisu textů nerodilých mluvčích viz www.c2j.cz.

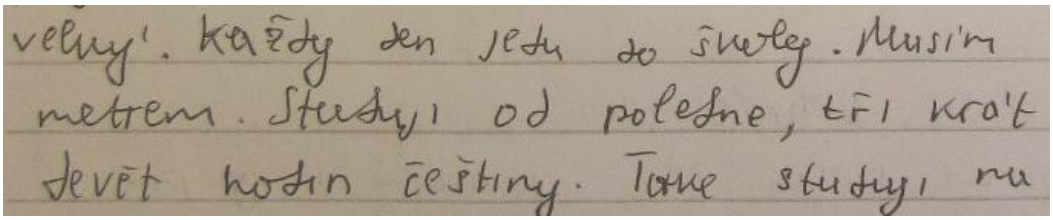
8.3.4.1 Koncept přepisu

Pro potřeby žákovského korpusu českého jazyka byla zavedena přepisovací pravidla, která v nejvyšší možné míře reflektují původní podobu textů, zároveň však umožňují jejich následné manuální i automatické zpracování, tj. chybovou anotaci i anotaci lingvistickou (morfologickou a slovnědruhovou). Původní návrh, který připravila autorka tohoto textu, byl částečně upravován v souladu s vývojem značkovacího programu FEAT a v souvislosti s potřebami dalších akvizičních korpusů AKCESu. V průběhu testování navržených pravidel pro přepis v jednotlivých korpusových skupinách (tj. při přepisu slohových prací českých žáků, při přepisu romských textů a při přepisu projevů nerodilých mluvčích) docházelo k dílčím modifikacím podle jednotlivých typů materiálů. Dále zde budu komentovat pouze podobu přepisovacích pravidel, která se užívají pro přepis textů nerodilých mluvčích v žákovském korpusu CzeSL.

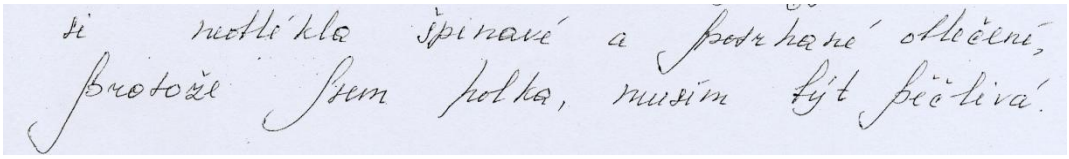
Ve fázi přepisu rukou psaných materiálů nerodilých mluvčích češtiny bylo rozhodnuto zaznamenávat pouze grafické aspekty, formální strukturaci textu a formalizovat autorské zásahy do textu. Toto rozhodnutí vychází představy, že problematika ortografie v širokém slova smyslu patří k základu chybového značkování, neměla by být tedy specifikována na úrovni přepisu. Podrobnější kategorizace široce chápaných ortografických chyb (včetně chyb typografických, fonologických záměn, morfologického nadbytečného zpravidelňování (*overregularization)* apod.) může podpořit výzkumy nabývání druhého/cizího jazyka. Umožní srovnání ortografických chyb u mluvčích s odlišnými prvními jazyky např. v souvislosti s vnímáním fonizace. Vedle toho se analýzy na základě chybového značkování ortografie (v nejširším slova smyslu) můžou významně podílet na vývoji automatické detekce pravopisných chyb v textech nerodilých mluvčích a podstatně můžou ovlivnit i cizojazyčnou výuku.

V přepisech projevů nerodilých mluvčích češtiny se nezachycují volné varianty grafémů, resp. idiosynkratické varianty jednotlivých písmen jako projevy systematické variantnosti autora, ačkoli jsem si vědoma, že by mohly signalizovat např. vliv grafické podoby prvního jazyka studentů. Viz př. (18 a,b,c).

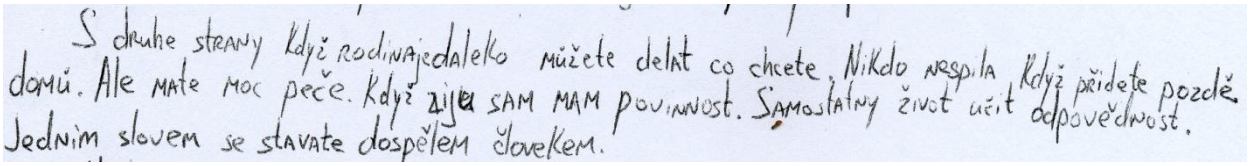
(18) a. Tři různé způsoby zápisu <d> (ukázka z textu KRIS_ AO_003 – II).


velky'. každy den jedu do školy. Musím
metrem. Studuji od poledne, tři krát
devět hodin češtiny. Tomé studuji mu

b. Idiosynkratické užívání zápisu <p> v pozici na začátku slova (ukázka z textu VOB_SA_020).

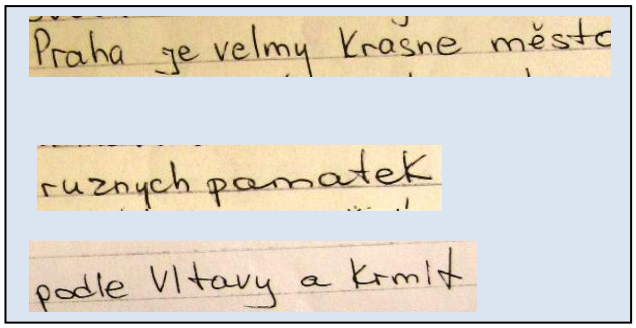

si motýlka špinavé a potrhané oblečení,
protože jsem holka, musím být pečlivá.

c. Konzistentní užívání zápisu <k>.
(ukázka z textu NEM_NI_009).


S druhé strany když rodina je daleko můžete dělat co chcete. Nikdo nespí, když přijdete pozdě
domů. Ale máte moc péče. Když žijete SAM MÁM povinnost. Samostatný život učí odpovědnost.
Jedním slovem se stáváte dospělém člověkem.

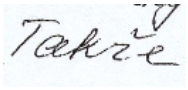
S tímto rozhodnutím je spojen problém obtížně vymezitelného rozdílu mezi (pravopisnou/*spellingovou)* chybou a grafickou anomálií, tj. přepisovač se musí rozhodnout, zda bude daný prvek považovat za jazykový nedostatek a v přepisu jej zaznamená, nebo jej bude regularizovat ve smyslu individuální varianty. Srov. ukázky v příkladu (19), kdy přepisovač může volit mezi zápisem pravopisné chyby *Praha je velmy* ***Krasne*** *město / ruznych* ***pamateK*** */ podle Vltavy a* ***Krmit,*** a mezi předpokladem varianty zápisu grafému „k“.

(19) Zápis grafému „k“ (KRIS_ AD_003)

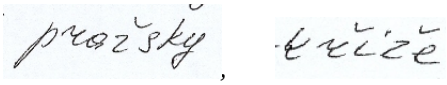

Praha je velmy krasne město
ruznych pamatek
podle Vltavy a krmit

Toto rozhodnutí je samozřejmě podstatně ovlivněno dvěma vnějšími faktory, za prvé zda má přepisovač k dispozici dostatečně obsáhlý grafický kontext pro porovnání záznamu dané grafické jednotky (srov. příklad 20); a za druhé zda zná první jazyk autora textu, resp. zda je obeznámen s jeho grafickým charakterem.

(20) Vliv grafického kontextu na přepis



Před srovnáním s odpovídajícími grafémy v grafickém kontextu přepisovač zaznamenal následovně: *Tak{ž|ř}e*



Po srovnání přepisovač rezignoval na variantní zápis a uvedl správně: *Takže*

8.3.5 Přepisovací pravidla[189]

8.3.5.1 Zásady přepisu

Formát přepisu, který byl zvolen pro digitalizaci textů nerodilých mluvčích češtiny, je lineární, jednovrstvý, a strukturní, resp. metajazykové informace jsou v něm prezentovány pomocí formálního aparátu. Primárním cílem přepisu dat pro korpus češtiny jako druhého jazyka je v maximální míře zachytit autentickou podobu zdrojového textu. Při přepisu jsou tedy zachovávány všechny jazykové nedostatky materiálu a je registrována většina jeho grafických specifik. Zároveň je třeba respektovat fakt, že možnosti vyhledávání dat jsou vždy determinovány i zvoleným způsobem transkripce. Snažili jsme se proto stanovit taková pravidla pro přepis, která by následně umožňovala manuální chybovou anotaci textu a která by nebyla překážkou pro aplikaci automatické lingvistické anotace v další fázi zpracování textů. Na rovině přepisu jsme rezignovali na ortografický přepis, který by usnadňoval lemmatizaci, morfologické i slovnědruhové značkování a aplikaci již existujících nástrojů, a tato problematika byla transponována na úroveň chybové anotace dat, kde se prezentuje rekonstrukce chybového textu, která je podrobena automatické proceduře v třírovinném anotačním schématu.

[189] Autorkou pravidel pro přepis textů nerodilých mluvčích českého jazyka je Barbora Štindlová. Technické zpracování a konkrétní podobu metaznaků navrhl ve spolupráci s autorkou a ÚTKL Jiří Hana, grafickou podobu manuálu pro přepis zpracoval Václav Lábus. Za konzultace, poznámky a návrhy k dané problematice děkuji týmu ÚTKL FF UK, Karlu Šebestovi a Svatavě Škodové.

Při přepisu textů nerodilých mluvčích češtiny jsou jen minimálně využívána doporučení TEI: neregularizuje se nestandardní pravopis, kódování nečitelných úseků a autorských rektifikací se maximálně zjednodušuje. Na druhou stranu se zavádí pro potřeby elektronizace rukou psaných textů nerodilých mluvčích jazyka množství kódů, které pravidla TEI nezahrnují.

8.3.5.1.1 Formát přepisu

K vytvoření přepisu jsou standardně užívány běžné textové editory a přepisy jsou ukládány ve formátu *.htm*. Pro další zpracování je nutné data převést pomocí konverzního programu do formátu XML.[190] Z toho důvodu je nezbytné, aby byl přepis textů dostatečně formalizován. Strukturace přepisu psaných dat je nesporně mnohem jednodušší, než je tomu v případě transkripce mluvených projevů. Chybové texty se však vyznačují, jak jsem již zmínila výše, řadou specifických atributů, které je nutné při převodu do elektronické podoby zachovat, proto byla k jejich zachycení definována množina symbolů a kódů, jejichž přehled je uveden v následujícím oddíle. Přepis textů nerodilých mluvčích, který žákovský korpus CzeSL používá, je usnadněn v tom smyslu, že nezahrnuje externí metadata, která jsou v systému ukládána separátně. Formát přepisu charakterizuje následující obecné schéma a konkrétní ukázka z přepisu dat pro korpus nerodilých mluvčí češtiny (příklad 21).

8.3.5.1.2 (Meta)znaky a kódy přepisu – přehled

V tomto oddíle uvádím souhrnný přehled všech přepisových znaků a kódů s následnou souvislou ukázkou přepisu.

Přepisové metaznaky:

dvě mezery	hranice mezi větami
prázdný řádek	oddělení odstavce, nadpisu
< >	kód (viz níže)
	odrážka
<bar>	přepis znaku |

[190] XML (Extensible Markup Language) je univerzální, rozšiřitelný značkovací jazyk pro popis strukturovaných dokumentů. Je mezinárodně standardizován a umožňuje snadné vytváření konkrétních značkovacích aplikací pro různé účely a různé typy dat. Blíže viz http://www.w3.org/XML/.

{ }	uzávorkování • variantních výrazů a výrazů opatřených kódem (není třeba u jednoslovných výrazů) • znaků/řetězců, jejichž přítomnost není v rukopisu jistá • přesunů • komentářů
[ ]	záznam jednoho grafému přepsaného více znaky, např. cedilla ç jako [c,]
XXX	nečitelný text
->	lokální přesun
~~*škrt*~~	škrtnutý řetězec
|	varianta
||	autorské rozdělení řetězců
+	autorské spojení řetězců

Přepisové kódy:

tr	přesun textu nebo jeho části
in	vsuvka do textu
st	alternativa od autora textu
pd	škrt provedený třetí osobou, pokud znečitelnil text (např. oprava učitele)
ni	interpretace zcela nečitelného textu
gr	text zapsán jiným písmem než latinkou
img	obrázek
co	komentář

Pro přepis textů se používá několik vyhrazených znaků (viz výše). Jedná se především o tři typy závorek: lomené, složené a hranaté. Funkce lomených závorek je metatextová, kódy v nich uváděné specifikují charakter zaznamenávaných aspektů textu. Pomocí lomených závorek se vymezují tři odlišné typy jevů: autorské zásahy do textu (přesuny, vsuvky), nejednoznačné interpretace (varianty), výskyt nejazykových grafických prostředků (obrázky, symboly). Složené závorky (srov. ukázka výše) slouží na textové rovině k vymezení rozsahu konkrétních problémových úseků. Jsou v nich vyjádřeny varianty v případě, že rukopis umožňuje různé čtení;

je v nich zaznamenán znak, resp. řetězec, jehož přítomnost není v rukopisu zřetelná; ve složených závorkách je uveden i přesunutý výraz či blok výrazů. Hranaté závorky jsou monofunkční, slouží výhradně k zaznamenávání cizích grafémů, obsahujících netradiční diakritická znaménka (např. tildu u *ñ* lze zaznamenat jako *[n~]* apod.), příp. k uvedení grafémů, které se v češtině nevyskytují, např. *[j']*, tj. *j* s čárkou apod. Registraci těchto atypických jevů pokládáme za důležitou např. pro pozdější analýzy variant fonologického zápisu v češtině nerodilých mluvčích v souvislosti s jejich mateřským jazykem.

(21) a. Obecné schéma strukturovaného přepisu (Nem-GD-008)

<*odrážka*>Text text text text text ~~škrt~~ text. *konec věty*Text{text*vložení*}<in> text text, text text text text text text text. *konec věty*Text text text text text, text text text text, text text text text text text text text. *konec věty*Text, text text text text, text text text text, text text text {*přesun* text text-> text text} text text text text text text {*vložení*text text text text text}<in>. *konec věty*Text text text text text text text text, text text text text text. *konec věty*Text text text text, text, text text text text ~~škrt~~ tex~~škrt~~t text text text text text text text text text.

odstavec

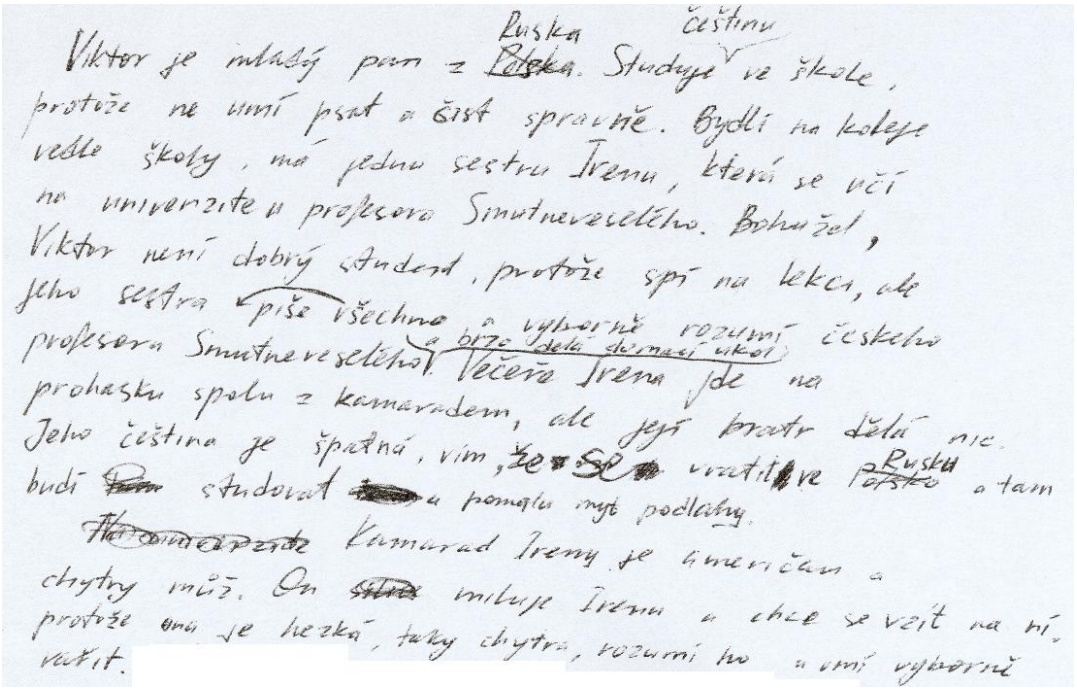
<*odrážka*>Text text text {T|t*varianta*}ext text text text. *konec věty*Text text text text text text text text. *konec věty*text text text text, text text, text text text text text text.

b. Modelová ukázka přepisu

Viktor je mladý pan z ~~Polska~~ Ruska. Studuje {češtinu}<in> ve škole, protože ne umí psat a čist spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra {piše všechno -> všechno piše} a vyborně rozumí českeho profesora Smutneveselého {a brzo delá domací ukol}<in>. Večeře Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve ~~Polsko~~ Rusk~~o~~u a tam budí studovat u pomalu myt podlahy.

Kamarad Ireny je {A|a}meričan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborně vařit.

c. Originál přepsaného textu



Viktor je mladý pan z Ruska. Češtinu Studuje ve škole, protože ne umí psát a číst správně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra píše všechno a výborně rozumí českého profesora Smutneveselého. a brzo dělá domácí úkol Večera Irena jde na prohaszku spolu z kamaradem, ale její bratr dělá nic. Jelo čeština je špatná, vim, že se vrátit ve Rusku Polsko a tam budí studovat a pomalu mýt podlahy. Kamarad Ireny je Američan a chytry muž. On miluje Irenu a chce se vzít na ní, protože ona je hezka, taky chytra, rozumí ho a umí výborně vařit.

8.3.5.2 Přepis dílčích jevů

8.3.5.2.1 Záznam autorských rektifikací

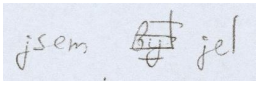
Sledování vlastních oprav *(self-repairs,* příp. *self-corrections)* žáků je tématem mnoha teoretických výzkumů zaměřujících se jak na rodilé, tak nerodilé mluvčí. V analýzách nabývání druhého jazyka je žákova vyvíjející se schopnost k sebekorekci signálem rostoucí jazykové kompetence a metalingvistického povědomí (Fathman*ová*, 1980; Verhoeven, 1989; van Hest*ová*, 1996; Kormos*ová*, 1998).[191] Rektifikace jsou chápány jako indikátory učení, poskytující vhled do vývoje žákova mezijazyka. Objevují se v situaci, kdy nerodilý mluvčí rozpozná, že jeho output je chybný nebo v nějaké ohledu nepřijatelný. Foster*ová* a Ohta*ová* (2005: 420) definují vlastní opravu jako „spontánně spuštěnou sebekorekci objevující se, když student opravuje svou vlastní promluvu, aniž by k tomu byl podněcován jinou osobou."

Vzhledem k povaze dat, která se sbírají (tj. heterogenní rukopisy), není možné využívat moderní technické prostředky, které slouží ke zkoumání sebekorekčních strategií, např. software, který by

[191] FATHMAN, A. K. Repetition and correction as an indication of speech planning and execution processes among second language learners. In *Towards a Crosslinguistic Assessment of Speech Production.* Eds. H. W. Dechert , M. Raupach. Frankfurt: Peter D. Lang, 1980, s. 77–85.
VERHOEVEN, L. T. Monitoring in childern's second language speech. *Second Language Research,* 1989, no. 5, s. 141-155.
VAN HEST, E. *Self-repair in L1 and L2 production.* Tilburg: Tilburg University Press, 1996.

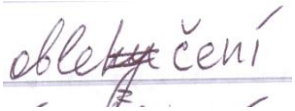
v případě homogenního sběru elektronických materiálů zachycoval všechny provedené přesuny, psaní, vymazávání apod., srov. Smith, 2008. Přesto chápeme zaznamenávání zjevných oprav[192] na úrovni textu, tj. v našem případě evidování autorských škrtů, přesunů a vsuvek, jako přínosné pro analýzy nabývání češtiny jako cizího jazyka, tj. přínosné pro sledování, jak nerodilí mluvčí testují hypotézy o cílovém jazyce.[193] Autorské rektifikace jsou na úrovni chybové anotace skryté a anotátoři pracují při značkování s textem „poslední ruky“. Protože jsou však sebeopravy formalizovány, jsou prohledávatelné a využitelné k analýzám, viz příklad 22.

(22) Ukázka záznamu škrtu

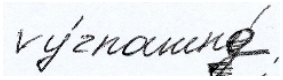


jsem ~~Byl~~ jel

t~~o~~uristy.



oble~~kyč~~ení



význam~~n~~~~ý~~é

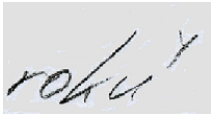
8.3.5.2.2 Varianty

Běžným znakem žákovských textů je uvedení alternativ v případě, že si student není jistý formou výrazu v daném kontextu (viz př. 23). Tyto alternativy jsou také chápány jako sebekorekční varianty ve smyslu výše vymezené zjevné opravy. Autorské alternativy jsou při přepisu zaznamenávány pomocí kódu <st>.

[192] K termínu viz Levelt (1983).

[193] Jsme si vědomi, že výzkumy vlastních oprav jsou primárně zaměřeny na mluvenou komunikaci (viz vlivný model řečové produkce Willema Levelta; resp. jeho popis struktury sebekorekcí pomocí pravidel *constituent rule* a *well-formedness rule).* Srov. Levelt (1983) a LEVELT, W. J. M. *Speaking – From Intention To Articulation.* Cambridge: MIT Press, 1989.
Data sbíraná pro korpus češtiny jako druhého jazyka jsou z větší části nepřipravené jazykové projevy a vykazují množství atributů charakteristických pro mluvené komunikáty. Předpokládáme tudíž, že závěry plynoucí z odborné diskuse o sebeopravách lze v modifikované podobě aplikovat i na ně. Tento názor opíráme o analýzy sebekorekcí v psaných komunikátech u Smithe (2008) a Kowalové (1999).

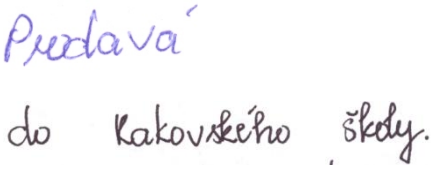
(23) Alternativní zápis autora textu



roku{u|y}<st>

V případě, že rukopis neumožňuje jednoznačné čtení, navrhuje přepisovač možné varianty zápisu, viz příklad (24). Ani v tomto případě tedy neprovádí regularizaci. Cílem tohoto postupu je minimalizovat normalizační zásahy přepisovače a omezit tak míru subjektivity přepisu.

(24) Ukázka záznamu variant



pr{o}dává

do {R|K}akovského školy.

Zaznamenáváním variant je samozřejmě znesnadněno následné zpracování dat. A to jak chybovou anotaci, protože se nabízí možnost dvojího značkování, tak aplikaci automatické lingvistické anotace (slovnědruhové a morfologické). V rámci navrženého anotačního schématu nemůžeme zpracovávat alternativní anotaci podle zvolené varianty. Srov. oddíl 8.4 o anotačním schématu pro korpus češtiny jako druhého jazyka. Varianty, autorské i od přepisovače, eliminuje podle zadaných pravidel anotátor, ale díky formalizaci jsou stejně jako autorské rektifikace dostupné k následným výzkumům.

8.3.5.2.3 Nečitelné řetězce

Nečitelné řetězce, resp. úseky textu se zapisují třemi křížky, tj. jako XXX, viz příklad (25). Je-li přepisovač schopen identifikovat množství nečitelných slov, uvede přepisovací znak v příslušném počtu.

Specifický případ nastává, je-li řetězec nečitelný, ale přepisovač je schopen na základě kontextu nebo jiných skutečností daný úsek rekonstruovat. V tom případě je tato interpretace značkována kódem <ni> a při chybové anotaci s ní pracujeme jako s inherentní součástí textu.

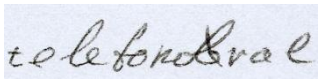
(25) Nečitelné úseky textu

dům o XXX XXX

8.3.5.2.4 Vliv jiných grafických systémů

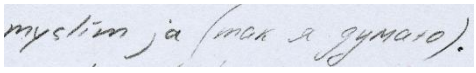
Pro potřeby přepisu textů nerodilých mluvčích se zaznamenávají i příp. užití jiných grafických systémů. Byly zvažovány tři alternativy evidence tohoto problému (viz příklad 26). První dvě testovaná řešení se ukázala jako technicky velmi náročná pro přepisovače, a proto jsme od nich ustoupili. Za prvé se uvažovalo o zaznamenávání problematických úseků textu odpovídajícím grafickým systémem (např. azbukou). Přepisovač by v tomto případě musel být schopen odhalit typ jazyka, resp. užitého grafického systému, což v případě, kdy nemá k dispozici metadatové informace, může být komplikované. Zároveň by musel disponovat různými druhy fontů, resp. sadami znaků různých druhů abeced. Druhým možným řešením byla transkripce do latinky. Předpokládat však, že anotátor ovládá různé typy písem a je zároveň schopen je adekvátně transkribovat se ukázalo jako nereálné.[194] Konečným řešením v případě, že text není zaznamenán latinkou, je volba znaku pro nečitelný text (XXX) a uvedení odpovídajícího kódu <gr>.

(26) Zaznamenání užití jiného grafického systému



a) telefono~~в~~val

myslím ja (так я думаю).



b) telefono~~ѵ~~val<gr>

myslím ja (tak ja dumaju)<gr>

c) telefono~~XXX~~val<gr>

myslím ja (XXX XXX XXX)<gr>

[194] Jedním z možných řešení by bylo využívat konkrétní přepisovače na konkrétní typy textů (resp. texty od respondentů s jedním mateřským jazykem). Toto řešení je však z hlediska celkového konceptu a zaměření projektu nepřijatelné.

8.4 Anotace korpusu nerodilých mluvčích češtiny

Jak vyplývá z analýzy světových žákovských korpusů, kterou představuji v kapitolách 5 a 6, přibližně polovina korpusů jazyka nerodilých mluvčích není chybově anotována (tj. 46 %, srov. odd. 5.1.3 v této práci, graf 4). Zároveň se řada anotací uplatňovaných v korpusech žákovského jazyka zaměřuje pouze na dílčí jazykové jevy, např. korpusy ISLE a LeaP se soustředí na problémy výslovnostní, TLEC na chyby ortografické, ALeSKO značkuje jevy z oblasti syntaxe a diskursu apod. Dále viz odd. 5.2, tab. 3, sl. 6 v této práci. Žákovský korpus češtiny se svou ambiciózní představou propojení manuální i automatické chybové anotace a anotace lingvistické řadí k malé skupině korpusů, které se pokoušejí o komplexní zachycení charakteru žákovského jazyka (jako např. ICLE, FRIDA, FALKO, NICT JLE).

Texty zařazené do korpusu nerodilých mluvčích češtiny budou značkovány gramaticky (slovnědruhově a morfosyntakticky) a lemmatizovány pomocí automatické procedury. Zároveň budou data manuálně značkována chybově.

Pro potřeby žákovského korpusu CzeSL byl vytvořen specifický anotační formát a systém chybové anotace.[195] Budování anotačního modelu pro žákovský korpus češtiny jako cizího jazyka zohledňovalo několik základních požadavků: (1) reflektovat charakter češtiny jako vysoce flektivního jazyka se specifickým slovosledem, (2) umožnit budoucí rozšíření chybové taxonomie, ať již toto rozšíření chápeme jako propojení s automatickou procedurou nebo jako zásah do podoby navržené chybové taxonomie, (3) vytvořit anotaci dostatečně podrobnou, ale zároveň zvládnutelnou pro anotátory. Při značkování korpusu CzeSL se totiž primárně nepočítá s malým, omezeným počtem úzce specializovaných anotátorů, jak je tomu v některých jiných světových korpusech (např. MELD, PiKUST apod.), ale s využitím a spoluprací většího množství anotátorů primárně z řad studentů bohemistiky.

8.4.1 Anotační formát

Pro žákovský korpus CzeSL byl vybudován třírovinný anotační formát, který operuje s jedinou hypotézou, ale aplikuje dvoustupňovou emendaci chyb. Z důvodů, které uvádím v oddíle 6.2.1 této práce, se při plánování chybového značkování rezignovalo na vkládanou lineární anotaci, kterou užívá většina chybově anotovaných žákovských korpusů (ICLE, JEFLL, NICT JLE, NOCE aj. Srov. kapitolu o typech anotací zde, odd. 6.2). V lineárním anotačním modelu totiž

[195] Koncept chybové anotace navrhl tým ÚTKL FF UK (včetně autorky této práce), anotační program FEAT vytvořil Jiří Hana. Dále k anotaci viz Hana, Rosen, Škodová, Štindlová (2010).

nelze zachytit některé dílčí informace o charakteru žákovského textu, které je žádoucí pro potřeby korpusu nerodilých mluvčích češtiny zachovat. V rámci jednorovinné anotace nelze (resp. lze jen s obtížemi) zaznamenat především postupné emendace a problémy nespojitých či překrývajících se řetězců.

Výraznou inspirací v úvahách o podobě anotačního formátu pro CzeSL byla podoba víceúrovňové distanční anotace německého žákovského korpusu FALKO (viz i zde, odd. 7.5), která ovšem pracuje s několikaúrovňovým typem anotace odlišným způsobem. Využívá možnost alternativních cílových hypotéz a rozšiřujícího počtu anotačních rovin, zároveň však zachycuje vztahy mezi elementy na jednotlivých anotačních rovinách pouze implicitně a korespondence mezi nimi může být při aplikaci tohoto typu anotace ztracena.

V modelu vytvořeném pro žákovský korpus češtiny jsou vztahy mezi elementy kódovány explicitně jako spoje mezi jednotlivými úrovněmi anotace. V prvním kroku, tj. při opravě z tzv. roviny 0 (R0), která reprezentuje transkribovanou podobu originálního textu bez chybových značek, na rovinu 1 (R1), se opravují chyby primárně ortografické a morfologické, resp. řeší se opravy izolovaných tvarů, neexistující formy se opravují na formy existující a výsledkem je řetěz existujících českých forem. Druhým krokem, tzn. při emendaci z R1 na rovinu 2 (R2), je oprava chyb, které plynou ze syntaktických vztahů, chyb lexikálních, v užití jazykového jevu, chyb slovosledných, stylových apod., jejímž výsledkem je gramaticky korektní věta.

Navržený anotační formát umožňuje identifikovat a opravit chyby v jednotlivých tvarech, ale také na větvících se řetězcích, které mohou být v některých případech nespojité. Emendace, tj. oprava textu, jak jsme již zmínili výše, probíhá na dvou rovinách a výhodou zvoleného anotačního schématu je možnost zřetelného značení vztahů mezi odpovídajícími si elementy jednotlivých rovin. Chybové značky jsou vkládány do uzlů na spojnicích mezi korespondujícími slovy nebo skupinami slov a od chybové značky je možno vést tzv. odkaz zaznamenávající důvod opravy.

8.4.2 Chybová taxonomie

Lze říci, že pro žákovský korpus češtiny je navržena středně hrubá taxonomie chyb.[196] Je definováno dvacet tři manuálních značek a devět chybových značek přiřazovaných automaticky. Do budoucnosti se počítá s rozšířením automatického chybového značkování, především při specifikaci ortografických chyb na první rovině emendace. Na rozdíl od chybové taxonomie,

[196] Toto tvrzení vychází ze současné podoby taxonomie, dá se však předpokládat, že ta se bude ještě částečně vyvíjet. K podrobným chybovým taxonomiím patří např. taxonomie FreeText, která zahrnuje přibližně 100 různých chybových značek, Díaz-Negrill*ová* a Fernández-Domínguez (2006: 89).

která se uplatňuje např. v korpusu CLC a NICT JLE, není chybová anotace pro CzeSL vystavěna na kategorizaci slovních druhů. Slovnědruhová klasifikace, obdobně jako morfosyntaktické značkování, bude v post-processingu zpracována automaticky.

Chybová taxonomie pro manuální značkování chyb v žákovském korpusu češtiny je hierarchická, založená na třístupňové klasifikaci chyby.[197] Srov. graf 8.

Graf 8: Zobrazení způsobu anotace v žákovském korpusu češtiny

Jazykové nedostatky jsou nejprve klasifikovány buď jako chyby individuálních slovních tvarů, které jsou, jak zmiňuji výše, emendovány a značkovány na R1, nebo jako chyby plynoucí z mezislovních vztahů a kontextového zapojení, které se emendují a značkují na R2.[198] Obě tyto široké klasifikační oblasti jsou dále zpřesňovány v doménách vymezených na základě lingvistických kategorií (příp. kategorií reflektujících povrchovou realizaci). Chyby v izolovaných slovních tvarech se člení na chyby v širokém slova smyslu ortografické, tj. v bázi nebo flexi daného tvaru *(incor)* a v hranicích slov *(wbd)*, a chyby v užití neexistujících výrazů *(fw)*, ať již se jedná o tzv. autorský neologismus, cizí výpůjčku nebo neidentifikovatelný řetězec. Chyby značkované na R2 se dělí na chyby syntaktické (tj. ve shodě *agr*, v široce chápané syntaktické závislosti vyjádřené jinak než shodou *dep* a v referenci *ref)*, chyby lexikální *(lex)*, chyby v uzuálním užití *(use)*, ve slovosledu *(wo)*. Dále jsou na této rovině značkovány chyby ve slovesných tvarech (resp. v analytických slovesných tvarech a složených predikátech, tj. *vbx)*,

[197] Přehled chybových kódů viz příloha 5.
[198] Některé chybné výrazy mohou být značkovány na obou rovinách při tzv. postupné emendaci/anotaci, např. *dvá další musea* – R1: $dva_{incorInfl}$ *další musea* – R2: $dvě_{agr}$ *další musea* nebo *nikdy ne pracovala* – R1: *Petr nikdy* $nepracovala_{wbdPre}$ - R2: *Petr nikdy* $nepracoval_{agr}$

včetně deagentivních, pasivních a rezultativních konstrukcí, chyby v reflexivních výrazech *(rflx)* a v negaci *(neg).* Zároveň je možné označkovat i zcela rozvrácenou, tedy neopravitelnou nebo velmi obtížně opravitelnou konstrukci *(disr)*. Specifickým typem značky, která slouží jako komplementární doplnění k již určenému typu chyby, je tzv. chyba sekundární *(sec),* jež signalizuje vynucenou opravu výrazů, které jsou v původním textu lokálně správně, emendovány jsou však v závislosti na opravě chybného tvaru vyskytujícího se v jejich okolí. Například v konstrukci *děti hledaly barevných vajec* je adjektivum ve jmenné frázi ve správném tvaru vzhledem k formě řídícího substantiva, toto substantivum je však v nesprávném pádu vzhledem k valenci slovesa. Při anotaci je nutné opravit a označit chybu ve slovesné rekci *(dep),* ale zároveň také „chybu" ve tvaru adjektiva, která je v tomto případě vynucená, tzv. sekundární. Specifickou sadou chybových kódů jsou značky, jež lze použít na obou rovinách anotace jako doplňkové k vyznačenému typu chyby. Jde o nedostatky stylové *(styl)* a o případy, kdy si anotátor není jist typem chyby, který vyznačil *(problem).*

Některé výše zmíněné domény se dále člení na dílčí subkategorie. Tyto subkategorie jsou také značkovány buď manuálně, nebo automaticky. Např. na rovině 1 jsou ortografické chyby podrobněji specifikovány na chyby ve slovotvorné struktuře a/nebo v derivační morfologii, tj. v kořeni a slovotvorných afixech *(incorStem)* a na chyby v tvarotvorných afixech *(incorInfl).* Na rovině 2 jsou například chyby ve slovesných tvarech dále zpřesňovány na chyby v analytickém slovesném tvaru *(cvf),* chyby v konstrukci s modálním nebo fázovým slovesem *(mod)* a chyby ve verbonominálních predikátech, opisném pasivu a rezultativních konstrukcích *(vnp).*

8.4.3 Automatické zpracování

Možnosti využití žákovského korpusu budou znásobeny doplněním jazykových informací pomocí automatického značkování. (Jelínek, 2010) Na obou anotačních rovinách bude provedena lemmatizace, slovnědruhové označení a značkování morfosyntaktických kategorií (na rovině 1 jako nejednoznačné, na rovině 2 diambiguované). Zároveň bude provedeno automatické doplnění chybové anotace, tedy značkování těch chyb, pro jejichž identifikaci není nezbytně nutný anotátor. Tyto chyby jsou v procesu manuální anotace záměrně ignorovány, což usnadňuje práci anotátorům a anotaci výrazně zrychluje. Automatické značkování bude zpřesňovat typ chyby na R1 srovnáním původního a opraveného řetězce, i na R2 porovnáním morfosyntaktických tagů. Na rovině 1 budou automaticky upřesněny např. chyby typu *incor,* u kterých bude doplněna informace, zda jde o chybu v diakritice, resp. délce vokálu, chybu v palatalizaci, znělosti, záměnu *i* a *y* apod. Automatická procedura na rovině 2 umožní detailnější

specifikaci manuálně vymezených typů chyb (např. u *vbx* či *rflx)* a zároveň doplní anotační značky u jevů, které byly emendovány, avšak záměrně neoznačeny (např. interpunkce). Součástí automatické chybové anotace bude i formální popis chyby, který je typickým atributem při charakterizaci chyb ve světových chybově značkovaných žákovských korpusech. Podle typu povrchové realizace[199] bude možné identifikovat chybu jako tzv. vynechání, přidání a chybný slovosled.

Vzhledem k tomu, že korpus češtiny nerodilých mluvčích je stále ve stavu budování, je pravděpodobné, že při aplikaci automatické anotace dojde následně ještě k rozšíření sady chybových značek, které budou pro anotaci chybných textů použity. Navržený anotační formát i chybová taxonomie to umožňují.

Kombinace manuální i automatické chybové anotace a automaticky vložených lingvistických informací, která je v kontextu existujících žákovských korpusů unikátní, společně v návaznosti na parametry reflektující mluvčího a text umožní podrobnou a funkční analýzu jazyka nerodilých mluvčích češtiny.

9 EVALUACE ANOTACE NAVRŽENÉ PRO ŽÁKOVSKÝ KORPUS ČEŠTINY

Manuální chybová anotace textů nerodilých mluvčích se musí vyrovnávat s jistým metodologickým rozporem.[200] Na jedné straně je vyžadována anotace, která povede k výzkumným výsledkům, jež budou nezávislé na anotátorovi dat a které budou replikovatelné. V takovém případě je anotátor chápán jako univerzální entita, kterou lze libovolně zaměnit a jejíž individualita by měla ovlivňovat podobu značkování co nejméně. Srov. např. Krippendorff (2004), Lombard et al. (2002), Hayes a Krippendorff (2007), kteří tvrdí, že i anotace speciálních korpusů (např. multimodálních, ale i chybových) by měla, resp. mohla být konzistentní a nevykazovat neshody, protože neshoda je znakem chybovosti ve značkování a omezuje možnosti dosáhnout při opakovaném výzkumu stejných výsledků. Na druhou stranu je manuální značkování korpusu vždy záležitostí rozhodování anotátora a v tomto smyslu je subjektivní záležitostí. Toto platí ještě výrazněji pro chybovou anotaci žákovských korpusů, kterou prozatím není možné opřít o automatické postupy detekce chyb nerodilých mluvčích. Viz i zde, oddíl 6.1.

[199] Ke kategorizaci chyb dle povrchové realizace viz v této práci odd. 1.6.3.3.1.
[200] Obdobně jako např. anotace multimodálních korpusů.

Obecně platí, že úspěšné využití anotovaného korpusu vyžaduje dostatečnou úroveň kvality značkování. V případě manuálně značkovaných korpusů to především znamená, že by anotátoři měli dobře chápat koncept zvoleného anotačního přístupu. Z toho důvodu je třeba soustředit se na dvě věci: za prvé musí být anotátoři dostatečně proškoleni, za druhé je nutné nepodcenit zpětnou vazbu a evaluovat zvolený anotační koncept, ať už na povrchové rovině instrukce (manuálu), či ve vnitřní struktuře.

Dále by anotátoři neměli být v rozporu se směrnicemi, které prezentuje anotační manuál. Tato podmínka nabývá na důležitosti především při práci s chybovým textem a s požadavkem rekonstruovat originální podobu jazykového projevu. Únava, neporozumění nebo prostě nepozorná práce způsobují, že anotátoři kromě chybné distribuce anotačních značek vnášejí do chybové anotace množství vlastních nedostatků a tím znehodnocují výsledky značkování. Srov. i Reidsma (2008: 2). Výše zmíněné jevy snižují úroveň značkování a omezují platnost závěrů plynoucích z odborných analýz.

Vyhodnocování kvality anotace se opírá o snahu kvantifikovat chyby, kterých se anotátoři dopouštějí. Jedna z metod této kvantifikace vychází z představy, že anotátoři anotující totožné penzum dat nedělají shodné chyby, resp. nechybují stejně. Srovnání několikanásobně anotovaných dat tedy prezentuje úroveň anotátorské shody *(inter-annotator agreement*[201]*,* IAA). V případě, že anotace, resp. anotátor vykazují vyšší procento neshody, mohou být označeni za nespolehlivé. Nespolehlivá anotace pak zakládá nevalidní a nereplikovatelné výsledky analýzy. Mezianotátorská shoda je proto velmi účinným nástrojem pro ověřování funkčnosti a platnosti anotačního schématu.

9.1 Mezianotátorská shoda

Od šedesátých let minulého století se objevila řada návrhů, jak měřit spolehlivost datových zdrojů, a stejně tak lze dohledat mnoho textů, které se zabývají odlišnostmi, podobnostmi, výhodami a nevýhodami jednotlivých měr. Podrobný přehled uvádí např. Lombard et al. (2002), Krippendorff (2004), Hayes a Krippendorff (2007), Artstein a Poesio (2008). V polovině devadesátých let se v lingvistice především na základě rozvoje sémantických analýz a výzkumů diskursu rozvinula odborná diskuse o vlivu subjektivity na budování značkovaných zdrojů dat. Jedním z hlavních témat této diskuse byla otázka míry spolehlivosti manuálně anotovaných korpusů a tedy i zvoleného anotačního schématu, resp. požadavek na vytvoření jednotného

[201] Někdy zmiňována jako *inter-rater reliability.*

konceptu pro hodnocení mezianotátorské shody v případě, kdy anotátoři aplikují na totožná data shodný kódovací manuál. Jeden ze základních způsobů, jak vypočítat míru mezianotátorské shody, představila poprvé v kontextu komputační a korpusové lingvistiky Carletta*ová* (1996), která adaptovala pro měření spolehlivosti tzv. koeficient shody κ.[202] Tento způsob měření spolehlivosti, nový v jazykovědných disciplínách, měl v té době již dlouhou tradici v kontextu jiných vědních oborů, srov. základní Cohenovu stať z roku 1960 (Cohen, 1960), ve které představuje koeficient shody κ pro statistickou analýzu v psychologii, Krippendorff (1980) se zabývá výpočtem míry shody v obsahové analýze, Fleiss[203] řeší shodu posuzovatelů v oblasti lékařství, resp. psychiatrii (viz též Grove et al., 1981[204]). Návrh Carletta*ové* byl velmi progresivní a koeficient κ se v komputační lingvistice stal standardním způsobem pro měření shody nejen v oblasti diskursivní analýzy, ale i v jiných odvětvích lingvistického výzkumu (např. Stevenson a Gaizauskas 2000, Mieskes*ová* a Strube 2006). Zároveň se také rozvinula debata o míře shody, kterou lze považovat za přijatelnou pro validitu výsledků výzkumů (viz Di Eugenio*vá* 2000 a dále zde tab. 6), o nevhodnosti používání κ pro měření některých typů shod (Poesio a Vieira 1998, Stevenson a Gaizauskas 2000, aj.) a také o vzájemném vztahu exitujících koeficientů shody, tj. Scottovo π, Cohenovo κ, jeho rozšířená projekce K Fleisse (resp. Siegela a Castelana) a Krippendorffovo α. Viz Scott 1955, Fleiss 1971, Siegel a Castelan 1988, Krippendorff 1980.

Mezianotátorská shoda kvantifikuje míru shody při nezávislém posuzování daného materiálu, jevu nebo subjektu dvěma nebo více hodnotiteli. V případě manuálně značkovaných korpusů přiřazování odpovídajících značek jednotlivým jevům znamená, že anotátor musí někdy vybírat mezi kategoriemi, které se vzájemně odlišují jen minimálně, případně je kategorizace jevu nejednoznačná a anotátor musí jev klasifikovat na základě osobní (i když často pravidly řízené) interpretace. Objektivita tohoto rozhodování je ověřována evaluací spolehlivosti značkování, resp. jde o to, zda anotátoři dosahují uspokojivé úrovně shody při práci na stejném úseku textu. Přepokládá se, že čím vyšší je shoda v hodnocení, tím vyšší je spolehlivost *(reliability)* dat, resp. jejich klasifikace (Krippendorff 1980: 147, Artstein a Poesio 2008: 557). Pokud se hodnotitelé (anotátoři) ve výsledcích shodují, lze říci, že obdobně chápou koncept a způsob anotace, obdobně rozumí anotátorskému manuálu a tím je primárně zaručena celková konzistence anotace. Ne tak jednoznačně vidí vztah shody a spolehlivosti korpusových dat Carletta*ová*

[202] K poznámkám o terminologických nesrovnalostech při vymezování κ, resp. π viz Artstein a Poesio (2008: 255).
[203] FLEISS, J. L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 1975, vol. 31, no. 3, s. 651–659. ISSN 0006-341X.
[204] GROVE,W. M., ANDREASEN, N. C., MCDONALD-SCOTT, P., KELLER, M. B., SHAPIRO, R. W. Reliability Studies of Psychiatric Diagnosis. Theory and Practice. *Archives of General Psychiatry*, 1981, vol. 38, s. 408–413.

(1996), Di Eugenio*vá* (2002), Reidsma a Carletta*ová* (2008), a navazují tak na vlastní Krippendorffovu poznámku (1980: 12), že nelze obecně stanovit standard pro přijatelnou míru shody, protože ten vyplývá z typu analýzy a ze sledování charakteru neshod, resp. z vyhodnocení jejich vlivu na využitelnost dat.

9.1.1 Koeficient kappa (κ)

Jak již bylo zmíněno výše, v současnosti se za standardní evaluaci mezianotátorské shody při kódování, tj. především v oblasti značkovaných korpusů, považuje kalkulace koeficientu κ. Tento koeficient budu aplikovat i na vzorek dat žákovského korpusu češtiny (viz oddíl 9.2.2). Dalšími koeficienty shody, které Carletta*ová* (1996: 252) považuje za varianty kappa koeficientu, se tato práce nezabývá. Výpočet κ je založen na celkovém poměru shody a očekávaném poměru shody. κ je možno vypočítat ze vztahu

$$\frac{P(o)-P(e)}{1-P(e)}, \quad \text{kde} \quad P(o) = \frac{1}{t}\sum t_a \text{ a} \qquad P(e) = \frac{1}{t^2}\sum t_x * t_y \,.$$

Ve vzorcích je *P(o)* tzv. zjištěná (pozorovaná) shoda mezi anotátory, *P(e)* je očekává shoda mezi anotátory, tj. hodnota hypotetické pravděpodobnosti náhodné shody. t_a je celkový počet shodně přiřazených značek, t_x je počet značek přiřazených prvním anotátorem, t_y je počet značek přiřazených druhým anotátorem, *t* je celkový počet značek.

Hodnoty proměnných *P(o)* a *P(e)* jsou vypočítány z kontingenční tabulky, jejíž obecnou podobu uvádíme v příkladu 5. Dva anotátoři klasifikují dané jevy (v našem případě chyby) přiřazováním chybových značek, což je v tabulce značeno indexy A – D. Indexy A a D reprezentují shodu anotátorů v distribuci chybových značek, B a C indikují neshodu v užití značky.[205]

Tabulka 6: Obecný příklad kontingenční tabulky pro výpočet koeficientu κ

	ANO$_{\text{Anotátor2}}$	**NE $_{\text{Anotátor2}}$**
ANO$_{\text{Anotátor1}}$	A	B
NE $_{\text{Anotátor1}}$	C	D

[205] Celkovou informaci o výpočtu κ koeficientu viz např. SIEGEL, S., CASTELLAN, N. J. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, 1988.

V případě, že je pozorovaná mezianotátorské shoda stejná jako očekávaná shoda, κ se rovná 0. Pokud je mezianotátorská shoda absolutní, κ se rovná 1. Pokud je mezianotátorská shoda horší, než indikuje náhodná shoda, kappa může být i záporné. V tabulce 6 uvádím škálu hodnot κ koeficientu, jak ji prezentují Krippendorff (1980) a Rietveld a Van Hout [206], kteří se ve své klasifikaci shodují s vymezením u Landise a Kocha (1977).

Tabulka 7: Škála hodnot κ koeficientu

Krippendorff		Rietveld a Van Hout	
KAPPA INDEX	MÍRA SHODY	KAPPA INDEX	MÍRA SHODY
0,8 ≤ κ ≤ 1	jednoznačné závěry	1 – 0,81	perfektní
0,67 ≤ κ < 0,8	provizorní závěry	0,6 ≤ κ < 0,8	podstatná
κ < 0,67	neprůkazné závěry	0,4 ≤ κ < 0,6	průměrná
		0,2 ≤ κ < 0,4	mírná
		κ < 0,2	nepatrná

Carlettaová (1996: 252) i Reidsma (2008: 30n.) tvrdí, že Krippendorffova hranice spolehlivosti κ=0,8, resp. 0,67, která je v rámci komputační lingvistiky často chápaná axiomaticky, je pro značkování některých jevů nevhodná, protože některé oblasti jsou pro manuální kódování mnohem obtížnější (diskurs, dialog, ale např. i chybový text) než původní obsahová analýza, pro kterou byla tato hranice modelována. Carlettaová dále uvažuje, že kromě evaluace anotátorů, anotačního konceptu i spolehlivosti anotovaných dat umožňuje κ index standardizaci srovnávání výsledků vzešlých z dat, na něž byla aplikována různá kódovací schémata. Reidsma a Carlettaová (2008) zároveň dokazují, že ani pro strojové učení a vývoj automatického značkování *(machine learning)* není hranice κ=0,8 klíčová, ale že lze za určitých podmínek tolerovat data s nižším kvocientem spolehlivosti. Tyto závěry jsou velmi inspirativní pro srovnávání IAA v případě značkovaných projevů nerodilých mluvčích, protože z výzkumů vyplývá, že koeficient shody v těchto případech zmíněného standardu 0,8 nedosahuje.

[206] RIETVELD, T., VAN HOUT, R. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, 1993.

9.1.2 IAA a anotace žákovského korpusu

Ověřování konzistentnosti, platnosti a funkčnosti chybové anotace není v rámci korpusů nerodilých mluvčích standardizováno. Současná anotační praxe světových žákovských korpusů je nechat chybové texty značkovat vždy pouze jedním anotátorem, a to i přes to, že nyní probíhá řada výzkumů automatické detekce chyb, zaměřených primárně na angličtinu jako cizí jazyk, které vyžadují vymezení spolehlivého a prověřeného zlatého standardu *(gold standard)*.[207] Srov. výzkumy zaměřené na automatickou detekci chyb ve členech a předložkách (Eeg-Oolofsson a Knutsson, 2004; Han et al., 2006; De Felice*ová* a Pulman, 2008; Gamon et al., 2008; Tetreault a Chodorow, 2008), japonské výzkumy zaměřující na analýzu chyb u počitatelných a nepočitatelných substantiv (Nagata et al., 2006) a širší detekci žákovských chyb (Izumi et al., 2003 a 2004), vývoj automatické identifikace syntaktických chyb (Dickinson a Meurers, 2005; Dickinson a Ragheb, 2009; Rosén*ová* a De Smedt, 2010), detekce chyb v kolokacích, členech a předložkách (Leackok*ová* et al., 2010), ve jmenném čísle a slovesném času, způsobu a vidu (Lee a Seneff*ová,* 2006).

Otázkou validity manuální chybové anotace značkované pouze jedním anotátorem, která má sloužit jako vzor pro trénování automatických aplikací, se poprvé podrobněji zabývali Tetreault a Chodorow (2008), již prezentovali svůj výzkum hodnocení rodilých mluvčích při klasifikaci užívání předložek. Při něm došli k závěru, že nejvyšší zjištěná shoda dvou anotátorů – rodilých mluvčích na značkování chyb v předložkách na stejném úseku textu má hodnotu kappa κ=0,63. Dále se k problému ověřování platnosti chybové anotace vyjadřuje i Meurers (2009), který kritizuje nedostatek studií zaměřujících se na výzkum IAA při manuální anotaci textů nerodilých mluvčích, což chápe jako velkou překážku při tvorbě automatických anotačních nástrojů. Viz i Rozovskaya a Roth (2010) a Polio*vé* (1997) souhrnnou studii mapující měření jazykové správnosti *(linguistic accuracy)* ve výzkumu cizojazyčného psaní.

9.1.2.1 Příkladové studie hodnocení IAA

Oprava chyb v textech nerodilých mluvčích je náročná a stanovení cílové hypotézy vyžaduje od anotátora mnoho obtížných rozhodnutí. Je zřejmé, že anotace žákovského korpusu je značně

[207] Zlatý standard je velmi kvalitně manuálně anotovaný vzorek dat, který obvykle slouží ke srovnání s automaticky anotovaným stejným vzorkem textu pro ověření správnosti automatického nástroje nebo k jeho trénování. Čím přesněji pak automatická anotace odpovídá zlatému standardu, tím je lepší. Pro vytvoření zlatého standardu obvykle slouží několikanásobné manuální značkování.

závislá na subjektivním přístupu. Tento fakt je podporován i skutečností, že rodilí mluvčí se výrazně rozcházejí v tom, co vlastně konstituuje přijatelnost v užití daného gramatického jevu či dané jazykové formy (srov. Tetreault a Chodorow, 2008). To je také pravděpodobně hlavní důvod, proč je k chybové anotaci standardně využíván jen jeden anotátor a proč je aplikace chybových anotací zřídka podrobována nějaké formě evaluace.[208]

Po podrobném šetření dostupných pramenů lze konstatovat, že jen několik studií[209] se zabývá problematikou validity chybové anotace, resp. mezianotátorskou shodou při značkování žákovského korpusu, viz i Meurers (2011).[210] Fitzpatrick*ová* a Seegmiller (2004) provedli několik výzkumů analyzujících konzistenci emendace chybových textů žákovského korpusu MELD. Nejedná se však o evaluaci chybové taxonomie, protože korpus MELD není chybově značkován (viz zde oddíl 7.3). Studii vztahu mezianotátorské shody při stanovení cílové hypotézy a kategorizace chyby prezentovala Lüdeling (2008). Rozvádí myšlenky Fitzpatrickové a Seegmillera (ibid.) a v dotazníkovém šetření zkoumá míru rozdílnosti cílových hypotéz u pěti různých anotátorů. Evaluaci anotačního schématu pro značkování ortografických chyb *(spelling errors)* v pilotním korpusu korejštiny jako cizího jazyka zmiňují Seok et al. (2009). Analýzu, která řeší problematiku IAA v chybově značkovaných žákovských korpusech, předložili i Rozovskaya a Roth (2010). Ti hodnotí mezianotátorskou shodu při manuální kategorizaci chyb na plně anotovaném žákovském korpusu sestaveném částečně z dat korpusu ICLE a částečně z dat korpusu CLEC, obsahujícím šedesát tři tisíc slov od respondentů na vyšší úrovni znalosti cílového jazyka. Tento korpus slouží autorům jako základ pro budování vlastního programu pro automatickou detekci chyb. Savkov a Becker*ová* (2009) ve svém projektu ProjectX, který se zaměřuje na vývoj automatického anotačního schématu pro značkování žákovských chyb založeného na XML, měří IAA na malém korpusu sestaveném z dat žákovského korpusu NOCE. Fitzpatrick*ová* a Seegmiller (ibid.) se zaměřili na srovnání mezianotátorské shody při evaluaci u tzv. expertů (tj. obou autorů) a u skupiny učitelů angličtiny jako cizího jazyka. K výpočtu IAA však nezvolili koeficient kappa, nýbrž tzv. koeficient spolehlivosti *(reliability),* který předpokládá chápání jednoho označkovaného textu jako zlatého standardu. Z toho důvodu jsou závěry této studie obtížně porovnatelné s novějšími analýzami, které pro IAA kalkulují κ. Přesto lze shrnout, že výsledky měření IAA nebyly v porovnání se standardem měřené spolehlivosti

[208] Je pravděpodobné, že validitu navrženého anotačního schématu interně ověřuje řada chybově značkovaných žákovských korpusů, závěry z těchto měření však nejsou veřejně dostupné.
[209] Máme k dispozici pouze pět studií zabývající se problematikou IAA při značkování jazyka nerodilých mluvčích.
[210] Hodnocení IAA v souvislosti se značkováním chybových textů uvádí ve své studii i Lu (2010). Jeho výzkum je zaměřen na detekci syntaktické složitosti psaných textů pokročilých čínských uživatelů angličtiny. Text se však přímo nezabývá chybovou anotací a IAA je vyhodnocena pomocí míry F-score, proto není tato studie v oddílu 9.1.2.1 zohledněna.

uspokojivé a pohybovaly se v rozmezí hodnot od 0,27 po 0,60. Řešení této situace uplatňované v žákovském korpusu MELD, jež bylo výrazně inspirováno přístupem Carlisleovým (viz Polio, 1997: 116), tedy zavést institut třetího anotátora jako konzultanta pro případy anotátorské neshody, je však značně časově i ekonomicky náročné a pro většinu žákovských korpusů obtížně realizovatelné.

Lüdeling*ová* (ibid.) částečně navazuje na Fitzpatrickovou a Seegmillera a zastává názor, že by cílová hypotéza měla být explicitní součástí chybové anotace, protože podle jejích výsledků v případě implicitní hypotézy narůstá rozrůzněnost ve značkování. Pokud však anotační schéma pracuje s explicitní cílovou hypotézou, je pravděpodobná vyšší mezianotátorská shoda při přiřazení chybové značky. Tento přístup předpokládá dvoustupňovou navazující anotaci, tj. emendaci a následné přiřazení tagu. Ve své studii neprovedla Lüdelingová žádné statistické měření mezianotátorské shody, není tedy možné její výsledky porovnat s ostatními analýzami.

Seok et al. (ibid.) měřili mezianotátorskou shodu mezi dvěma anotátory, kteří měli k dispozici pět značek pro kategorizaci v širokém slova smyslu ortografických chyb a anotovali pilotní korpus korejštiny jako cizího jazyka v rozsahu deseti textů od respondentů stejné úrovně znalosti jazyka (středně pokročilí). Za těchto podmínek bylo dosaženo mezianotátorské shody κ=0,83 v případě distribuce tagů a κ=0,73 v případě označení chybného, resp. správného výrazu.

Rozovskaya a Roth (ibid.) popisují značkování provedené na úrovni věty, ve kterém se primárně soustředí na chyby ve členech a předložkách, vedle toho však klasifikují i chyby ve jmenném čísle, slovesné formě, slovním tvaru a chyby ortografické. Ostatní chyby jsou jednotně řazeny do společné kategorie doplnění/vynechání/záměna výrazu. Na značkování se podíleli celkem tři anotátoři. Rozovskaya a Roth nechali opravit původní chybné věty jedním z anotátorů a tyto emendace (vždy celkem sto od jednoho anotátora) pak předložili k posouzení zbylým dvěma hodnotitelům.[211] Mezianotátorskou shodu, vyjádřenou koeficientem κ, autoři počítali na dvou obecných kategoriích ‘správně’ a ‘chybně’, které přiřadili jednotlivým větám podle toho, zda byly, či nebyly dané konstrukce opraveny ve druhém kole anotace. Mezianotátorská shoda se v jejich výzkumu pohybovala mezi 56% a 78% s relativně nízkou hodnotou kappy 0,16 – 0,40.

Savkov a Becker*ová* (ibid.) pro potřeby svého výzkumu upravili chybovou taxonomii korpusu FRIDA, především zredukovali počet dílčích kategorií v gramatické doméně a do svého konceptu nezahrnuli domény `registr` a `styl`. Anotaci vypracovali dva anotátoři. Výpočet IAA provedli Savkov a Becker*ová* na několika rovinách: binárně; pro jednotlivé úrovně anotace,

[211] K proceduře podrobně viz Rozovskaya a Roth (2010: odd. 6).

tj. úroveň chybových kategorií, subkategorií a slovních druhů; a pro všechny parametry souhrnně. Mezianotátorská shoda se v jejich výzkumu pohybovala v rozmezí $0,53<\kappa<0,66$.[212]

Z uvedených výsledků vyplývá, že ačkoli byl výzkum v případě Rozovskaye a Rotha zacílen pouze na omezený počet žákovských chyb, a v případě Savkova a Beckerové byla chybová taxonomie přizpůsobena anotátorům a výzkumnému cíli, index mezianotátorské shody ani v jenom případě nedosahuje Krippendorfova standardu $\kappa=0,67$, resp. $\kappa=0,8$. K obdobným výsledkům dospěli i Fitzpatrick*ová* a Seegmiller při svém měření spolehlivosti dvoustranné emendace a intuitivně jej potvrzuje i Lüdeling*ová.* Jistou výjimkou je analýza, kterou předkládá Seok et al., zde jsou však vyšší výsledky mezianotátorské shody dosaženy s největší pravděpodobností díky malému testovacímu vzorku a malému počtu distribuovaných tagů. Je tedy na místě položit si otázku, jakých hodnot IAA může být v kontextu chybové anotace dosaženo, resp. jaká hodnota kappa je relevantní pro jednotlivé aspekty chybové anotace (srov. Meurers, 2011). Tento problém si zasluhuje další podrobnější výzkum.

9.2 Mezianotátorská shoda pro anotaci korpusu CzeSL

9.2.1 Hypotéza

Jak jsem již zmínila výše, pro ověření kvality manuální anotace žákovského korpusu češtiny určíme míru mezianotátorské shody pomocí koeficientu κ.[213] Vybraný vzorek textů bude nezávisle anotován dvěma anotátory a budeme sledovat, zda se anotátoři ve svých analýzách shodují. Lüdeling*ová* (2008) uvádí, že i pro chybovou analýzu by měla být cílem vysoká IAA, a kritizuje, že tento požadavek není v prostředí žákovských korpusů běžně reflektován. V návaznosti na tento požadavek je však třeba upřesnit, co znamená vysoká mezianotátorská shoda při značkování textů nerodilých mluvčích, resp. jaká hodnota κ indexu, pokud volíme tento způsob měření IAA, je pro tento typ dat relevantní a dosažitelná. Tím se však již Lüdeling*ová* ve své stati nezabývá. Zmínila jsem již, že lingvistické analýzy mají tendenci respektovat jako prahovou hodnotu pro evaluaci anotačních schémat $\kappa=0,67$, resp. 0,8, kterou poprvé uvedl Krippendorff (1980). Paradoxně tak ale nereflektují Krippendorffovu připomínku, že není možné jednoznačně stanovit hranici dostatečné spolehlivosti (ibid., s. 146), která by byla obecně platná, a že kalkulovaná míra spolehlivosti anotace dat je vždy dána cíli, pro které mají být data využívána. Srov. i Passonneau*ová* et al. (2006).

[212] K proceduře podrobně viz Savkov a Becker*ová* (2009: odd. 4).
[213] Technické zpracování výpočtu indexu mezianotátorské shody provedl podle požadavků autorky Jiří Hana.

Měřením kappa indexu se chci v následujících oddílech pokusit ověřit validitu anotačního schématu navrženého pro žákovský korpus CzeSL, resp. prozkoumat, do jaké míry lze pro značkování chybových textů akceptovat Krippendorffovu hranici přijatelnosti κ=0,67. Zároveň přepokládám, že podrobnější interpretace výsledků měření IAA pomůže identifikovat problémové části pilotní verze manuálu pro chybovou anotaci a současně zprostředkuje odhalení možné nežádoucí systematičnosti v nesprávné distribuci tagů. Lze očekávat, že vyšší hodnotu κ indexu budou mít značky reflektující morfosyntaktickou strukturu jazyka, naopak je pravděpodobné, že chybové kategorie, které jsou svou povahou interpretativní, budou zdrojem anotátorské neshody.

9.2.2 Vzorek dat

Data pro anotaci byla vybrána z databanky shromažďované pro žákovský korpus CzeSL (viz zde kapitola 8). Do vzorku bylo zařazeno sedmdesát čtyři textů mluvčích s různým jazykovým pozadím, nejčastěji zastoupenými jsou mluvčí ruštiny. Každý text má průměrný rozsah 133 slov, nejmenší zařazený text obsahuje 49 slov, nejrozsáhlejší text má 448 slov. Vzorek obsahuje celkem 9848 slov. Texty zařazené do vzorku klasifikujeme podle škály SERR do úrovně A2 a B1. Dále srov. přílohu 6.

9.2.3 Metoda

Vzorek byl anotován celkem čtrnácti anotátory. Anotátoři jsou studenty druhých a třetích ročníků FF UK v Praze, obor: český jazyk a literatura (8) a TU v Liberci, obor filologie (6). Část anotátorů měla předchozí zkušenost s problematikou češtiny jako cizího jazyka, buď jako lektoři nebo jako studenti specializovaných programů učitelství češtiny jako cizího jazyka. Všichni anotátoři prošli vstupním proškolením, byli seznámeni s prostředím anotačního programu FEAT a s pilotní verzí manuálu pro chybovou anotaci. Každý text byl anotován vždy dvakrát, tj. dvěma nezávislými hodnotiteli. Každý z anotátorů anotoval v průměru 1475 slov a 11 textů. Srov. tabulku v příloze 6. Anotátoři byli pro potřeby výpočtu mezianotátorské shody rozděleni do dvou skupin, jež jsem souhrnně označila jako skupinu ANOTÁTOR A a skupinu ANOTÁTOR B. Tato označení používám v oddíle 9.2.4 při prezentaci výsledků kalkulace IAA.

Při srovnání mezianotátorské shody se distribuce chybových značek porovnává zvlášť na rovině 1 a na rovině 2. Na R1 měli anotátoři k dispozici deset tagů pro klasifikaci morfologických a ortografických chyb, chyb v hranicích slov, pro označení „nečeských“ výrazů a doplňkové tagy

pro značkování stylových nedostatků, viz tabulka 8. Syntaktické, morfosyntaktické, lexikální a stylové chyby na R2 značkovali anotátoři pomocí patnácti tagů, viz tabulka 10). Zároveň se počítaly výsledky mezianotátorské shody pro doménové tagy (`incoSum`, `wbdSum`, `fwSum` a `stylSum`) souhrnně, tj. bez rozlišení na jednotlivé chybové typy. Ke každému anotovanému textu z vybraného vzorku vyplňovali anotátoři tzv. anotační dotazníky (viz příloha 7), které slouží jako zpětná vazba pro autory manuálu a jako doplňkový zdroj informací při zhodnocení výsledků měření IAA.

9.2.4 Výsledky[214]

9.2.4.1 Anotační rovina 1

V tabulce 8 uvádím přehled distribuce chybových značek na rovině 1. První sloupec udává počet užití daného tagu anotátorem A v případě, že jej neužil anotátor B. Ve druhém sloupci daný tag přiřadil anotátor B, nikoli však anotátor A. Třetí sloupec zaznamenává shodu obou anotátorů v distribuci dané chybové značky.

Tabulka 8: IAA – distribuce tagů na R1

CHYBOVÁ KATEGORIE	ANOTÁTOR A	ANOTÁTOR B	ANOTÁTOR A, B - SHODA
incorSum	168	130	894
incorStem	167	165	559
incorInfl	173	130	250
wbdSum	14	21	45
wbd	26	9	5
wbdPre	12	14	3
wbdComp	5	27	8
fwSum	25	17	18
fw	4	6	1
fwFab	23	13	3
fwNc	10	9	3
stylSum	17	3	2
stylColl	10	2	2
stylOther	3	0	0

[214] Podrobné výsledky kalkulace koeficientu mezianotátorské shody κ viz příloha 8.

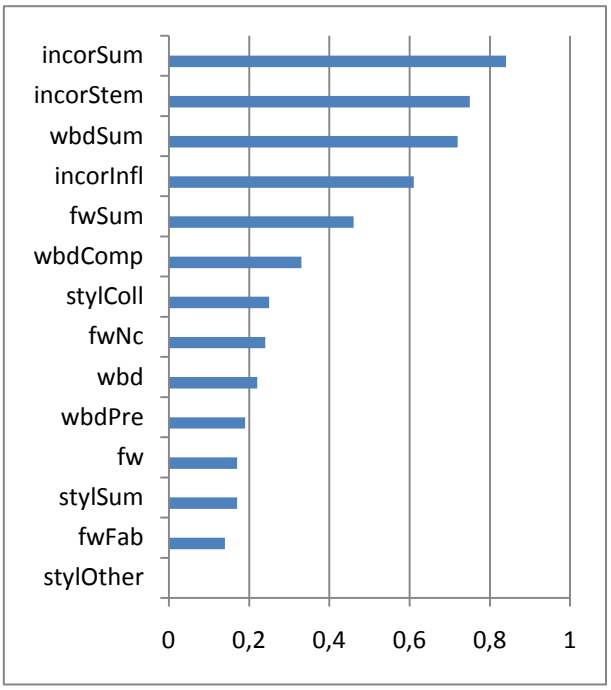
U některých značek je zřetelná asymetrie v distribuci u anotátora A a B, srov. `wbdComp` a `wbd`. Lze konstatovat, že v těchto případech je příčinou nejistota anotátorů v tom, jakou chybu značkovat jako `wbdComp` a které přiřadit bezpříznakovou značku `wbd`. Viz počty přiřazených značek anotátorem A a anotátorem B uvedené v tabulce: `wbdComp` 5 : 27, `wbd` 26 : 9. Příklady *ataké* nebo *mezi národní* značkoval anotátor A jako nesprávně rozdělené, resp. spojené kompozitum, kdežto anotátor B jako chybu v hranici slova bez podrobnějšího rozlišení. V tomto případě lze jako problém zakládající neshodu vymezit nepřesnost v anotačním schématu, resp. v chybové taxonomii, která nevede anotátora k jednoznačnému výběru tagu pro daný typ chyby. Dále z šetření mezianotátorské shody na R1 vyplývá, že minimální shoda panuje v přiřazování dílčích tagů z domén `fw` a `styl`. Tento výsledek byl očekávaný, protože jednoznačně souvisí s interpretací textu a postulováním cílové hypotézy.

Z následující tabulky (9) a grafu (9) vyplývá, že průměrnou a vyšší mezianotátorskou shodu, tj. κ>0,4, resp. κ>0,6, vykazují anotátoři při klasifikaci do doménových kategorií `incorSum`, `wbdSum` a `fwSum`. Stejně tak vykazuje vyšší shodu, tj. κ>0,6, značkování subkategorií chyb typu `incor`, tj. značkování chyb ve slovotvorné struktuře, derivační morfologii a tvarotvorných afixech. Nižší mezianotátorská shoda, tj. κ<0,4, se projevila při značkování interpretačně nejednoznačných kategorií s obtížnou emendací, jako jsou pokusy nerodilých mluvčích o vytvoření českého lexému (např. *mluvit bez přizůcku, můj synjačeky, programy nejlepšichanu, hodně jinaků* apod.) nebo užití cizích, příp. neidentifikovatelných řetězců (např. *priory u pero, jím rád eggs, jsem v truong).* Problematické bylo pro anotátory také vzájemné odlišení typů `fwFab` a `fwNc` (26 % z celkového neshodného užití těchto značek u anotátora A a B), a to především v případech, kdy anotátor nedokáže identifikovat původ daného lexému *(synjačeky, zeleněje).* Zcela zanedbatelné hodnoty mezianotátorské shody na R1, tj. κ<0,2, vykazuje doplňkové značkování stylově nevhodných výrazů. Z toho důvodu je na místě uvažovat, zda má značkování této chybové domény `(stylSum)` a jejích subkategorií na rovině 1 smysl, případně zda tuto doménu nezúžit pouze na příznak kolokviality.

Tabulka 9:
IAA – procentuální shoda a κ koeficient na rovině 1

CHYBOVÁ KATEGORIE	SHODA	KAPPA
incorSum	75 %	0,84
incorInfl	45 %	0,61
incorStem	63 %	0,75
wbdSum	56 %	0,72
wbd	12 %	0,22
wbdPre	10 %	0,19
wbdComp	20 %	0,33
fwSum	30 %	0,46
fw	9 %	0,17
fwFab	8 %	0,14
fwNc	14 %	0,24
stylSum	9 %	0,17
stylColl	14 %	0,25
stylOther	0 %	0

Graf 9:
Úspěšnost IAA podle κ koeficientu na rovině 1



9.2.4.2 Anotační rovina 2

V tabulce 10 uvádím přehled distribuce chybových značek na rovině 2. Stejně jako u tabulky 8 je v prvním sloupci uveden počet užití daného tagu anotátorem A v případě, že jej neužil anotátor B. Číslo ve druhém sloupci signalizuje, kolikrát daný tag přiřadil anotátor B, nikoli však anotátor A. Třetí sloupec zaznamenává shodu obou anotátorů v distribuci dané chybové značky.

Tabulka 10: Distribuce tagů na rovině 2

CHYBOVÁ KATEGORIE	ANOTÁTOR A	ANOTÁTOR B	ANOTÁTOR A, B - SHODA
agr	82	99	110
dep	99	118	87
ref	14	17	3
vbx	20	9	3
rflx	6	11	3
neg	11	9	9
oddObj	0	2	0
missObj	0	1	0
lex	107	131	74
use	60	74	19
sec	45	18	4
stylColl	14	14	10
stylOther	1	0	0
stylSum	19	14	10
disr	11	50	0
problem	8	10	0

Pro analýzu distribuce tagů pro chyby v povrchové realizaci (`oddObj` a `missObj`) a stejně i pro doplňkové značkování stylově nevhodného výrazu (jiného než obecněčeského) byl zvolený vzorek textů příliš malý, tagy `oddObj` a `missObj` byly použity pouze anotátorem B a to v prvním případě dvakrát a ve druhém případě jednou. Značku `stylOther` použil jednou anotátor A. Z toho důvodu je měření mezianotátorské shody u těchto kategorií nerelevantní a hodnota kappa indexu neprůkazná.

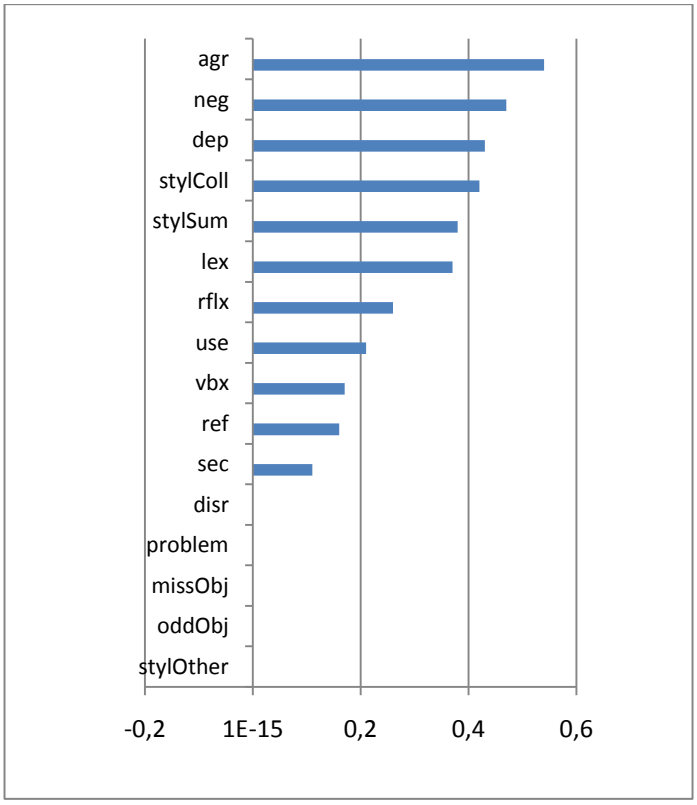
Očekávaně se anotátoři neshodli v přiřazování doplňkového tagu `problem`, který mohou distribuovat libovolně a který slouží pouze jako informace pro supervizora anotací o anotátorově nejistotě při přiřazení chybové značky. Výrazná neshoda v distribuci se projevila při značkování interpretačně náročné rozvrácené konstrukce (`disr`). Vzhledem k tomu, že jde o chybový typ, jehož značkování souvisí se subjektivním hodnocením anotátora, bylo lze předpokládat nízkou míru mezianotátorské shody. Z prezentované tabulky vyplývá, že shoda nenastala ani v jenom případě. Zajímavá je disproporce v přiřazování tohoto tagu u anotátora A a anotátora B. Po bližším prozkoumání anotovaného vzorku je možné konstatovat, že nadužívání značky `disr`

u anotátora B (11: 50) je způsobeno nevhodným způsobem zaznamenávání rozvrácených konstrukcí v anotačním programu FEAT, anotátor B značkuje pomocí tagu všechny dotčené individuální formy a neslučuje je do tzv. pavouka. Tím se nedrží instrukcí uvedených v manuálu (odd. 5.14). Například na úseku *oblékali se spolu protažni celého dnů*, který anotátor B považuje za nesrozumitelný, značkuje šest chyb `disr` namísto jedné, uvedené na sloučeném meziuzlu.

Tabulka 11
IAA – procentuální shoda a κ koeficient na rovině 2

CHYBOVÁ KATEGORIE	SHODA	KAPPA
agr	38 %	0,54
dep	27 %	0,43
ref	9 %	0,16
vbx	9%	0,17
rflx	15 %	0,26
neg	31 %	0,47
oddObj	0	0
missObj	0	0
lex	24 %	0,37
use	12 %	0,21
sec	7 %	0,11
stylColl	26 %	0,42
stylOther	0	0
stylSum	23 %	0,38
disr	0	0
problem	0	0

Graf 10
Úspěšnost IAA podle κ koeficientu na rovině 2



Z výše uvedené tabulky (11) a grafu (10) plyne, že průměrnou a vyšší mezianotátorskou shodu, tj. κ>0,4 vykazují anotátoři při klasifikaci syntaktických chyb narušení shody (`agr`) a chyb ve vyjádření syntaktické závislosti (`dep`), a také při značkování omezené a jednoznačně definované kategorie chyb v negaci (`neg`).

Velmi nízkou mezianotátorskou shodu, tj. κ<0,2 se vyznačuje značkování chyb v zájmenném odkazování (`ref`), v označování tzv. sekundárních, tj. vynucených chyb (`sec`) a

překvapivě také označování chyb v analytickém slovesném tvaru a složených predikátech (vbx), jež jsou identifikovatelné na základě formálních lingvistických kritérií. Ve všech třech případech se při podrobném rozboru charakteru distribuce tagů podpořeném přímým dotazováním u anotátorů (viz i anotační dotazníky v příloze 7) prokázala nedostatečnost anotačního manuálu, který anotátorům neposkytuje oporu pro jednoznačnou identifikaci zmíněných typů chyb, neilustruje dostatečně distinktivní rysy mezi charakterem chybového typu ref a agr, ref a dep (v obou případech tvoří vzájemná neshoda 19 % z celkového počtu neshodného užití značky ref), příp. také vbx a use, vbx a agr, vbx a dep, a zároveň nejasně specifikuje formální způsob značkování.

Mezianotátorská shoda při distribuci tagů označujících chyby lexikální a chyby v uzuálním užití se pohybuje v rozmezí $0,2<\kappa<0,4$. Obě tyto značky podstatně závisí na subjektivním posouzení anotátora, proto jsem přepokládala, že nelze očekávat vyšší hodnoty κ indexu. Výsledky $\kappa=0,37$ pro lex a $\kappa=0,21$ pro use toto očekávání potvrdily. Při analýze distribuce tagu lex se projevila systematičnost v jeho přiřazování: v případě, že jsou lexémy dostatečně významově rozdílné, anotátoři se na potřebě emendace a anotace ve většině případů shodnou. Pokud však jsou lexémy sémanticky blízké, nastává u anotátorů vysoká neshoda v názoru na potřebu emendace a tedy i následnou anotaci. Viz příklady (27) a (28).

(27) a. R0: *všichny rodiče chce když jejich dite bylo sčastne*
R2: An. A: ... *aby$_{lex}$ jejich dítě bylo šťastné*
An. B: ... *aby$_{lex}$ jejich dítě bylo šťastné*

b. R0: *Celou Cestu jsme slyšeli hudbu*
R2: An. A: *Celou cestu jsme poslouchali$_{lex}$ hudbu*
An. B: *Celou cestu jsme poslouchali$_{lex}$ hudbu*

c. R0: *mohou pojít na prochazku*
R2: An. A: *mohou jít$_{lex}$ na procházku*
An. B: *mohou jít$_{lex}$ na procházku*

(28) a. R0: *také kvůli její historii*
R2: An. A: *také kvůli její historii*

An. B: *také díky*$_{lex}$ *její historii*

b. R0: *když se divá na druhý kultury*
R2: An. A: *když se dívá na druhé*$_{agr+stylColl}$ *kultury*
An. B: *když se dívá na jiné*$_{lex+agr+stylColl}$ *kultury*

c. R0: *automaticky srovnám shody a neshody*
R2: An. A: *automaticky srovnám shody a neshody*
An. B: *automaticky srovnám shody a rozdíly*$_{lex}$

9.2.4.3 Problém neshodné emendace

Z analýzy značkovaných dat dále vyplývá, že neshoda v užívání chybových značek nemusí být nutně zapříčiněna anotátorskou chybou v jejich distribuci, ale může souviset s podobou zvoleného anotačního formátu. Srov. tabulku 12. Například ze 181 neshodných užití značky `agr` se 70 případů (tj. 39 %) liší na rovině 2 zároveň i v emendaci, viz př. (29).

(29) R0: *a kdyz stratil manzel*
R2: anotátor A *a když ztratí*$_{agr}$ *manžela*$_{dep}$
anotátor B *a když se ztratil manžel*

Obdobně např. i pro neshodu v užití značky pro chyby v lexiku, příklad (30).

(30) R0: *neměl s tím žádného setkání*
R2: anotátor A *neměl s tím žádné*$_{agr}$ *setkání*
anotátor B *neměl s tím žádné*$_{agr}$ *zkušenosti*$_{lex}$

Ze zbylých 111 případů neshodného užití značky pro narušení shody se jich dalších 28 (tj. 15 %) liší v emendaci již na rovině 1, viz příklad (31).

(31) R0: *tezki období*
anotátor A R1: *těžký*$_{incorStem+incoInfl}$ *období*
R2: *těžké*$_{agr+stylColl}$ *období*

anotátor B R1: *těžké* $_{incorStem+incoInfl}$ *období*

R2: *těžké období*

Ve všech uvedených případech je značkování v souvislosti se zvolenou emendací zcela v pořádku. V současnosti je zpracováván výzkum, který vychází z výše uvedeného zjištění a který by měl ověřit hypotézu o vlivu emendací na chybové značkování na jednotlivých rovinách. Již nyní však lze konstatovat, že požadavek Lüdelingové (Lüdeling, 2008) na explicitní vyjádření cílové interpretace v anotačním schématu je zcela opodstatněný a umožňuje reálně ověřit validitu tohoto schématu při kalkulaci mezianotátorské shody při distribuci tagů právě v závislosti na cílové hypotéze (srov. i Meurers, 2011).

Tabulka 12: Neshoda v emendaci

	Odlišná distribuce tagů	Odlišná emendace na R2		Odlišná emendace na R1		Odlišná emendace celkem
agr	181	70	39 %	28	15 %	54 %
dep	218	76	35 %	32	15 %	50 %
ref	32	7	22 %	3	9 %	31 %
vbx	30	17	57 %	6	20 %	77 %
rflx	18	12	67 %	0	0	67 %
neg	21	3	14 %	7	33 %	47 %
lex	239	147	62 %	10	4 %	66 %
use	135	66	49 %	10	7 %	56 %
sec	64	29	45 %	2	3 %	48 %

9.2.4.4 Srovnání vybraných skupin anotátorů

V následující tabulce srovnávám výsledky mezianotátorské shody mezi anotátory dvou skupin, jež se liší absolvovaným proškolením (srov. příloha 6). Anotátoři skupiny P prošli komplexním tréninkem,[215] včetně praktických cvičení, absolvovali testovací anotaci a obdrželi následnou zpětnou vazbu reflektující některé základní problémy, které se v této zkušební anotaci vyskytly. Všichni anotátoři této skupiny jsou až na jednu výjimku studenty druhého ročníku a pouze jeden z anotátorů skupiny P má zkušenost s problematikou češtiny pro cizince. Anotátoři skupiny L

[215] Problematice chybové anotace a školení anotátorů byl věnován celý semestrální kurz *Chybová anotace textů nerodilých mluvčích* v zimním semestru 2010/2011.

absolvovali intenzivní, ale krátkodobé výkladové proškolení[216] a nebyli prověřeni při zkušební anotaci. Všichni anotátoři této skupiny jsou studenty třetího ročníku a část z nich má zkušenost s výukou nerodilých mluvčích (celkem tři). Šest anotátorů skupiny P zpracovalo celkem 2056 slov, pět anotátorů skupiny L celkem 2430 slov.

Vzorek dat byl bohužel pro některé typy chyb příliš malý, proto pro ně nebylo možné spočítat hodnotu koeficientu shody κ (jde o tagy `fw`, `stylOther`, `oddObj`, `missObj`, `problem`). Tyto značky nejsou do následujících přehledů zařazeny.

Tabulka 13: Mezianotátorská shoda u více / méně proškolených anotátorů

Rovina 1

CHYBOVÁ KATEGORIE	SK. P	SK. L
	2056 slov	2430 slov
incorSum	0,88	0,76
incorInfl	0,67	0,42
incorStem	0,79	0,62
wbdSum	0,66	0,74
wbd	0,33	0
wbdPre	0	0
wbdComp	0,33	0,27
fwSum	0,25	0,34
fwFab	0,29	0
fwNc	0	0.44
stylSum	0	0
stylColl	0	0

Rovina 2

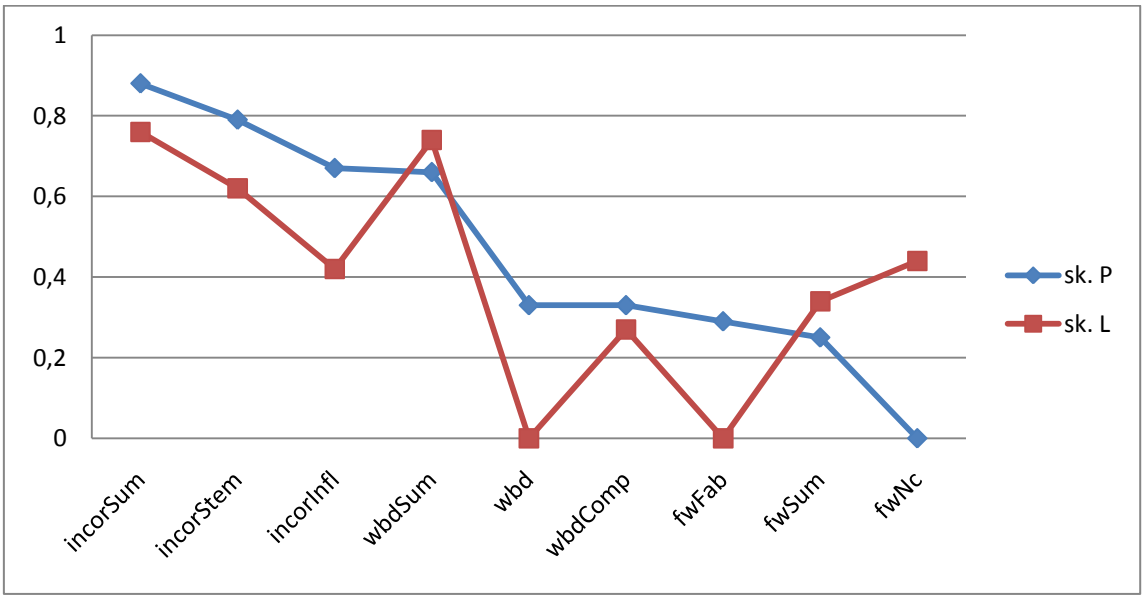
CHYBOVÁ KATEGORIE	SK. P	SK. L
	2056 slov	2430 slov
agr	0,61	0,35
dep	0,45	0,34
ref	0	0,2
vbx	0	0,18
rflx	0,5	1
neg	0,67	0
lex	0,35	0,17
use	0,04	0,08
sec	0,2	0
stylColl	0,35	0,29
stylSum	0,35	0,2
disr	0	0

Na obou rovinách je průkazná vyšší mezianotátorská shoda u anotační skupiny P. Viz grafy 11 a 12. Výjimkou na R1 je značkování chyb typu `wbdSum` (11:5:16[217] u sk. P, 6:4:14 u sk. L), v tomto případě vykazují obě skupiny obdobnou prostou, resp. pozorovatelnou shodu (*simple*,

[216] Celkem tři devadesátiminutové bloky.
[217] Zaznamenávám počet užití příslušného tagu: první pozice označuje počet užití pouze anotátorem A (zde 11), druhá pozice počet užití pouze anotátorem B (zde 5), třetí pozice vyjadřuje shodu obou anotátorů (zde 16).

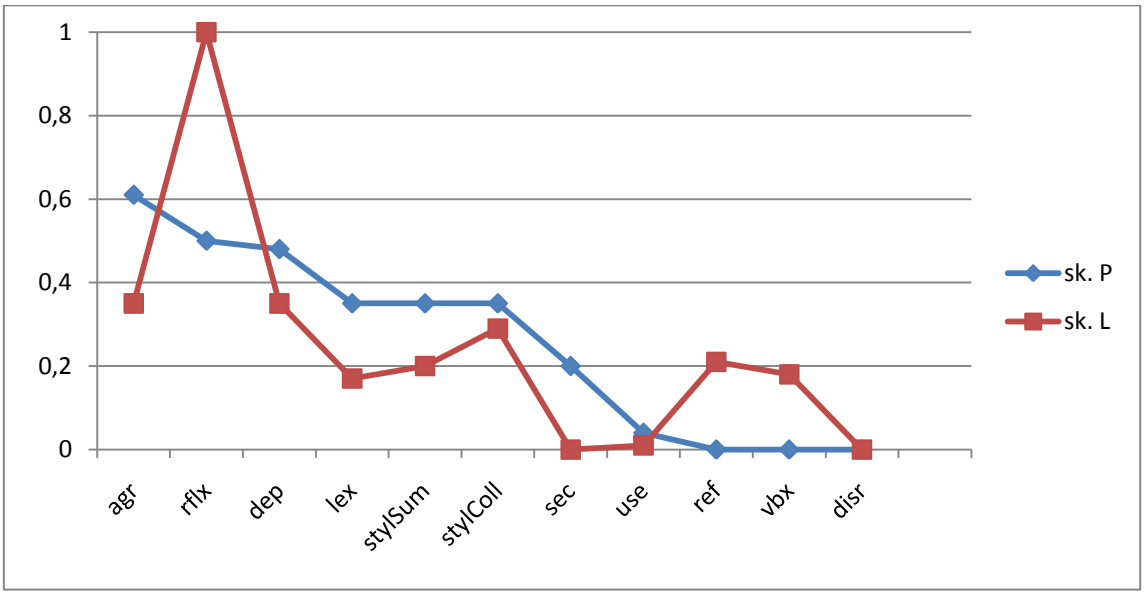
resp. *observed agreement*), tj. u skupiny P 50 %, u skupiny L 58 %, rozdíl v hodnotě κ je dán výraznějším rozdílem v distribuci wbd značek u skupiny P. Dále jde o chyby řadící se k souhrnnému typu `fwSum` (4:2:1 u sk. P, 2:17:5 u sk. L), kdy ve skupině P došlo ke shodě v užití pouze v jednom případě; a `fwNc` (0:1:0 u sk. P, 4:2:1 u sk. L), kde je nulová hodnota koeficientu κ u skupiny P dána téměř nulovým užitím dané chybové značky v daném vzorku textu.

Graf 11: Distribuce chybových značek na R1 u více / méně proškolených anotátorů



Na R2 vykazují vyšší hodnoty mezianotátorské shody u skupiny L pouze tři chybové typy: `rflx` (1:3:2 u sk. P, 0:0:1 u sk. L), kde je hodnota κ=1, tj. totální shoda značně zkreslená, protože ve skupině L byla daná značka přítomna právě jen v jediném případě; `vbx` (3:3:0 u sk. P, 1:8:1 u sk. L), kdy se anotátoři skupiny P neshodli při distribuci ani jednou a anotátoři skupiny L pouze v 10 % užití, tj. jednou; a `ref` (2:7:0 u sk. P, 11:4:2 u sk. L), v tomto případě je také viditelná diskrepance v používání značky u anotátora A a B u obou anotátorských skupin. Tj. u skupiny P užil anotátor A tag `ref` ve 22 % a anotátor B v 78 % z celkového počtu užití, ke shodě nedošlo ani v jednom případě; u skupiny L použil anotátor A tag `ref` v 65 % a anotátor B v 24 % z celkového počtu užití, shodli se dvakrát.

Graf 12: Distribuce chybových značek na R2 u více / méně proškolených anotátorů



V souvislosti se zjištěnými daty předpokládám, že odlišnosti ve výsledcích měření mezianotátorské shody u skupin P a L jsou značně závislé na způsobu proškolení anotátorů a vyšší úspěšnost skupiny P při kalkulaci mezianotátorské shody je dána délkou a mírou podrobnosti jejího předanotačního tréninku. Zároveň se domnívám, že velký vliv na podobu anotace, resp. na úroveň mezianotátorské shody má také míra zkušenosti anotátorů s jazykem nerodilých mluvčích. Tuto hypotézu je však třeba dále ověřit.

9.2.5 Závěry

Vyšší hodnota mezianotátorské shody a tedy ověření spolehlivosti anotačního schématu neznamená vždy výlučně také spolehlivou platnost výzkumných závěrů, srov. Krippendorff (1980: 213). Po výpočtu mezianotátorské shody by měly být výsledky měření interpretovány, protože teprve porozumění tomu, jak a proč se anotátoři při značkování neshodli (a také shodli) umožňuje objektivní využití anotovaných dat. Krippendorff (ibid.), v návaznosti na něj i Carletta*ová* (1996), Passonneau*ová* et al. (2006), Artstein a Poesio (2008) uvádějí, že je v procesu ověřování spolehlivosti anotace nutné analyzovat strukturu anotátorských neshod a uvažovat o tom, jak mohou případné pravidelnosti v neshodě ovlivnit spolehlivost dat. Odborné lingvistické studie, které jsem měla k dispozici, však k této interpretaci mezianotátorské neshody neposkytují žádné vhodné metodické instrukce, srov. i Reidsma (2008: 19).

Výzkumy v oblasti mezianotátorské shody prokázaly, že při manuálním značkování se (lidští) anotátoři nikdy kompletně vzájemně neshodnou v názorech na to, co a jak anotovat (tzv. *inter-annotator disagreement*), a zároveň, že v některých případech nejsou anotátoři ve značkování konzistentní ani jako individua (tzv. *intra-annotator disagreement*). Obecně lze říci, že většinu neshod mezi anotátory způsobují čtyři základní nedostatky: (A) za prvé jde obecně o nedostatky anotačního schématu, (B) druhým důvodem jsou procesní nedostatky anotace, (C) třetí příčina neshody tkví ve víceznačné povaze anotovaných dat a (D) čtvrtým důvodem je individualita anotátora, jeho schopnosti, znalosti a zkušenosti, zájem a motivace. K tomu viz i Krippendorff (1980), Weber (1983), Reidsma (2008), Zhang (2010) aj.

Pro potřeby validního výzkumu založeného na anotovaných datech je nutné si uvědomit, že některé typy neshod mezi anotátory jsou systematičtější než jiné. Tyto systematické neshody jsou pak větším problémem pro využití dat, než chyby 'náhodné' *(noise-like).*

9.2.5.1 Příčiny mezianotátorské neshody a doporučení k její minimalizaci

(A) 1. Nevhodný teoretický rámec pro anotační schéma

Výběr vhodného teoretického rámce je základním krokem při budování korpusu i při plánování konceptu anotace. V případě, že anotátor dostatečně nerozumí teoretickému pozadí zvolené anotace, nebo dokonce aplikuje odlišný teoretický model, budou výsledky značkování zavádějící. Domnívám se, že tento nedostatek se při analýze mezianotátorské shody pilotního značkování žákovského korpusu CzeSL neprojevil.

2. Nepřesné anotační schéma včetně nedostatků v chybové taxonomii

Teoretický základ anotace může být chybně zpracován v navrženém anotačním schématu. Takové schéma pak může obsahovat nerelevantní kategorie, některé potřebné kategorie v něm můžou chybět, a vymezení jednotlivých kategorií může anotátora vést k chybnému výběru. V případě zkušebního značkování vzorku dat z korpusu CzeSL se ukázalo takových nedostatků několik.

Na rovině 1: za neopodstatněné typy chyb, které by však podle instrukcí měly být na této rovině značkovány, považuji chyby typu `stylOther` a `sec`. Nedostatečně je diversifikována doména `wbd`, některé chyby (např. *atak, navýletu*) nelze jednoznačně zařadit ani k typu `wbdPre`, ani k typu `wbdComp`. V tomto případě doporučuji rozšířit podtypy domény `wbd`, a zároveň upřesnit instrukce v anotačním manuálu. Nejednoznačná charakteristika chybových kategorií vedoucí

k jejich záměně je jednoznačně prokazatelná u chybových typů `fwFab` a `fwNc`. Zároveň se nabízí k úvaze, zda nepřehodnotit oddělení pravopisných chyb, které se manuálně neznačkují, od chyb spadajících do domény `incor`. Anotátoři zde nemůžou spoléhat na jednoznačnou intuitivní hranici mezi oběma druhy chyb a z toho důvodu se pak v některých případech neshodují v distribuci značek domény `incor`.

Na rovině 2: výraznou mezianotátorskou neshodu vykazuje značkování interpretačních chybových domén `lex`, `use`, `disr` a `sec`. Domény `use` a `sec` vyžadují podrobnější a přesnější specifikaci, v případě domény `disr` je třeba přehodnotit její vymezení a způsob jejího formálního záznamu v anotačním programu. Domnívám se, že by bylo vhodné uvažovat o rozdělení chybové domény `lex` na dílčí podtypy, které by mohly reflektovat např. slovnědruhovou kategorizaci na autosémantika a synsémantika. Zároveň je třeba jednoznačně odlišit chybové typy, jejichž značkování není vždy intuitivní a jež jsou často v anotaci vzájemně zaměňovány, jako např. chybový typ `ref` od chyby typu `agr` a `dep`. Pro anotátory by toto odlišení mohlo mít např. podobu databanky problémových příkladů.

Za problematický aspekt navrženého anotačního schématu navrženého pro žákovský korpus nerodilých mluvčích češtiny považuji i ne zcela vyjasněný vztah manuální a automatické chybové anotace, který vyžaduje zpřesnění.

(B) 3. Nedostatečné proškolení anotátorů

V případě, že je nevhodně zpracován manuál pro anotaci, nebo když anotátoři nejsou dostatečně proškoleni, nemůžou na data určená ke značkování správně aplikovat anotační schéma. Možné odlišnosti v anotaci založené za charakteru školení pro anotátory uvádím v oddíle 9.2.4.4 této práce. Podstatnou otázkou v tomto kontextu zůstává, jaký je ideální rozsah a způsob anotátorského školení, abychom předešli problémům, které plynou z nedostatečného, či naopak z přílišného výcviku. Srov. i Hovy (2010: 71). Podotýkám také, že by pro potřeby značkování dat žákovského korpusu češtiny bylo vhodné prověřit, jak ovlivňuje spolehlivost anotace fakt, že anotátor má, příp. nemá zkušenosti s češtinou jako cizím jazykem.

Některé ze způsobů práce s anotátory by bylo vhodné aplikovat i na značkování žákovského korpusu češtiny. Jde především o nácvik anotace na adjudikatorním textu,[218] o vytvoření konzultačního anotátorského fóra, o ustavení tzv. supervizora anotace nebo-li superanotátora, který bude analyzovat a rozhodovat sporné případy, a o revizi anotačního manuálu, jež by v žádném případě neměl být zakonzervován příliš brzy (viz i Hovy, 2010: 61).

[218] Tento text má obdobnou funkci jako tzv. zlatý standard, slouží však anotátorů především pro nácvik anotace a následnou diskusi.

4. Tzv. administrativní chyby[219]

Tyto chyby jsou ovlivňovány především externími okolnostmi anotace. V případě testu anotace žákovského korpusu češtiny se týkaly především technických problémů s anotačním programem FEAT (např. viditelnost *ď* a *ť* na monitoru počítače), příp. nedostatku času na anotaci atd.

(C) 5. Víceznačné výrazy

Některé jazykové jevy či výrazy jsou víceznačné, a proto možné neshody zapříčiněné touto víceznačností nelze považovat za chybu. Podle zjištění, které vychází z anotovaného vzorku dat žákovského korpusu CzeSL, se však jedná pouze o jednotlivé případy.

(D) 6. Úroveň intersubjektivity

Některé anotace vyžadují od anotátorů značnou míru interpretace a tato interpretace se může anotátor od anotátora z mnoha důvodů lišit (věk, pohlaví, vzdělání apod.). Tato skutečnost pak vede k mezianotátorské neshodě při anotaci. Odborníci v oblasti anotací nezaujímají k tomuto problému jednotné stanovisko. Někteří subjektivní anotace ve výzkumu zcela odmítají, pro jiné to znamená, že typ anotace částečně založený na subjektivitě anotátora může vykazovat nízkou hodnotu koeficientu mezianotátorské shody.

Tento problém je velmi podstatný pro žákovský korpus češtiny, který pracuje s dvoustupňovou anotací, tj. emendací a přiřazováním chybové značky. V oddíle 9.2.4.3 jsem prokázala, že neshodná emendace zakládá neshodu ve značkování, tuto neshodu však nelze považovat za chybu či nedostatek anotačního schématu. Výhodou anotačního formátu korpusu CzeSL je přítomnost explicitní cílové hypotézy. Výpočty koeficientu mezianotátorské shody (v našem případě κ) by tento fakt měly zohledňovat. V případě, že bychom chtěli uvažovat o narovnání neshod v emendaci, nabízí se možnost definování striktnějších pravidel pro opravy na rovině 1, např. zrušení principu postupné emendace. Přísnější řízení emendací na rovině 2 by bylo komplikovanější.

[219] Termín viz Reidsma (2008).

ZÁVĚR

Předkládaná disertační práce si v úvodu stanovila několik výzkumných cílů: podat aktuální přehled současného stavu bádání v oblasti žákovských korpusů se zaměřením především na jejich chybovou anotaci; dále představit pravidla pro transkripci rukopisných textů, která byla navržena pro vznikající žákovský korpus češtiny jako druhého jazyka; posledním vytyčeným úkolem bylo prověřit navržený koncept anotačního schématu žákovského korpusu CzeSL, analyzovat problémy, které se při zkušební anotaci projeví, a navrhnout možnosti řešení. Dané cíle se podařilo téměř bezezbytku naplnit, řada dílčích úkolů však zůstává k dalšímu řešení.

Kapitoly 2, 3 a 4 mapují problematiku výstavby žákovského korpusu a konfrontují specifické rysy tohoto typu korpusu se standardní podobou korpusu národního.

V kapitole 5 je charakterizován reprezentativní vzorek 57 žákovských korpusů a souhrnně je prezentována aktuální analýza dat, která staví na předcházejících výzkumech jiných odborníků, ale také na dotazníkovém šetření, navrženém konkrétně pro tuto práci. Na základě analýzy získaných dat uvádíme, že 65 % světových žákovských korpusů se zaměřuje na angličtinu jako cizí jazyk, 30 % korpusů se věnuje jiným cílovým jazyků a zbylé korpusy jsou multilingvální. Celkem 56 % žákovských korpusů shromažďuje texty od respondentů na pokročilé a středně pokročilé úrovni znalosti cílového jazyka, 35 % korpusů sbírá texty od žáků na všech úrovních znalosti a ostatní korpusy se zaměřují na začátečníky, mírně pokročilé, či děti. Poměr chybově anotovaných a neanotovaných žákovských korpusů je vyrovnaný, chybovou anotaci, ať již komplexní, nebo částečnou obsahuje 45 % korpusů, 46 % korpusů anotaci nemá. Žákovské korpusy v 74 % sbírají psané materiály, 17 % korpusů se soustředí na mluvené projevy, 7 % korpusů sbírá oba typy komunikátů a 2 % jsou multimediální. Nejvíc žákovských korpusů dosahuje velikosti od 100 tisíc po 1 milion slov (celkem 39 %). Nad 2 miliony slov má 18 % žákovských korpusů.

Kapitola 7 rozebírá charakter chybové anotace a na vybraných korpusech demonstruje různá anotační schémata chybové taxonomie. Na základě zjištění prezentovaných v této kapitole lze tvrdit, že žákovské korpusy aplikují lineární anotační formát s vkládanou anotací, některé korpusy zároveň reflektují i potřebu explicitní anotace. V případě jednorovinného formátu je však obtížně řešitelné např. zaznamenávání chyb na nespojitých řetězcích, chyb ve slovosledu, příp. zaznamenávání odlišných cílových hypotéz. Z toho důvodu německý žákovský korpus FALKO jako první použil vícerovinnou distanční anotaci, které umožňuje snazší řešení

uvedených problémů. Vícerovinnou anotaci v současné době testuje více korpusů a jedním z nich je i korpus češtiny nerodilých mluvčích CzeSL.
Komplexní přehled uvedený ve zmíněných kapitolách podává jasnou představu o charakteru světových žákovských korpusů. Bude-li v průběhu dalšího výzkumu rozšířen o další informace týkající se nově vznikajících projektů, stane se velmi cenným východiskem pro budoucí analýzy dané problematiky.

Kapitola 8 se zaměřuje na žákovský korpus CzeSL a představuje problematiku sběru a anotace dat. Podrobně se pak soustředí na otázku přepisu rukou psaných textů a představuje přepisovací pravidla navržená speciálně pro potřeby přepisu textů nerodilých mluvčích, resp. v modifikované podobě pro potřeby akvizičních korpusů skupiny AKCES. Tato podrobná přepisovací pravidla pro rukopisy jsou v českém kontextu novátorská a mohou sloužit jako vzor pro další projekty, které musí řešit digitalizaci rukou psaných jazykových projevů.

Kapitola 9 se věnuje evaluaci anotačního schématu navrženého pro žákovský korpus češtiny. Jako evaluační metoda byl vybrán výpočet míry mezianotátorské shody (κ), který je v komputační lingvistice od 90. let minulého století chápán jako standardní prostředek pro ověřování platnosti anotací. Při analýze testovacího anotovaného souboru dat, který obsahoval téměř 10000 slov, se projevily některé dílčí nedostatky aplikovaného anotačního schématu, které vyplývají z nedokonalé formulace pravidel anotace uvedených v anotačním manuálu a také z neadekvátního proškolení anotátorů, resp. z jejich předpokladů k velmi specifickému a náročnému úkolu.. Většina mezianotátorských neshod však pramenila především ze dvou zdrojů – z odlišné emendace na rovině 1 i na rovině 2, která přirozeně zakládá i rozdílnou anotaci, a z očekávané neshody na interpretaci některých chyb, především takových, které patří do interpretačních chybových kategorií (jako je např. `lex`, `use`, `fwFab` apod.).
Při evaluaci se ukázalo, že zvolený vzorek není dostatečně velký pro hodnocení shody u některých typů chyb (např. `stylOther`, `fwNc`, `oddObj`, `missObj`), proto nebyly tyto chyby do kalkulace mezianotátorské shody zařazeny. Z toho důvodu by bylo vhodné rozšířit zkušební anotační vzorek a znovu jej otestovat. K tomu je třeba upravit anotační manuál (některá doporučení jsou uvedena v oddíle 9.2.5) a adekvátně proškolit anotátory. Protože jistou nevýhodou výpočtu spolehlivosti pomocí koeficientu κ je, že měří stupeň spolehlivosti pouze globálně, ale nevypovídá nic o riziku záměny v klasifikaci, měla by být dalším krokem výzkumu kalkulace hodnoty oddsRatio pro porovnání záměnného užívání některých chybových značek (např. `dep` a `agr`, `lex` a `use` apod.). A konečně by také měl být dokončen výzkum, který

přinese podrobnější informace o vlivu cílové hypotézy na distribuci chybový značek. Pak bychom pravděpodobně mohli odpovědět na otázku, jaká je přijatelná hodnota koeficientu κ při výpočtu mezianotátorské shody chybové anotace.

PŘÍLOHY

PŘÍLOHA 1

Průzkumný dotazník pro vytvoření přehledu světových žákovských korpusů

LEARNER CORPORA SURVEY	
1. Name of the learner corpus:	
2. What size is your learner corpus at present?	
3. What kind of annotation is applied to your corpus?	
➢ POS	for the text as a whole only for errorneous expressions
➢ linguistic	for the text as a whole only for errorneous expressions
➢ error	for the text as a whole only for specific features
➢ none	
4. How large part of the corpus is annotated?	
➢ for POS	
➢ linguistically	morfology parsing lemma other
➢ for errors	
➢ none	
5. Have you build an error taxonomy? If yes, what kind of taxonomy is it?	
A. Taxonomy type:	**B. Taxonomy based on:**
➢ partial: intent on you research questions (which is?)	linguistic cathegories
➢ systematic: general	target modifications

➢ other	communicative competence
➢ none	other
6. How many taggs has your error taxonomy?	
7. What kind of procedure do you use for annotating?	
➢ manual	
➢ authomatic	
➢ combination of both	
➢ none	
8. What is the proficiency level of the authors of the samples inserted in the corpus?	
➢ advanced (C1-C2)	
➢ intermediate (B1-B2)	
➢ beginner (A1-A2)	
➢ various / all	
9. Why have you decided to gather data from this particular proficiency level // from the all levels of knowledge?	
10. What size have to be a sample you insert into the corpus (do you have a limit)?	

PŘÍLOHA 2

Ukázky vyplněného průzkumného dotazníku

A. Korpus CALES

LEARNER CORPORA SURVEY		
11. Name of the learner corpus:	**CALES – Corpus Archive of Learner English in Sabah-Sarawak**	
12. What size is your learner corpus at present?	**480,000+**	
13. What kind of annotation is applied to your corpus?		
➢ POS	for the text as a whole only for errorneous expressions	
➢ linguistic	for the text as a whole only for errorneous expressions	
➢ error	for the text as a whole only for specific features	**YES** **YES**
➢ none		
14. How large part of the corpus is annotated?		
➢ for POS		
➢ linguistically	morphology parsing lemma other	
➢ for errors	**Approximately 9%**	
➢ none		
15. Have you build an error taxonomy? If yes, what kind of taxonomy is it? I am using Granger's UCLEE taxonomy (Dagneaux et al, 1998)		
A. Taxonomy type:	**B. Taxonomy based on:**	
➢ partial: intent on you research questions (which is?)	linguistic cathegories	**YES**
➢ systematic: general **YES**	target modifications	**YES**

➢ other	communicative competence
➢ none	other
16. How many taggs has your error taxonomy?	**40**
17. What kind of procedure do you use for annotating?	
➢ manual	
➢ authomatic	
➢ combination of both **YES – using specialised editing software**	
➢ none	
18. What is the proficiency level of the authors of the samples inserted in the corpus?	
➢ advanced (C1-C2)	
➢ intermediate (B1-B2) **YES**	
➢ beginner (A1-A2)	
➢ various / all	
19. Why have you decided to gather data from this particular proficiency level // from the all levels of knowledge?	Because this reflects the students in the institution where I teach (in Sarawak state, East Malaysia)
20. What size have to be a sample you insert into the corpus (do you have a limit)?	No limit

B. Korpus FLLOC

LEARNER CORPORA SURVEY		
21. Name of the learner corpus:	French Learner Language Oral Corpora (FLLOC), available at www.flloc.soton.ac.uk	
22. What size is your learner corpus at present?	2+ million words	
23. What kind of annotation is applied to your corpus?		
➢ POS	✓for the text as a whole	
➢ linguistic	×for the text as a whole ×only for errorneous expressions	
➢ error	×for the text as a whole ×only for specific features	
➢ none		
24. How large part of the corpus is annotated?		
➢ for POS ✓ All of it		
➢ linguistically	×morfology ×parsing ×lemma ×other	
➢ for errors ×		
➢ none ×		
25. Have you build an error taxonomy? NO **26. If yes, what kind of taxonomy is it?**		
A. Taxonomy type:	**B. Taxonomy based on:**	
➢ partial: intent on you research questions (which is?)	linguistic cathegories	
➢ systematic: general	target modifications	
➢ other	communicative competence	

➢ none	other
27. How many taggs has your error taxonomy?	NOT APPLICABLE
28. What kind of procedure do you use for annotating?	
➢ manual	
➢ authomatic	
➢ combination of both ✓ WE USE THE PROCEDURES OF CHILDES/ CLAN TO TAG DATA SEMI-AUTOMATICALLY	
➢ none	
29. What is the proficiency level of the authors of the samples inserted in the corpus?	
➢ advanced (C1-C2) ✓	
➢ intermediate (B1-B2) ✓	
➢ beginner (A1-A2) ✓	
➢ various / all ✓	
30. Why have you decided to gather data from this particular proficiency level // from the all levels of knowledge?	WE ARE INTERESTED IN MORPHOSYNTACTIC DEVELOPMENT OF L2 FRENCH LEARNERS FROM BEGINNERS TO ADVANCED
31. What size have to be a sample you insert into the corpus (do you have a limit)?	SAMPLES VARY IN SIZE BUT MOST ARE TRANSCRIPTIONS OF SEVERAL MINUTES OF SPOKEN FRENCH

PŘÍLOHA 3

Příklady současných studií vycházejících ze žákovského korpusu[220]

V lexikální oblasti proběhl např. výzkum slov s vysokou frekvencí užívání (Ringbom, 1998), který prokázal, že pokročilí studenti cizího jazyka ve srovnání s rodilými mluvčími nadužívají vysoce frekventovaných slov (např. výraz *think* užívají nerodilí mluvčí pětkrát více než rodilí). Z analýzy intenzifikace adjektiv (Lorenz, 1998) vyplývá, že studenti cizího jazyka intenzifikující adjektiva znějí rodilým mluvčím nepřirozeně proto, že často intenzifikují adjektiva v tematické pozici na rozdíl od rodilých mluvčích, kteří intenzifikují adjektiva v pozici rematické. Analýza frazikonu (de Cock et al., 1998) vyvrátila hypotézu, že nerodilí mluvčí pracují spíš se slovy, než s prefabrikovanými frázemi. Podle tohoto výzkumu se cizinci opírají o prefabrikáty kvantitativně více než rodilí mluvčí, mají jich však k dispozici jen omezené množství, které nadužívají.

Další výzkumy založené na žákovských korpusech se zaměřovaly např. na frázová slovesa (Waibel, 2008), vid (Rogatcheva, 2009), používání časů (Granger, 1999), výběr prepozic v jazyce nerodilých mluvčích (Barlow – Smith, 2005), na funkci ukazovacího zájmena jako anaforického ukazatele (Leńko-Szymańska, 2004), používání adverbiálních konektorů

[220] Aijmer K. *I think* as a marker of discourse style in argumentative Swedish student writing. In *A Wealth of English. Studies in Honour of Göran Kjellmer,* ed. K Aijmer. Göteborg: Acta Universitatis Gothoburgensis. 2005, s. 247-257.

Altenberg, B., Tapper, M. The use of adverbial connectors in advanced Swedissh learnerers' written English. In *Learner English on Computer,* ed. S. Granger. London: Longman, 1998, s. 80-93.

Barlow M. – Smith S. *Analysing preposition choice in learner English*. Paper presented at the workshop "Linking Up Contrastive and Learner Corpus Research", University of Santiago de Compostela, 2005.

Belz, J., Vyatkina, N. 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.* 62.1, 2005, s. 17-48.

de Cock, S., Granger, S., Leech, G., McEnery, T. An automated approach to the phrasicon of EFL learners. In *Learner English on Computer,* ed. S. Granger. London: Longman, 1998, s. 67-79.

Granger, S. Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In *Out of corpora. Studies in honour of Stig Johansson*, eds. H. Hasselgård, S. Oksefjell. Amsterdam: Rodopi. 1999, 191-202.

Leńko-Szymańska A. Demonstratives as anaphora markers in advanced learners' English. In *Corpora and Language Learners,* eds. G. Aston, S. Bernardini, D. Stewart. Amsterdam: Benjamins. 2004, 89-107.

Lorenz, G. R. *Adjective Intensification - Learners versus Native Speakers. A Corpus Study of Argumentative Writing.* Amsterdam: Rodopi, 1999.

Rogatcheva, S. „I've only found the answer a few days ago": aspect use in Bulgarian and German EFL writing. In *New Trends and Methodologies in Applied English Language Research. Diachronic, Diatopic and Contrastive Studies,* eds. C. Prado-Alonso, L. Gómez-García, I. Pastor-Gómez, D. Tizón-Couto, Frankfurt: Peter Lang, 2009, s. 255-278.

Ringbom, H. Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In *Learner English on Computer,* ed. S. Granger. Harlow: Longman. 1998, s. 41-52.

Waibel B. *Phrasal verbs. German and Italian learners of English compared*. Saarbrücken: VDM, 2008.

(Altenberg – Tapper, 1998), analýzu diskurzu v textech pokročilých studentů (Aijmer, 2001), pragmatickou kompetenci nerodilých mluvčích (Belz-Vyatkina, 2005).

Zkoumání jazyka nerodilých mluvčích opírající se o korpusová data je stále teprve v počátcích. Existující studie se primárně zaměřují na jednotlivosti a zřídka dovozují obecné závěry zapojené do širšího kontextu. Zcela jistě však můžeme předpokládat, že analýzy žákovských korpusů významně podpoří výzkum akvizice cizího jazyka, včetně popisu mezijazyka.

Pedagogické aplikace žákovských korpusů

Vliv žákovských korpusů na pedagogickou praxi a jejich uplatnění při tvorbě didaktických materiálů nejsou prozatím příliš velké. V zásadě lze říci, že v současné době jsou žákovské korpusy systematicky využívány pouze pro tvorbu výkladových slovníků (viz např. *Logman Dictionary of Contemporary English* (2003), *Macmillan English Dictionary for Advanced Learners* (2007) a *Cambridge Advanced Learner´s Dictionary* (2003)). Tyto slovníky jsou vystavěny na analýzách žákovského jazyka v komerčních žákovských korpusech CLC a LLC a v návaznosti na chybovou analýzu zahrnují také sekce *common learner error*, příp. *warning notes* či *help-boxes.* Druhou oblastí, která s žákovskými korpusy aktivně pracuje, jsou elektronické didaktické nástroje, tzv. CALL programy[221]. Jde například o nástroj zaměřený na problémy psaní (WordPilot, Milton, 1998), program pro explanované korekce gramatických chyb (ESL Tutor, Cowan et al., 2003) nebo tréninkový program pro učitele cizího jazyka (TeleNex, Allan, 2002).

Dále viz i Štindlová (2011).

[221] CALL – computer assisted language learner

PŘÍLOHA 4

A. Anamnestický dotazník o autorovi textu

K textům

Text číslo:	Text číslo:	Text číslo:	Text číslo:

V následující tabulce doplňte nebo vyberte z nabízených charakteristik tu informaci, která se týká autora textu.

<table>
<tr><td colspan="5">POVINNÉ POLOŽKY (prosím vyplňte):</td></tr>
<tr><td>VĚK</td><td colspan="2">…….</td><td>POHLAVÍ</td><td>muž žena</td></tr>
<tr><td>PRVNÍ JAZYK</td><td colspan="4">vietnamština angličtina jiný:
ruština němčina
ukrajinština polština …….</td></tr>
<tr><td>STUDIUM ČJ
KDE</td><td colspan="4">základní nebo střední škola v ČR
kurz na vysoké škole v ČR (včetně přípravných)
kurz v komerční jazykové škole v ČR
jiný typ organizovaného kurzu v ČR
JAKÝ? …….
jazykový kurz mimo ČR
JAKÝ? …….
samostudium
individuální lekce (one-to-one)
jiné: …….</td></tr>
<tr><td></td><td colspan="4">…….</td></tr>
<tr><td>JAK DLOUHO

POČET HODIN TÝDNĚ</td><td colspan="4">do 3 hodin týdně do 15 hodin týdně více než 15 hodin týdně
komentář: …….</td></tr>
<tr><td>POUŽÍVANÉ UČEBNICE</td><td colspan="4">Communicative Czech
New Czech Step by Step
Easy Czech Elementary
Chcete mluvit česky? (Do you want to speak Czech? / Chotite govorit´po-češsky?)
Čeština pro cizince
Czech for Foreigners
Čeština pro malé cizince
Jiná: ……</td></tr>
<tr><td>ÚROVEŇ ZNALOSTI ČJ (PODLE SERR)</td><td>A1
A1+</td><td>A2
A2+</td><td>B1
B2</td><td>C1
C2</td></tr>
<tr><td colspan="5">FAKULTATIVNÍ POLOŽKY (prosím vyplňte, pokud je tato informace k dispozici):</td></tr>
<tr><td>ZNALOST DALŠÍCH JAZYKŮ</td><td colspan="4">angličtina ruština jiný:
němčina polština
francouzština španělština …….</td></tr>
<tr><td>BILINGVNÍ</td><td colspan="2">ano ne
jaké jazyky: ……</td><td>DÉLKA POBYTU NA ÚZEMÍ ČR</td><td>kratší než 1 rok
1 – 2 roky
více než 2 roky</td></tr>
<tr><td>V RODINĚ UMÍ ČESKY</td><td colspan="4">nikdo z rodiny oba rodiče bratr / sestra
otec matka partner / partnerka</td></tr>
</table>

B. Průvodka k nasbíranému materiálu

TEXT ČÍSLO:	**DATUM SBĚRU:**

Z následujících charakteristik vyberte tu, která se vztahuje ke konkrétnímu získanému textu, např. (ANO) / ~~NE~~.

Zadán **ČASOVÝ LIMIT**	ANO 10 minut 15 minut 20 minut 30 minut 45 minut 60 minut 90 minut Jiný: …… NE
Při psaní **POVOLEN SLOVNÍK**	ANO NE
Text je **SOUČÁSTÍ ZKOUŠKY**	ANO Typ zkoušky: …… NE
ROZSAH TEXTU ZADÁN	ANO Jaký?* …… NE
VSTUPNÍ AKTIVITA PŘED PSANÍM	žádná obrázek k tématu cvičení k tématu brainstorming slovní zásoby k tématu diskuse na téma zadaná osnova jiná Jaká? ……
ZPŮSOB ZADÁNÍ „TÉMATU“	1 téma určené učitelem výběr z několika témat volné téma jiné zadání Jaké? ……

* uveďte počet slov / počet vět / počet stran dle zadání

C. Manuál pro přepis rukou psaných textů (varianta 10-07)

PROJEKT: Inovace ve vzdělávání v oboru čeština jako druhý jazyk
KLÍČOVÁ AKTIVITA 05: Inovace didaktických metod 1 – budování jazykové databanky se zapojením studentů

MANUÁL PRO PŘEPIS PSANÝCH MATERIÁLŮ

Barbora Štindlová **Václav Lábus**

Tereza Hrdličková

OBSAH:

1. Formát přepisu

a. Přepisovat v programu **Word 2000 a vyšší** (obsahuje znakovou sadu UTF-8).
b. Ukládat a ve formátu **RTF**.

2. Praktická informace

- Každý text k přepisu je označen jedinečným kódem (v tomto formátu: Vob_AR_005)
- Jeho přepis uložte jako zvláštní **rtf soubor** se stejným názvem jako je původní text: např. Vob_AR_005
- **Každý text bude tedy samostatný soubor**!

3. Obecné zásady

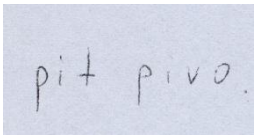
- Úkolem přepisu je **ZACHOVAT AUTENTICKOU PODOBU** původního textu; přepisovač tedy v žádném případě **NEOPRAVUJE CHYBY** ani **NEUVÁDÍ SPRÁVNOU VERZI**!
- Přepisovač využívá pro přepis všech možností textového editoru Word 2000 a vyššího, pouze

v případě problematických úseků textu (nečitelný text apod., viz oddíl 5.5) doprovází přepis smluvenými **znaky** a **kódy** (viz dále oddíl 4).

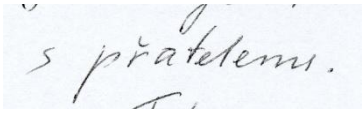
- Při přepisu **VYPNĚTE AUTOMATICKOU KONTROLU PRAVOPISU!** (Nástroje – Možnosti – Pravopis – v záložce Automatická kontrola pravopisu zrušit označení).
- Dávejte pozor na opravování chyb (tj. na to, abyste bezděčně nevkládali správnou variantu např. u diakritiky, interpunkce ap. Tzn. nepište *ve škole* místo originálního *v škole; máma* místo *mama* atd.).
- Přepis po sobě zkontrolujte.

4. Zásady transkripce

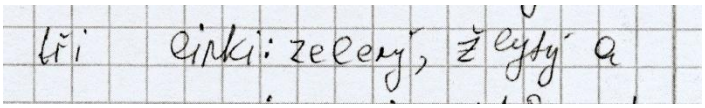
- Přepisuje se reálná podoba textu, **přepisovač neopravuje chyby.**



se přepíše jako „pit pivo“



se přepíše jako „s přatelemi“



se přepíše jako „tři linki: zelený, žlytý a“

- **Věty** se oddělují **<u>dvěma mezerami</u>** (jsou-li hranice vět rozeznatelné), tj. dvakrát mezerník.
  - **POZNÁMKA 1:** V ojedinělých případech se může stát, že hranice mezi větami nejsou rozeznatelné v celém textu (pisatel například neužívá interpunkci). Takový text zachováme v jeho původní podobě, tedy bez interpunkce.
  - **POZNÁMKA 2:** Pokud v textu pisatel zapomene napsat tečku za větou a je zřetelný předěl mezi větami (mezera, velké písmeno na začátku nové věty), zaznamená se to pomocí kódu <.> (tečka ve špičatých závorkách). Pak následují dvě mezery, jako by tam tečka byla.

    Pokud jasný předěl není znát, přepíše se to tak, jak to je.
- **Odstavce** se oddělují **jedním prázdným řádkem**, tj. dvakrát enter. Totéž platí pro oddělení **nadpisu**, je-li v textu přítomen. Pokud je nadpis podtržený, použijeme podtržené písmo.
- **Odrážky** ve všech grafických podobách („bullets“ apod.) se přepisují kódem <li> (o kódech viz dále). Funkce textového editoru Word pro formátování seznamů nelze použít. Kód se píše vždy na začátek věty a s mezerou před prvním slovem. Každá odrážka s textem má pak podobu odstavce a odděluje se jedním prázdným řádkem, tj. dvakrát enter.
- Přepisovač dodržuje **psaní velkých a malých písmen přesně podle originálu**, a to včetně všech chyb; pokud je celý text psaný velkými písmeny, přepíše se také tak; není-li jasné, zda jde o velké, nebo malé písmeno, použije zápis varianty (viz níže oddíl 5.1).

  Pozor na individuální rukopis autora!
- Pokud se objeví v textu „**i“ a „j“ bez tečky**, bude to přepisovač uvádět na pravou míru a přepisovat s tečkou.
- Pokud se v textu psaném na PC objeví **typografické chyby** (např. po tečce na konci věty není mezera, za čárkami nejsou mezery), bude to přepisovač uvádět na pravou míru a zaznamenávat podle pravidel

v manuálu (oddělování vět dvěma mezerami, …).

- **Dělení slov na konci řádku** se nezaznamenává.
- Text obsahuje **tečky** (autor tím dává najevo, že přemýšlí o tématu, či jde o nedokončenou větu; může se vyskytovat i větší počet teček)
  - zapíše se vždy 3 tečkami
- **Předtištěný text:** žák navazuje na již předtištěný text
  - přepsat pomocí kódu <dt>, předtištěný text bude ve složených závorkách, za ně se připíše kód <dt>
  např: {Ahoj Jakube}<dt>
  - pokud se předtištěný text člení do odstavců, přepisovač to přepíše do jednoho odstavce!
- Pro **SPORNÉ A PROBLEMOVÉ JEVY** (viz dále část 5.) se používají tyto vyhrazené **znaky**:

  | || { } [ ] < > XXX

  - Tyto znaky se užívají pro zápis:
    - variant |
    - přepis rektifikace, která dělí slovo ||
    - uzávorkování variantních výrazů a výrazů opatřených kódem (to není nutno u jednoslovných výrazů) { }
    - uzávorkování znaků/řetězců, jejichž přítomnost není v rukopisu jistá { }
    - rektifikací (tj. oprav), uzávorkování rektifikací { }
    - jednoho grafému přepsaného více znaky [ ]
    - chybového kódu <kód>
    - nečitelného textu XXX
  - Na české klávesnici Windows (QWERTZ) se tyto znaky zadávají současným stisknutím klávesy „pravý Alt“ (někdy popsané jako „AltGraph“) a další klávesy:

    < pravý Alt + čárka
    > pravý Alt + tečka
    | pravý Alt + w

    { pravý Alt + b
    } pravý Alt + n
    [pravý Alt + f
    ] pravý Alt + g

Uvedené znaky **SE VKLÁDAJÍ PŘÍMO K PŘÍSLUŠNÉMU ŘETĚZCI,** mezera navíc se nevkládá.

- Vyskytnou-li se znaky <, >, {, }, [a] v rukopisu, je třeba je v přepisu zdvojit:
  <<, >>, {{, }}, [[a]]. [[ ]]
- Výjimkou je znak | . Je-li v textu, nezdvojuje se, ale přepisuje kódem <bar>

- Pro **identifikaci vybraných jevů/chyb** se používají tzv. **CHYBOVÉ KÓDY** (viz dále):

  - **tr** – přesun textu nebo jeho části
  - **in** – vsuvka do textu
  - **pd** – přeškrtnutí učitelem (pokud text znečitelnilo)
  - **img** – obrázek v textu
  - **ni** – přepisovačova interpretace zcela nečitelného textu
  - **co** – komentář přepisovače
  - **bar** – přepis znaku |

- **li** – přepis odrážek (píše se před odstavec, viz výše)
- **gr** – text zapsán jiným písmem než latinkou
- **st** – nabídka alternativ od autora textu
- **dt** – předtištěný text
- . – chybějící tečka (přepíše se <.>)

Chybový kód se uvádí ve většině případů **ve špičatých závorkách** vždy **ZA PŘÍSLUŠNÝM VÝRAZEM BEZ MEZERY,** např. pivo<in>.
Víceslovné výrazy musí být ve složených závorkách, např. {levné pivo}<in>
(Před příslušným výrazem před závorkou, která ho uzavírá zleva, mezera být musí.

5. Problémové jevy

5.1 Varianty

V případě, že rukopis neumožňuje jednoznačné čtení, může přepisovač navrhnout varianty.
Příklady:

Zde není jasné, jestli jde o „Han“, nebo „Hau“. Přepisovač vypíše obě varianty.
Varianty lze uvádět takto:

- Ha{n|u}

V případě, že variantami jsou celá slova, lze závorky vynechat:

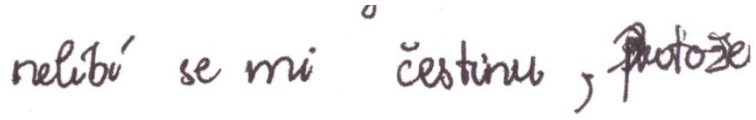
- Han|Hau

Lze-li jednu z variant označit za správnou, uvede se na prvním místě.

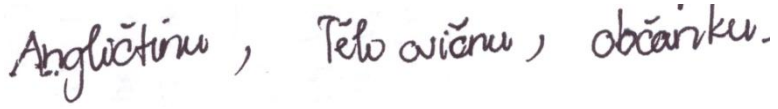
Jde o Můj, nebo Maj?
M{ů|a}j
Můj|Maj

Není jasné, zda písmeno „o“ v rukopisu je, nebo není.
pr{o}dává
prdává|prodává

Lze interpretovat jako jako „t s háčkem“, „t s čárkou“ nebo „n s čárkou“.
Vie{ťn|t'n|tn'}amský
Vieťnamský|Viet'namský|Vietn'amský



Velké, nebo malé „p“?
{P|p}rotože
Protože|protože



Jedno, nebo dvě slova? Zde nemusí být jasné, jestli jde o Tělo cvičnu, nebo Tělocvičnu.
V těchto případech je nutné při přepisu vždy použít závorky:
tělo{ }cvičnu
{tělo cvičnu|tělocvičnu}

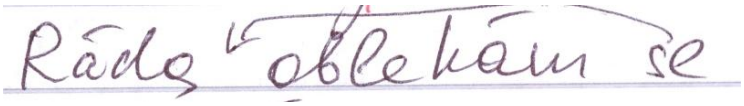
Znak, jehož přítomnost v rukopisu je nejistá, je uzavřen ve složených závorkách { }.
Mezera se chápe jako každý jiný znak, je tedy třeba zapsat
tělo{ }cvičnu (ve složených závorkách je mezera!), nikoli tělo{}cvičnu.

5.2 Rektifikace

Pro maximální výpovědní hodnotu přepisu **je třeba zachovat i pisatelovy vlastní opravy**.

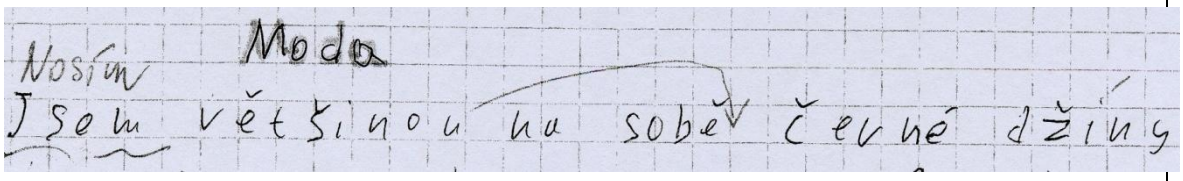
5.2.1 Přesuny

- Přesuny na krátké vzdálenosti (znak ->):
  - přesuny písmen: Če{ks->sk}á
  - lokální přesuny slov, např.



Ráda {oblekám se -> se oblekám}

- Přesuny (skupin) slov na větší vzdálenosti (kód „tr“):



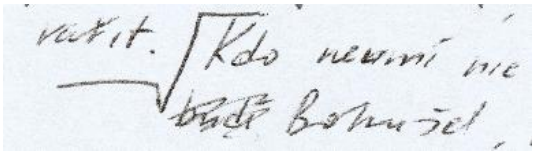
~~Jsem~~ Nosím {většinou}<tr-odsud-1> na sobě <tr-sem-1> černé džíny
Vysvětlivky:

- {většinou} = uzávorkované slovo, které se má přesunout, za něj (bez mezery) se uvede příslušný chybový kód: <tr-odsud-1>
- tr – identifikace kódu pro přesun
- odsud – identifikace, že právě odsud se předchozí slovo/řetězec přesouvá jinam
- sem – identifikace, že příslušné slovo/řetězec má přesunout právě sem
- 1 – číselná identifikace konkrétní dvojice (nutné v případě, že je v jedné větě více přesunů – v takovém případě by bylo dále <tr-odsud-2> ... <tr-sem-2>)

Další příklad (opačný směr):

já <tr-sem-1> hrozně moc těšil {jsem se}<tr-odsud-1>

- Oprava v odsazení odstavce se nijak neznačí. Bere se v úvahu jako nový odstavec:


se přepíše jako:

vařit.

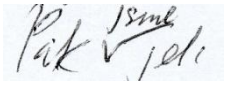
Kdo neumí nic …

5.2.2 Rozdělování slov (znak ||)

ve||městě

5.2.3 Vsuvka do textu (kód „in“)

- **Vsuvka slova**:



Pak jsme<in> jeli

Pak {jsme}<in> jeli

- **Vsuvka písmene**:

„pro{c}<in>házela

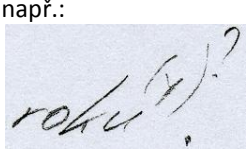
- **Vsuvka pomocí hvězdičky na konci textu:**

- tam, kam student vsouvá text, přepsat s kódem „in“

Je-li do textu vsouváno 2 a více slov, musí být všechna tato slova vždy umístěna do složených závorek, tj. např. Pak {jsme všichni}<in> jeli

5.2.4 Nabídka alternativ od autora textu (kód „st“)

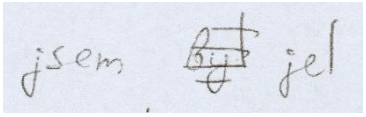
Nabízí-li autor textu (nejde tedy o alternativu od přepisovače) alternativu, pak se označí kódem <st>, např.:



se přepíše jako: roku{u|(y)}<st> (kulaté závorky proto, že jsou v originále)

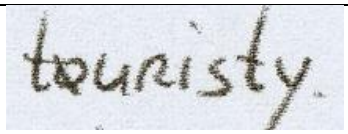
5.2.5 Škrtání

Škrtnutý řetězec označit jako přeškrtnutý funkcí Wordu Formát – Písmo – Přeškrtnuté:



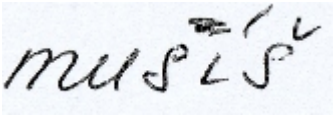
jsem ~~Byl~~ jel

Je-li přeškrtnutá pouze část slova, v přepisu bude vyznačena přeškrtnutým písmem jen ta.

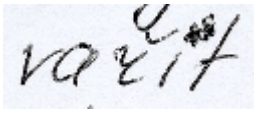


to~~u~~risty.

Škrty diakritických znamének se přepisují tímto způsobem:

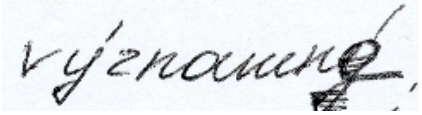


mu~~s~~síš



vař~~i~~it

Přepis, pokud je čitelný, se přepíše stejně jako přeškrtnutí:



významn~~ý~~é

Začmárané a zcela zaškrtané, tudíž **nečitelné řetězce se nepřepisují**.

5.2.6 Škrt učitele (kód „pd“)

Veškeré učitelovy škrty, vpisky a opravy v textu ignorujte, nijak je při přepisu nezachycujte.

Pouze v případě, že původní žákova slova jsou kvůli učitelovu zásahu nečitelná, označte příslušnou pasáž XXX (viz oddíl 5.5) a chybovým kódem.

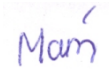
Např. jsem XXX<pd> jel

nebo jsem {XXX XXX}<pd> jel

5.3 Diakritická znaménka

Řadu grafémů lze ve Wordu zapsat pomocí funkce Vložit/Symbol, je-li k dispozici písmo, které tento grafém obsahuje.

Některé grafémy lze zapsat kombinací diakritického znaménka a mezery, např. *m s čárkou* nebo *j s čárkou* atp.

Příklad:  "m s čárkou"

1. současně stisknout klávesy "pravý Alt" a "f"
2. stisknout klávesu "m"
3. stisknout klávesu s diakritickým znaménkem (exotická znaménka jsou dostupná přes klávesy s čísly při současném stisknutí klávesy pravý Alt)
4. stisknout mezerník
5. současně stisknout klávesy "pravý Alt" a "g"

[m´]

Výsledkem je zápis [m'], podobně i [j']. Závorky jsou důležité kvůli odlišení od výskytu diakritického znaménka v pozici za písmenem: [j'] (čárka nad j) vs. j' (čárka za j).

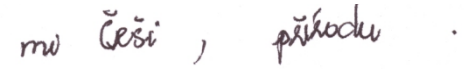
Nedokáže-li přepisovač zapsat grafém přímo (viz výše), zapíše písmeno bez diakritiky a další diakritická znaménka hned za něj, vše uzavřené do hranatých závorek: [<písmeno><znaménko>]

Některá exotická diakritická znaménka lze zadat zjednodušeným způsobem:

přehláska	ä	[a“]	tilda nad písmenem	ñ	[n~]
dlouhá přehláska	ű	[u:]	šikmá čárka	ł	[L/]
cedilla	ç	[c,]	těžký akcent	è	[e`]
ogonek	ą	[a;]	vodorovná čárka	ā	[a-]
tečka nad písmenem	ż	[z.]	kroužek nad písmenem	å	[ao]
stříška nad písmenem	â	[a^]			

5.4 Interpunkce

- Mezera před interpunkčním znaménkem nebo mezera za ním se přepisovat nebude, tj. přepisovač bude tuto chybu sám uvádět na pravou míru.

Např. 

mi Češi, přírodu

- Dvě a více interpunkčních znamének za sebou se přepíše podle rukopisu.

Např. 

školy.,

5.5 Nečitelné řetězce (znak XXX, kód „ni“)

- Totálně nečitelná slova/řetězce zapsat jako XXX.

  Např. Pracuju jako XXX

- Je-li přepisovač schopen identifikovat počet nečitelných slov, uvede XXX v příslušném počtu.

  Např. Pracuju jako XXX XXX

  Pozn.: znak XXX je jeden celek, který se užívá jak pro písmeno, skupinu písmen, tak pro slovo; jen X se nikdy nepíše.

- Pokud je slovo/řetězec nečitelný, ale přepisovač je schopen na základě kontextu nebo jiných skutečností interpretovat nějakou možnost čtení, může ji zapsat, a to tímto způsobem (kód „ni“):

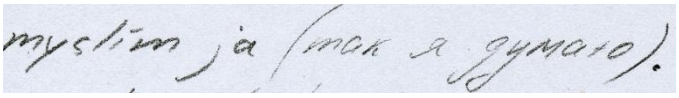
  Např. Pracuje jako {dělník}<ni>

  (V nejzazším případě, nabízí-li se ve zcela nečitelném úseku více interpretací, zapište jako variantu s kódem pro nečitelnost.

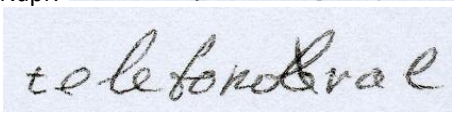
  Např. Jí rád {rybu|ryby}<ni>

5.5.1 Azbuka a jiné grafické systémy (kód „gr“)

Není-li text zaznamenán latinkou, zvolí se znak pro nečitelný text – XXX (viz výše) a doplní se kódem **gr**

Např. 

myslím ja (XXX XXX XXX<gr>)



telefono~~XXX~~val<gr>

XXX v těchto případech nahrazuje:
- jeden nečitelný znak, který je součástí slova
- jeden řetězec nečitelných znaků, který je součástí slova

- nečitelné slovo

5.6 Obrázky v textu (kód „img“)

Vyskytne-li se v textu obrázek (vlepený, kreslený, emotikon apod.), uveďte na příslušném místě kód se stručným popisem obrázku, i víceslovným. Emotikony se přepisují obvyklým způsobem, jen je třeba se vyvarovat znaku >.

Např.:

šel tam pes <img pes>

mám ráda chlupaté knedlíky <img chlupaté knedlíky>

Ukázka přepisu emotikon:

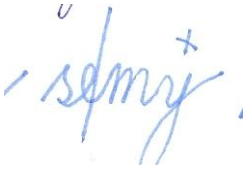
<img :-)> <img :)> <img :-(> <img :-|> <img :-]> <img :{>

Emotikony nesmí být ve formě obrázku (☺)!!

Není-li obrázek přímo v textu, umístit kód za nejbližší odstavec nebo na konec textu.

6. Komentář přepisovače

Potřebuje-li přepisovač přidat k nějakému jevu komentář/popis atp., použije kód „co“, např.:



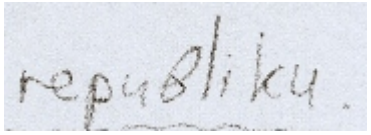
se||my{přeškrtnutá čárka nad y}<co>

Text komentáře je vždy ve složených závorkách a **komentář se vždy vztahuje jen k jednomu slovu**.

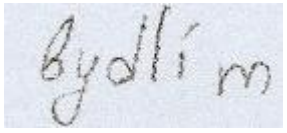
7. Respektování rukopisu autora

Při přepisu je nutné brát v potaz celý text a respektovat způsob psaní autora. Různé, zejména grafické odchylky (např. od psacího písma latinky) není třeba hodnotit jako chyby nebo varianty, ale projev individuálního způsobu psaní autora textu. Příklady viz níže. Není třeba hledat problémové jevy tam, kde nejsou!

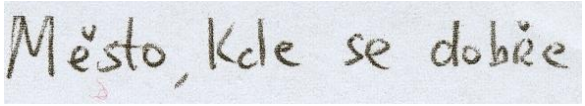
např.

 se přepíše jako

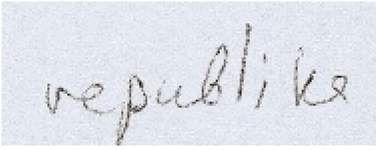
repu**b**liku, nikoli repuBliku (malé „b“, srov. další slova v textu apod.)

 se přepíše jako

bydlí**m**, nikoli {bydlím|bydlí m}

 se přepíše jako

Město, **k**de se dob**ř**e, nikoli Město, Kde s dobŘe

 se přepíše jako

republik**a,** nikoli republik{a|e} – viz úzus psaní „e“ a „a“ na konci slov (srov. další slova v textu)

8. Kódování vlastních jmen

- Kvůli anonymitě je nutné osobní jména (křestní i příjmení), **která se týkají pisatele textu či jeho rodiny**, nahrazovat:

- křestní jméno a příjmení se bude nahrazovat pouze křestními jmény takto:

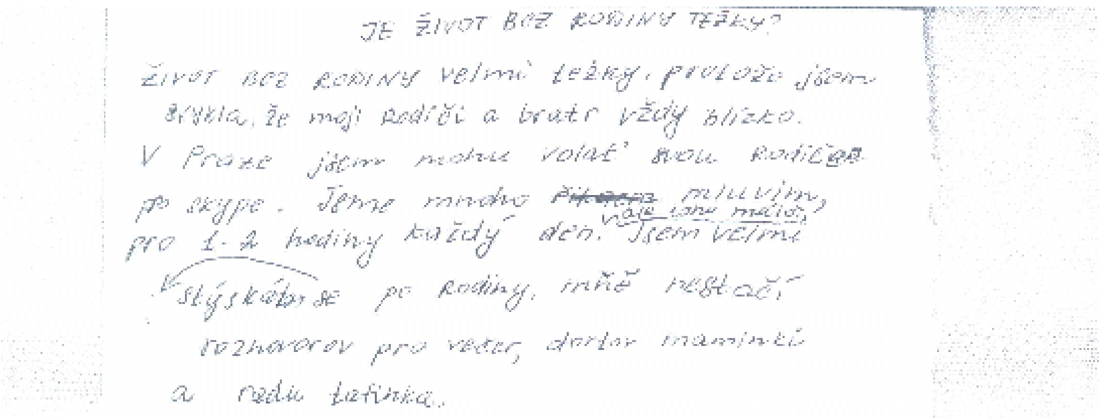
- mužské jméno (křestní a příjmení): **Adam**
- ženské jméno (křestní a příjmení): **Eva**
- jméno, u něhož není možné identifikovat rod: **Sin**

POZNÁMKA: zástupné jméno bude v příslušném pádě jako v originálním textu; pokud je z kontextu jasné, že jde o špatně skloněné jméno, použije se nominativ (1.pád) (např. „viděl jsem Murat“ místo „viděl jsem Murata“, přepíše se jako „viděl jsem Adam“)

- Ostatní jména (určená v zadání slohové práce, jednoznačně fiktivní, jména zvířat, přezdívky) se přepisují podle původního textu.
- Pokud se objeví chyba v osobním jméně, které se týká pisatele textu, zakóduje se podle pravidla výše (tedy Adam, Eva, Sin), chybu nezaznamenáváme.
- Pokud se objeví v textu dvě jména stejného rodu (např. jméno otce a bratra), která je nutno zakódovat, použije se pouze zástupné jméno (bez jakéhokoliv číslování či jiného odlišování), tedy např. Adam a Adam.
- Zdrobněliny u osobních jmen, která se nahrazují zástupnými jmény, používáme také ve zdrobnělém tvaru (Evička, Adámek).
- V textu se také mohou vyskytovat jména pisatelů textu, která ale nejsou součástí textu (jde o klasický podpis slohové práce), ty se nepřepisují; přepisují se pouze tehdy, jestliže jsou součástí textu (podpis u dopisu, životopisu, …).
- Místní jména

- jména měst, částí měst, obcí, ulic, … se nekódují, nechávají se v textu v původním znění
- pokud se vyskytuje v textu číslo popisné – nepřepisuje se!

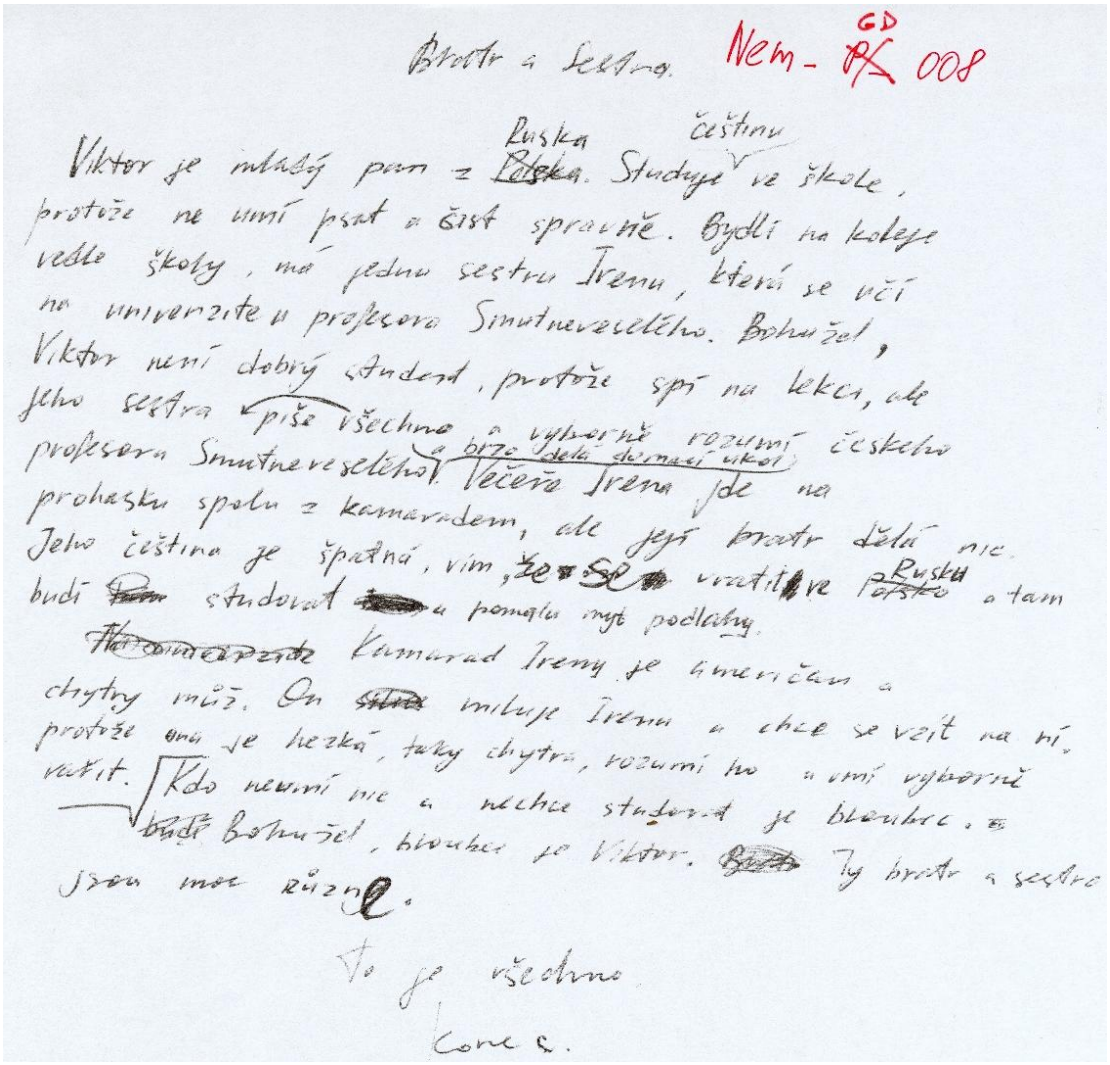
9. Ukázky přepsaných textů



JE ŽIVOT BEZ RODINY TEŽKY?

Život bez rodiny velmi težky, protože jsem svykla, že moji rodiči a bratr vždy blízko. V Praze jsem mohu volať svou rodiče po|pro skype. Jsme mnoho ~~říkam~~ mluvim, pro 1-2 hodiny každý den. {ale toho málo.}<in> Jsem velmi

{stýskám se -> se stýskám} se po rodiny, mně nestačí rozhovorov pro večer, dortov maminki a radu tatinka.



Bratr a Sestra.

Viktor je mladý pan z ~~Polska~~ Ruska. Studuje {češtinu}<in> ve škole, protože ne umí psat a čist spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra {piše všechno -> všechno piše} a vyborně rozumí českeho profesora Smutneveseleho {a brzo delá domací ukol}<in>. Večeře Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve ~~Polsko~~ Rusk~~o~~u a tam budí studovat u pomalu myt podlahy.

Kamarad Ireny je {A|a}meričan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborně vařit.

Kdo neumí nic a nechce studovat je bloubec. ~~budi~~ Bohužel, bloubec je Viktor. Ty bratr a sestra jsou moc různ~~y~~e.

To je všechno.

Konec

PŘÍLOHA 5

Kódy chyb pro anotaci žákovského korpusu CzeSL[222]

TYP CHYBY	PODTYP CHYBY	POPIS CHYBY	ROVINA	PŘÍKLADY	
incor		nesprávný tvar (obecná kategorie pro neanotované případy oprav)	R1		
	incorInfl	nesprávná flexe	R1	*V **šestě** třídě, nosit **sukné,** to jsou drahý **prsteni**, **jetu** na kole*	M
	incorBase	nesprávný kmen	R1	*Bydlím s **mátkou,** lidé jsou **mérný**, kočka se **jmemuje, libili** se mi, mluvim **česka,***	M
	incorOther	ostatní nesprávné tvary	R1		A
fw		nově utvořené / cizí slovo	R1		
	fwFab	nově utvořené slovo	R1	*Anglicky jidlo je strasny, všechno je **chlebasa.** Nechce slyšet **smášky.***	M
	fwNc	cizí / neidentifikovatelné slovo	R1	*Jím rád **eggs**, jdu do **shopu**, musim **dopaintovat***	M
	flex	přítomnost flexe	R1		M
wbd		chybná hranice slov	R1		
	wbdPre	prefix oddělený mezerou a předložka bez mezery	R1	***Při pravit,** já jsem to rychle **roz vázal,** Petr už **jel od**; **dolesa, veškole***	M
	wbdComp	neoprávněně rozdělená kompozita	R1	*Sádro karton; šesti násobný; **anglo-americká** literatura*	M
	wbdOther		R1	*Socialis mus, mochezky*	A
agr		narušení shody	R2	*Jen jedna **holky** zůstala doma, Petr **vařím** oběd, máme **hezkých** psa*	M
dep		chyba ve vyjádření syntaktické závislosti	R2	*Věřím **učitelku,** dal obraz **dole**, doufali ve **vitězí**, řekl mi **napsat** to*	M
ref		chyba v zájmenném odkazu	R2	*Dal jsem to jemu i **jejího** bratrovi, lidé podle **té** žijou.*	M
vbx		chyba v analytickém slovesném tvaru a složeném přísudku	R2	↓	M
	cvf	chyba v analytickém slovesném tvaru	R2	*Zítra budu **napsat** dopis, **učím** jsem se, zítra bude **pršeli, by jsi** přišel*	A
	mod	chyba v konstrukci s modálním nebo fázovým slovesem	R2	*Musím **píšu** domácí úkol, začnu **přečíst** celou knížku*	A

[222] Uvedenou chybovou tabulku zpracovaly pro interní potřeby anotátorů Svatava Škodová a Barbora Štindlová.

	vnp	chyba ve sponově-jmenném přísudku (vč. pas. a rez.)	R2	***Byl*** *mu teplo, je třeba* ***udělám,*** *zahrada je* ***sousedovi,*** *dostal* ***vynadat***	A
rflx		chyba v reflexivním výrazu	R2	*Smála* ***si****, narodila jsem v Petrohradu, dnes* ***si*** *neučí*	M
neg		chyba v negaci	R2	*Půjdu* ***neráno****, on* ***ne*** *velký, máma ani táta* ***kouří, mám*** *žádný čas*	M
odd		nadbytečné slovo	R2	↓	
	oddObj	nadužití osobního zájmena ve funkci předmětu	R2	*Petra jsem* ***ho*** *viděl,* ***kamarádce*** *koupil jsem jí zmrzlinu*	M
	oddSubj	nadužití osobního zájmena ve funkci podmětu	R2	***On*** *nesnídá, protože* ***on*** *nemá hlad*	A
	oddOther	ostatní nadbytečná slova	R2		A
miss		chybějící slovo	R2	↓	
	missObj	chybějící výraz ve funkci předmětu	R2	*To je Petr. Znáš?(/Znáš* ***ho?)***	M
	missPred	chybějící spona	R2		A
	missOther	ostatní chybějící slova	R2		A
wo		chybný slovosled	R2	*Koupil jsem* ***auto nové,*** *to by velmi vadilo* ***mi***	A
lex		chyba v lexiku a frazeologii	R2	*Bydlíme v domě* ***i*** *nebydlíme v bytě. Vladimír je* ***lékařka.***	M
use		chyba ve významovém užití gramatické kategorie	R2	*Nemám rád* ***třešni,*** *včera* ***bude sněžit,*** *pošta je* ***nejvíc blízko***	M
sec		sekundární, "zavlečená" chyba	R2	*Dívám se na* ***americkém*** *filmu.*	M
disr		rozvrácená konstrukce	R2	*Zkušební důvtip muže te řidit, mám rád packu viečkem*	M
styl		obecněčeský, knižní, nářeční tvar	R1,R2		
	stylColl	potenciální obecněčeský tvar	R1,R2	*Viděli jsme* ***hezký*** *holky.*	M
	stylOther	knižní, nářeční, slangový ap. tvar/výraz	R1,R2	*Jedeme* ***šalinou, pláči*** *nad* ***vejdělkem****.*	M
problem		problémová chyba	R1,R2		M

Legenda:

R1 **Teplé barvy**	**R2** **Studené barvy**	**R1+R2** **fialová**	**A** **Automaticky šedá**

PŘÍLOHA 6

Přehled anotátorů a jejich podílu na anotaci vzorku

ANOTÁTOR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	CELKEM:
ROK STUDIA	2.	2.	3.	2.	2.	2.	2.	2.	3.	3.	3.	3.	3.	3.	
PROŠKOLENÍ (dle skupiny)	P	P	P	P	P	P	P	P	L	L	L	L	L	L	
ZKUŠENOST S ČCJ:	NE	NE	NE	NE	NE	NE	ANO	NE	NE	ANO	ANO	ANO	NE	NE	
POČET ANOTOVANÝCH SLOV:	1602	1609	1529	1620	1335	1808	1462	1588	1202	1934	1822	1245	1012	883	20651
POČET ANOTOVANÝCH TEXTŮ:	14	14	14	14	14	14	13	14	9	11	9	9	6	5	160

12 textů (tj. 955 slov) bylo pro technické nedostatky v programu FEAT z anotačního vzorku vyloučeno.

PŘÍLOHA 7

A. Anotační dotazník[223]

Slouží jako zpětná vazba pro autory manuálu a jako doplňkový zdroj informací při zhodnocení výsledků měření IAA.

	TEXT Č.:
1. váhám mezi emendací na.. a na …	
2. nevím, jak mám emendovat	
3. nevím, jestli je to chyba	
4. nevím, jakou značku přiřadit	
5. váhám mezi značkami	
6. chybí mi značka	
7. nevím, jestli mám odkazovat	
8. nevím, kam mám odkázat	
9. v manuálu nerozumím	
10. technické aspekty anotace (např. nemůžu udělat, protože mi chybí …)	
11. jiné:	

[223] Autorkou tohoto dotazníku je B. Štindlová.

B. Příklady vyplněného dotazníku

(1)

	TEXT Č.: HRD_JG_007_t_1
1. váhám mezi opravou na.. a na …	
2. nevím, jak mám opravit	Věta: „Je to jazyk úplně jiný od francouzsky.“ Mám problém určit, zda můj zásah do textu není už přespříliš… Rozhoduji se pro opravu: „Je to jazyk úplně jiný než francouzský.“ Váhám nyní, jak změnit „y“ na „ý“. Původně jsem to udělala už na R1 (s tím, že byla jiná intence pisatele), ale nakonec jsem usoudila, že k tomu není důvod, jelikož slovo „francouzsky“ existuje…
3. nevím, jestli je to chyba	Věta: „Čeština je jazyk, který zapomínáme rychle, pokud musíme poslouchat rádia, číst časopisy, mít český kamarády.“ Sdělení, že autorčina znalost češtiny se horší, když přichází do styku s českými mluvčími atd., nedává – zdá se – příliš smysl. Váhám proto, zda bych neměla „zapomínáme“ nahradit jiným slovesem s tím, že autorka si spletla význam slova… Abych do textu nezasahovala příliš a vyhnula se velkým změnám, rozhoduji se prozatím větu interpretovat tak, že autorka pro samou zábavu nemá čas na systematické osvojování jazyka, a proto se její znalosti horší.:)
4. nevím, jakou značku přiřadit	
5. váhám mezi značkami	Věta: „Chtěla jsem se učit češtinu, když jsem přijela poprvé v Praze.“ Opravuji na „do Prahy“. Předložku „do“ označuji jako chybu v lexiku a frazeologii, avšak váhám, co v případě „Prahy“. V jednom z textů, jež jsme anotovali společně v semináři, jsme měli výraz „od pondělí do pátek“. „Pátek“ jsme klasifikovali jako chybu ve vyjádření syntaktické závislosti. Myslím ale, že toto není ten samý případ, protože ve spojení „v Praze“ je syntaktická závislost v pořádku, a tak značím „Prahu“ ve shodě s „do“ jako chybu v lexiku a frazeologii.
6. chybí mi značka	
7. nevím, jestli mám odkazovat	
8. nevím, kam mám odkázat	
9. technické aspekty anotace (např. nemůžu udělat, protože mi chybí …)	Věta: „Chtěla jsem se učit češtinu, když jsem přijela poprvé v Praze.“ Opravuji předložku „v“ jako chybu v lexiku a frazeologii a myslím, že by bylo dobré vést od opravené předložky odkaz k přísudku „přijela“. To však není možné (nebo mně to nejde), protože autorka z neznámého důvodu píše text jako poezii (přičemž neplatí, že nový řádek = nová věta: autorka začíná nové řádky zcela libovolně). A mně se každý řádek zobrazuje zvlášť… Lze to nějak udělat, aby se to v anotátorském okně propojilo a daly se vést odkazy alespoň v rámci té samé věty?
10. v manuálu nerozumím	
11. jiné:	Nejsem si jistá, jestli jsem se správně vypořádala se situací, kde v přísudku „naučila jsem“ chybělo zvratné zájmeno „se“. „Se“ jsem doplnila na jeho místo ve větě a teď: Mohu ho tam nechat „jen tak“? Nebo od něj musím vést nějaké hrany, někam vést odkaz…? Mohu ho v tomto případě označit jako chybu v reflexivním výrazu i jako chybějící objekt?
	Proč se mi na rovině oprav zobrazují všechny infinitivy napsané takto: čís(r/t), mí(r/t)…? Jde o zkrat při přepisu, nebo autorce v tu chvíli opravdu vypadlo,

	jaká je česká infinitivní koncovka, a zvolila tuto formu zápisu? Jelikož není v textu nikde žádná zkratka, která by signalizovala, o co se jedná, rozhodla jsem se považovat to opravdu za nejistotu autorky a na R1 opravuji do správných tvarů…

(2)

	TEXT Č.: UJA_A1_002_t_1.b
1. váhám mezi opravou na.. a na …	
2. nevím, jak mám opravit	
3. nevím, jestli je to chyba	„Mám moc ráda dárky“ – pro dostatečnou srozumitelnost neopravuji, ale myslím, že autorka spíš měla na mysli význam „Moc se mi líbí dárky“ (pokud je z anglicky mluvících zemí, pak by tu byl jasný vliv anglického slovesa „like“).
4. nevím, jakou značku přiřadit	
5. váhám mezi značkami	U opravy R1 „všechny“ na R2 „všichni“ váhám, zda přidat i tag „sec“ a odkaz ke „kamarádi“ (z důvodu toho, že čistě spojení „všechny kamarády“ je gramaticky správně)
6. chybí mi značka	
7. nevím, jestli mám odkazovat	
8. nevím, kam mám odkázat	
9. technické aspekty anotace	
10. v manuálu nerozumím	Vysvětlení opravy „sec“
11. jiné:	Spojka „a“ v takové koncentraci, jako u tohoto textu, působí značně rušivě, zajímalo by mne, jestli i v tomto případě je nutno se držet zásady, že cílem není stylisticky vybroušená věta…
	V manuálu chybí zmínka o tom, jsou-li anotátoři oprávněni rozdělovat příliš dlouhé větné celky.

(3)

	TEXT Č.: UJA_BM_001_t_1
1. váhám mezi opravou na.. a na …	Autor v textu dvakrát oddělil věty čárkami, avšak po nich použil velké písmeno, jako kdyby se jednalo o tečku. Ani v jednom případě se nejedná o souvětí podřadné, věty jsou spojeny volně, tudíž by obě interpunkční znaménka byla možná. Váhám, zda je větší zásah do textu, když opravím čárku na tečku a souvětí roztrhnu (není možné, že autorovi „ujelo pero“ a znaménko se při přepisu chybně interpretovalo?), nebo když opravím velké písmeno na

	malé. (Z důvodu opakování chyby nakonec volím opravu velkého písmena na malé.)
2. nevím, jak mám opravit	
3. nevím, jestli je to chyba	V textu je věta: „Ale cena takového bytu je hrozně drahá.“ Opravuji zde na R2 slovo „drahá“ na „vysoká“ jako chybu v lexiku a frazeologii, ačkoliv si nejsem jistá, zda tato má oprava není nadbytečná. Vlastní jazyková zkušenost mi říká, že se toto spojení nepoužívá a SSČ uvádí jen „vysokou cenu“ a pak případně „koupit něco za drahé PENÍZE“…
4. nevím, jakou značku přiřadit	Věta: „Budeme se přestěhovat za týden.“ Opravila jsem zde „přestěhovat“ na „stěhovat“, avšak stále nemohu přijít na to, jak označit, když pouze měním vid u slovesa…
5. váhám mezi značkami	
6. chybí mi značka	
7. nevím, jestli mám odkazovat	
8. nevím, kam mám odkázat	
9. technické aspekty anotace	
10. v manuálu nerozumím	
11. jiné:	Věty: „A pak jsme našli panelák, který je 3 plus 1, docela daleko od centra. Ale ten panelák je docela hezký.“ Opravuji zde na R2 „panelák“ na „byt“, jelikož panelák zjevně nemůže být o zmiňované velikosti. V následující větě váhám, zda „panelák“ opravit z důvodu první opravy na „byt“ také, ale jelikož v tomto případě věta smysl dává, rozhoduji se ponechat původní výraz.

PŘÍLOHA 8

IAA – kalkulace kappa koeficientu

```
--- Kappas for layer 1 (i.e. layer a) ---

noOfTags  13

Tag  1   incorInfl
   n (nr of tokens) 9848
   Table  9295 173
          130 250
   p      0.9438465 0.01756702
          0.01320065 0.02538587
   po     0.9692323 ,  pe  0.9217754
   kappa         0.6066753

Tag  2   incorStem
   n (nr of tokens) 9848
   Table  8957 167
          165 559
   p      0.9095248 0.01695776
          0.01675467 0.0567628
   po     0.9662876 ,  pe  0.8636015
   kappa         0.7528388

Tag  3   fw
   n (nr of tokens) 9848
   Table  9837 4
          6 1
   p      0.998883 0.0004061738
          0.0006092608 0.0001015435
   po     0.9989846 ,  pe  0.9987822
   kappa         0.1661728

Tag  4   fw:fwFab
   n (nr of tokens) 9848
   Table  9809 23
          13 3
   p      0.9960398 0.002335500
          0.001320065 0.0003046304
   po     0.9963444 ,  pe  0.9957438
   kappa         0.1411295

Tag  5   fw:fwNc
   n (nr of tokens) 9848
   Table  9826 10
          9 3
   p      0.997766 0.001015435
          0.0009138911 0.0003046304
   po     0.9980707 ,  pe  0.9974646
   kappa         0.2390357

Tag  6   wbd
   n (nr of tokens) 9848
   Table  9808 26
          9 5
   p      0.9959383 0.00264013
          0.0009138911 0.0005077173
   po     0.996446 ,  pe  0.9954395
```

```
   kappa         0.2206958

Tag  7   wbd:wbdPre
   n (nr of tokens) 9848
   Table  9819 12
          14 3
   p      0.9970552 0.001218522
          0.001421608 0.0003046304
   po     0.9973599 ,  pe  0.9967559
   kappa         0.1861830

Tag  8   wbd:wbdComp
   n (nr of tokens) 9848
   Table  9808 5
          27 8
   p      0.9959383 0.0005077173
          0.002741673 0.0008123477
   po     0.9967506 ,  pe  0.9951353
   kappa         0.3320475

Tag  9   styl:stylColl
   n (nr of tokens) 9848
   Table  9834 10
          2 2
   p      0.9985784 0.001015435
          0.0002030869 0.0002030869
   po     0.9987815 ,  pe  0.9983763
   kappa         0.2495428

Tag  10   incorSum: 1 incorInfl, 2 incorStem
   n (nr of tokens) 9848
   Table  8656 168
          130 894
   p      0.8789602 0.0170593
          0.01320065 0.09077985
   po     0.96974 ,  pe  0.8106067
   kappa         0.840227

Tag  11   fwSum: 3 fw,  4 fw:fwFab,  5 fw:fwNc
   n (nr of tokens) 9848
   Table  9788 25
          17 18
   p      0.9939074 0.002538587
          0.001726239 0.001827782
   po     0.9957352 ,  pe  0.9921106
   kappa         0.4594202

Tag  12   wbdSum: 6 wbd, 7 wbd:wbdPre, 8 wbd:wbdComp
   n (nr of tokens) 9848
   Table  9768 14
          21 45
   p      0.9918765 0.001421608
          0.002132413 0.004569456
   po     0.996446 ,  pe  0.9873874
   kappa         0.7182173

Tag  13   stylSum: 10 styl, 11 styl:stylColl, 12 styl:stylOther
   n (nr of tokens) 9848
   Table  9826 17
          3 2
   p      0.997766 0.001726239
```

```
          0.0003046304 0.0002030869
   po     0.9979691 ,  pe  0.997565
   kappa         0.1659962

Summary of kappas for each tag
0.6066753 0.7528388 0.1661728 0.1411295 0.2390357 0.2206958 0.1861830
0.3320475 0.2495428 0.840227 0.4594202 0.7182173 0.1659962

--- Kappas for layer 2 (i.e. layer b) ---

noOfTags  17

Tag  1   agr
   n (nr of tokens) 9848
   Table  9557 82
          99 110
   p      0.9704509 0.008326564
          0.01005280 0.01116978
   po     0.9816206 ,  pe  0.9601086
   kappa         0.539265

Tag  2   dep
   n (nr of tokens) 9848
   Table  9544 99
          118 87
   p      0.9691308 0.01005280
          0.01198213 0.008834281
   po     0.977965 ,  pe  0.9610828
   kappa         0.4337993

Tag  3   ref
   n (nr of tokens) 9848
   Table  9814 14
          17 3
   p      0.9965475 0.001421608
          0.001726239 0.0003046304
   po     0.9968522 ,  pe  0.99625
   kappa         0.1605957

Tag  4   vbx
   n (nr of tokens) 9848
   Table  9816 20
          9 3
   p      0.9967506 0.002030869
          0.0009138911 0.0003046304
   po     0.9970552 ,  pe  0.9964517
   kappa         0.1700995

Tag  5   rflx
   n (nr of tokens) 9848
   Table  9828 6
          11 3
   p      0.9979691 0.0006092608
          0.001116978 0.0003046304
   po     0.9982738 ,  pe  0.997667
   kappa         0.2600463

Tag  6   neg
   n (nr of tokens) 9848
   Table  9819 11
          9 9
```

```
   p      0.9970552 0.001116978
          0.0009138911 0.0009138911
   po     0.9979691 ,  pe  0.9961488
   kappa         0.4726696

Tag  7   odd:oddObj
   n (nr of tokens) 9848
   Table  9846 0
          2 0
   p      0.999797 0
          0.0002030869 0
   po     0.999797 ,  pe  0.999797
   kappa         0

Tag  8   miss:missObj
   n (nr of tokens) 9848
   Table  9847 0
          1 0
   p      0.9998985 0
          0.0001015435 0
   po     0.9998985 ,  pe  0.9998985
   kappa         0

Tag  9   lex
   n (nr of tokens) 9848
   Table  9536 107
          131 74
   p      0.9683184 0.01086515
          0.01330219 0.007514216
   po     0.9758327 ,  pe  0.9615694
   kappa         0.3711431

Tag  10   use
   n (nr of tokens) 9848
   Table  9695 60
          74 19
   p      0.9844639 0.006092608
          0.007514216 0.001929326
   po     0.9863932 ,  pe  0.982686
   kappa         0.2141128

Tag  11   sec
   n (nr of tokens) 9848
   Table  9781 45
          18 4
   p      0.9931966 0.004569456
          0.001827782 0.0004061738
   po     0.9936028 ,  pe  0.9928126
   kappa         0.1099315

Tag  12   styl:stylColl
   n (nr of tokens) 9848
   Table  9810 14
          14 10
   p      0.9961413 0.001421608
          0.001421608 0.001015435
   po     0.9971568 ,  pe  0.9951378
   kappa         0.4152416

Tag  13   styl:stylOther
   n (nr of tokens) 9848
```

```
   Table  9847 1
          0 0
   p      0.9998985 0.0001015435
          0 0
   po     0.9998985 ,  pe  0.9998985
   kappa         0

Tag  14   disr
   n (nr of tokens) 9848
   Table  9787 11
          50 0
   p      0.9938058 0.001116978
          0.005077173 0
   po     0.9938058 ,  pe  0.9938172
   kappa         -0.001834471

Tag  15   problem
   n (nr of tokens) 9848
   Table  9830 8
          10 0
   p      0.9981722 0.0008123477
          0.001015435 0
   po     0.9981722 ,  pe  0.9981739
   kappa         -0.000903424

Tag  16   missSum: 8 miss:missPred, 9 miss:missObj
   n (nr of tokens) 9848
   Table  9847 0
          1 0
   p      0.9998985 0
          0.0001015435 0
   po     0.9998985 ,  pe  0.9998985
   kappa         0

Tag  17   stylSum: 13 styl, 14 styl:stylColl, 15 styl:stylOther
   n (nr of tokens) 9848
   Table  9805 19
          14 10
   p      0.9956336 0.001929326
          0.001421608 0.001015435
   po     0.996649 ,  pe  0.9946325
   kappa         0.3756935

Summary of kappas for each tag
0.539265 0.4337993 0.1605957 0.1700995 0.2600463 0.4726696 0 0 0.3711431
0.2141128 0.1099315 0 0.4152416 0 -0.001834471 -0.000903424 0 0.3756935
```

PŘEKLADOVÝ SLOVNÍK TERMÍNŮ

Následující překladový slovník vznikl primárně pro potřeby této práce. Předpokládám však, že může sloužit jako základ výkladového slovníku, který bude reflektovat odbornou terminologii z oblasti žákovských korpusů a oboru češtiny jako cizího jazyka. Řada cizojazyčných termínů nemá doposud zavedený český ekvivalent, proto jsou některé návrhy uvedené v tabulce určeny k další odborné rozpravě.

	ANGLICKÝ TERMÍN	ČESKÝ EKVIVALENT
A	*accuracy*	1. přesnost, 2. správnost
	acquired (system)	osvojený (systém)
	acquisition hierarchy	1. hierarchie osvojování jazyka, 2. akviziční hierarchie
	acquisition: language a.	1. nabývání, 2. akvizice, 3. osvojování (jazyka)
	▪ *second language acquisition (SLA)*	nabývání , osvojování, akvizice druhého jazyka
	▪ *foreign language acquisition (FLA)*	nabývání, osvojování, akvizice cizího jazyka
	agreement	shoda
	▪ *chance agreement*	▪ náhodná shoda
	annotation	anotace
	▪ *embedded a.*	vkládaná anotace
	▪ *flat a.*	1. plochá anotace, 2. jednorovinná a.
	▪ *inline a.*	1. vkládaná anotace, 2. inline a.
	▪ *linear a.*	1. lineární anotace, 2. jednorovinná a.
	▪ *multi-layer a.*	1. vícerovinná anotace, 2. víceúrovňová a.
	▪ *stand-off a.*	1. distanční anotace, 2. stand-off a.
	annotation format	anotační formát
	annotation layer	rovina anotace
	annotation model	anotační model
	annotation scheme	anotační schéma
	annotation span	anotační interval
	▪ *conflicting annotation span*	▪ anotační překryv
	annotator	anotátor
	appropriateness: error of a.	1. přiměřenost, 2. vhodnost
	aproximative system	aproximativní systém
	authoritative interpretation	1. autoritativní interpretace, 2. směrodatná i.
	avoidance	1. neužívání, 2. vyhýbání
B	*backlash*	zpětné přenesení
	built-in sylabus	zabudovaný sylabus
C	*chunk*	úsek textu
	classroom-oriented	zaměřený na vyučování
	coeficient of agreement	koeficient shody
	comparative fallacy	srovnávací omyl
	competition model	kompetiční model
	computer aided error analysis	počítačem podporovaná chybová analýza
	computer-mediated student interaction	1. počítačem podporovaná studentská interakce, 2. interakce studentů prostřednictvím počítače
	content analysis	obsahová analýza
	contrastive analysis	kontrastivní analýza
	Contrastive Analysis Hypothesis (CAH)	hypotéza kontrastivní analýzy (CAH)

	contrastive interlanguage analysis	kontrastivní analýza mezijazyka
	corpus-based	založený na korpusu
	corpus-driven	korpusem řízený
	co-text	ko-text
	cross-linguistic influence	mezijazykové působení
	cross-sectional corpus	průřezový korpus
D	*data architecture*	datová architektura
	dispreffered form	nepreferovaná forma
	double marking	dvojité označení
E	*emendation*	1. emendace, 2. oprava
	encoding formats	kódovací formáty
	error	1. systémová chyba, 2. chyba
	▪ *absolute e.*	▪ prostá chyba
	▪ *communication-strategy e.*	▪ chyba v komunikační strategii
	▪ *developmental e.*	▪ vývojová chyba
	▪ *induced e.*	▪ 1. vynucená chyba, 2. indukovaná chyba
	error analysis	chybová analýza
	error exponent	1. nositel chyby, 2. chybový exponent
	error gravity	závažnost chyby
	error weightening	míra závažnosti chyb
	exposure to a language	1. vystavení jazyku, 2. působení jazyka na (žáka)
	external data (metadata)	1. metadata, 2. externí data, 3. správní anotace
F	*false concept*	falešný koncept
	fluency	plynulost
	foreign language acquisition	1. nabývání cizího jazyka, 2. osvojování c. j.
	foreign language teaching	vyučování cizímu jazyku
	foreign-soundingness	aspekt cizosti
G	*goofs*	chyby
	grammatical tagging	gramatické značkování
H	*hierarchy of difficulty*	hierarchie obtížnosti
I	*idiosyncratic dialect*	idiosynkratický dialekt
	inhibition	inhibice
	▪ *proactive i.*	proaktivní inhibice
	▪ *retroactive i.*	retroaktivní inhibice
	initial state	počáteční stav
	input	1. vstup, 2. input
	intake	1. zpracovaný input, 2. příjem
	inter-annotator agreement	mezianotátorská shoda
	interference	interference
	▪ *inhibitive i.*	inhibiční i.
	▪ *intrusive i.*	1. zjevná i., 2. rušivá i.
K	*keyboarding*	přepisování klávesami
L	*langage acquisition device (LAD)*	1 modul jazykového vývoje, 2. ústrojí osvojování jazyka
	learneability	naučitelnost
	learner	žák
	learner corpus	žákovský korpus
	learner corpus research	výzkum založený na žákovském korpusu
	learner language	1. žákovský jazyk, 2. jazyk žáků
	learning	učení (se)
	▪ *language learning*	jazykové učení
	learning strategy	strategie učení
	▪ *exploiting redundancy*	1. zneužití redundance, 2. zanedbání gramatického prvku

	▪ *false analogy*	mylná analogie
	▪ *hypercorretion*	hyperkorektnost
	▪ *incomplete rule application*	aplikace nekompletního pravidla
	▪ *misanalysis*	chybná hypotéza
	▪ *overgeneralization*	přílišné zobecnění
	▪ *overlooking co-occurance restrictions*	ignorance současně se vyskytujících omezení
	linguistic annotation	lingvistická anotace
M	*Markedness Differential Hypothesis*	hypotéza diferenciálu příznakovosti
	markedness theory	teorie příznakovosti
	mark-up	značkovat, (formátové)značkování
	metalinguistic awareness	metajazykové povědomí
	mistake	nesystémová chyba
	multimodal corpora	multimodální korpus
N	*native language*	rodný jazyk
	native speaker (NS)	rodilý mluvčí
	native speaker norm	norma rodilého mluvčího
	natural order hypothesis	hypotéza přirozené posloupnosti
	non-native speaker (NNS)	nerodilý mluvčí
	non-nativness	nerodilost
O	*order of acquisition*	1. posloupnost nabývání, 2. posloupnost akvizice
	output	1. výstup, 2. output
	overregularization	přílišné zpravidelnění
	over-teaching	opětovný výklad
	overuse	nadužívání
P	*performance analysis*	performanční analýza
	plain (raw) text	1. holý text, 2. surový text
	processing: langauge p.	zpracování (jazyka)
	proof-reader	1. korektor, 2. supervizor
	pseudoprocedure	pseudoprocedura
	purpose-oriented	účelový
Q	*quasi-longitudinal corpus*	kvazilongitudinální korpus
R	*register*	registr
	reliability	1. spolehlivost, 2. reliabilita
	reusable	opakovaně použitelný
S	*sample*	1. vzorek, 2. sonda
	self-correction	1. sebeoprava, 2. sebekorekce
	self-repair	1. vlastní rektifikace, 2. sebeotrava, 3. sebekorekce
	string	řetězec (slov)
	▪ *discontinuous s.*	▪ nespojitý řetězec
	▪ *spanning string*	▪ větvící se řetězec
	superficial alternation	1. povrchová alternace, 2. povrchová modifikace
	surface strategy	povrchová realizace
	▪ *addition*	▪ 1. nadbytečné užití, 2. přidání
	▪ *blends*	▪ 1. blendy, 2. směšování, 3. hybridizace
	▪ *misformation*	▪ chybný tvar
	▪ *misordering*	▪ 1. chybné pořadí, 2. chybný slovosled
	▪ *misselection*	▪ chybný výběr
	▪ *ommission*	▪ 1. vynechání, 2. chybějící element
	systematicity	systematičnost
T	*tag*	1. značka, 2. tag
	▪ *error tag*	▪ chybová značka
	tagging	1. značkování, 2. tagování

	▪ *error tagging*	▪ chybové značkování
	target hypothesis	1. cílová hypotéza, 2. rekonstrukční hypotéza
	target modification	povrchová modifikace
	task	úloha
	task conditions	realizační faktory
	taxonomy	taxonomie
	▪ *communicative effect t.*	▪ t. dle komunikačního efektu
	▪ *comparative t.*	▪ komparativní t.
	▪ *error t.*	▪ chybová t.
	▪ *generic error t.*	▪ komplexní chybová t.
	▪ *surface strategy t.*	▪ t. dle povrchové realizace
	▪ *target modificaton t.*	▪ t. dle povrchové modifikace cílového jazyka
	telecollaborative corpus	telekolaborativní korpus
	token	1. slovní výskyt, 2. token
	transcriber	přepisovač
	transfer	transfer (jazykový t.)
	▪ *transfer analysis*	▪ analýza transferu
	▪ *borrowing t.*	▪ zpětný t.
	▪ *negative t.*	▪ negativní t.
	▪ *positive t.*	▪ pozitivní t.
	▪ *substratum t.*	▪ transfer základu
	transitional competence	přechodná kompetence
	transitional form	přechodná forma
	translation equivalence	ekvivalence překladu
U	*underuse*	1. podužívání, 2. nedostatečné užívání
	usage	úzus
	use	užití
	▪ *accurate u.*	▪ správné užití
V	*validity*	1. platnost, 2. validita
	variables	proměnné
	version	verze
	▪ *moderate v. (CAH)*	▪ umírněná v.
	▪ *strong v.*	▪ silná v.
	▪ *weak v.*	▪ 1. redukovaná v., 2. slabá v.
W	*well-formedness*	1. správnost, 2. správná utvářenost
	writing	psaní
	▪ *calligraphy*	▪ písmo
	▪ *orthography*	▪ pravopis

BIBLIOGRAFIE

AIJMER, K.; ALTENBERG, B.; JOHANSSON, M. (eds). *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies Lund 4-5 March 1994*. Lund: Lund University Press, 1996.

ARTSTEIN, R.; POESIO, M. Inter–coder agreement for computational linguistics. *Computational Linguistics.* 2008, vol. 34, no. 4, s. 555–596. Dostupný z WWW: http://www.aclweb.org/anthology/J/J08/J08–4004.pdf

ATKINS, S.; CLEAR, J.; OSLER, N. *Corpus design criteria.* 1991. Dostupný z WWW: http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf

ATWELL, E.; HOWARTH, P.; SOUTER, C. The ISLE Corpus: Italian and German Spoken Learners English. *ICAME Journal.* 2003, no. 27, s. 5–18.

AXELSSON M. W. Project USE (Uppsala Student English). *ASLA Information.* 1999, vol. 25, no. 2, s. 25–26.

BARTNING, I. The advanced learner varieties: ten years later. In MYLES, F.; LABEAU, E. (eds.). *The advanced learner varieties: The case of French L2.* Bern : Peter Lang, 2009, s. 11–40.

BELL, R. Error analysis*:* a recent pseudo procedure in applied linguistics. *ITL Review of Applied Linguistics.* 1974, s. 25–35.

BELZ, J. A.; REINHARDT, J.; RINE E. *Telekorp: The telecollaborative learner corpus of English and German.* 2005.

BIRD, S. An Integrated Framework for Treebanks and Multilayer Annotations. In *Proceedings of the Third International Conference on Language Resources and Evaluation.* Paris : European Language Resources Association, 2002.

BIRD, S.; LIBERMAN, M. 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, 1999, s. 1–10. Dostupný z WWW: www.ldc.upenn.edu/acl/W/W99/W99–0301.pdf

BLEY–VROMAN, R. The comparative fallacy in interlanguage studies: the case of systematicity. *Language Learning.* 1983, vol. 33, s. 1–17.

BLEY–VROMAN, R. The logical problem of foreign language learning. *Linguistic Analysis.* 1990, vol. 20, s. 3–49.

BOYD, A. EAGLE: an Error–Annotated Corpus of Beginning Learner German. In *Proceedings of LREC 2010.* ELRA : Malta, 2010. Dostupný z WWW: http://www.ling.ohio–state.edu/~adriane/papers/boyd–lrec–2010.pdf

BRAND, CH.; KÄMMERER, S. The Louvain International Database of Spoken English Interlanguage (LINDSEI): Compiling the German component. In BRAUN, S., KOHN, K., MUKHERJEE, J. (eds.). *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main : Peter Lang, 2006, s. 127–140.

BRILL, E. Transformation–based error–driven learning and natural language processing: A case study in part–of–speech tagging. *Computational Linguistics*. 1995, vol. 21, no. 4, s. 543–566.

BROWN, D. H. *Principles of Language Learning and Teaching.* Prentice–Hall Regents, 1987.

BURNARD, L.; BAUMAN, S. (eds.) *Guidelines for Electronic Text Encoding and Interchange (TEI P5).* The TEI Consortium, 2007. Dostupný z WWW: *http://www.tei–c.org/Guidelines/P5*

CARLETTA, J. C. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*. 1996, vol. 22, no. 2, s. 249–254. Dostupný z WWW: http://acl.ldc.upenn.edu/J/J96/J96–2004.pdf

CARLETTA, J. C.; MCKELVIE, D.; ISARD, A. Supporting linguistic annotation using XML and stylesheets. In SAMPSON, G.; MCCARTHY, D. (eds.) *Corpus linguistics: readings in a widening discipline*. London & New York : Continuum Interpretations, 2002. Dostupný z WWW: homepages.inf.ed.ac.uk/jeanc/revised.but.like.readings–in–corpling.pdf

CARLETTA, J. et al. 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 (NLPXML–2003)), Budapest, 2003.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, vol. 20, no. 1, s. 37–46.

COOK, V. *Linguistics and second language acquisition.* New York : St. Martin's press, 1993.

COOK, V.; BASSETTI, B. (eds.) *Second Language Writing Systems*. Clevedon : Multilingual Matters, 2005.

CORDER, S. P. *Error Analysis and Interlanguage.* Oxford : Oxford University Press, 1981.

CORDER, S. P. Error analysis, Interlanguage and second language acquisition. *Language Teaching and Linguistics Abstracts.* 1975, vol. 8, s. 201–218.

CORDER, S. P. Idiosyncratic Dialects and Error Analysis. In RICHARDS, J. (ed.). *Error analysis: Perspectives on Second Language Acquisition.* Essex : Longman, 1974, s. 158–171.

CORDER, S. P. The Significance of Learner´s Errors. *IRAL.* 1967, vol. 5, no. 4, s. 161–170.

ČERMÁK, F. Korpus, informace a lingvistika. In *Přednášky z XLVIII. běhu LŠSS UK*. Praha : Karolinum, 2005. s. 19–20.

ČERMÁK, F., BLATNÁ R. (eds.) *Manuál lexikografie.* Jinočany : H&H, 1995.

ČERMÁK, F.; SCHMIEDTOVÁ, V. Český národní korpus: základní charakteristika a širší souvislosti. *Národní knihovna*. 2004, sv. 15, č. 3, 2004, s. 152–168.

DAGNEAUX, E.; DENNESS, S.; GRANGER, S. Computeraided error analysis. *System,* 1998, vol. 26, s. 163–174.

DANEŠ, F. a kol. *Český jazyk na přelomu tisíciletí*. Praha : Academia, 1997.

DAVIES, E. E. Error evaluation: the important viewpoint. *ELT Journal,* 1983, vol. 37, no. 4, s. 304–311.

DE FELICE, R.; PULMAN, S. G. A classifier–based approach to preposition and determiner error correction in L2 English. In *COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1.* Stroudsburg : Association for Computational Linguistics, 2008, s. 169–176. Dostupný z WWW: http://portal.acm.org/citation.cfm?id=1599103

DE HAAN, P. The optimum corpus sample size? In LEITNER, G. (ed.) *New Directions in English Language Corpora.* Berlin/New York : Mouton de Gruyter, 1992, s. 3–19.

DI EUGENIO, B. On the usage of Kappa to evaluate agreement on coding tasks. In *Proceedings of the Second International Conference on Language Resources and Evaluation,* 2000, s. 441– 444. Dostupný z WWW: http://www.cs.uic.edu/~bdieugen/PS–papers/lrec00.pdf

DI EUGENIO, B.; GLASS, M. The Kappa Statistic: A Second Look. *Computational Linguistics.* 2004, vol. 30, no. 1, s. 95–101. Dostupný z WWW : http://www.mitpressjournals.org/doi/abs/10.1162/089120104773633402.

DÍAZ–NEGRILLO, A.; FERNÁNDEZ–DOMÍNGUEZ, J. Error Tagging Systems for Learner Corpora. *Resla.* 2006, no. 19, s. 83–102.

DICKINSON, M.; MEURERS, W. D. Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05).* 2005, s. 322–329. Dostupný z WWW: http://aclweb.org/anthology/P05–1040.

DICKINSON, M.; RAGHEB, M. Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT–8).* Milan, 2009. Dostupný z WWW: http://jones.ling.indiana.edu/_mdickinson/papers/dickinson–ragheb09.html.

DIPPER, S. XML–based Stand–off Representation and Exploitation of Multi–Level Linguistic Annotation. In ECKSTEIN, R.; TOLKSDORF, R. (eds.) *Proceedings of Berliner XML Tage*. 2005, s. 39–50. Dostupný z WWW: http://www.ling.uni–potsdam.de/~dipper/papers/xmltage05.pdf

DOOLITTLE, S. Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basie–renden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenen– Architektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner. Magisterarbeit HU–Berlin, 2009.

DOUGHTY, C. J.; LONG, M. (eds.) *The Handbook of Second Language Acquisition.* Oxford : Blackwell, 2004.

DULAY, H.; BURT, M. (a) Natural Sequences in child second language acquisition. *Language Learning*. 1974, vol. 24, s. 37–53.

DULAY, H.; BURT, M. (b) You can't learn without goofing. In RICHARDS, J. C. (ed.) *Error analysis.* London : Longman, 1974.

DULAY, H.; BURT, M.; KRASHEN, S. *Language Two.* Oxford : OUP, 1982.

DUŠKOVÁ, L. On Sources of Errors in Foreign Language Learning. *IRAL.* 1969, vol. 7, s. 11–36.

DUŠKOVÁ, L. Similarity – An aid or hindrance in foreign language learning? *Folia Linguistica.* 1984, vol. 18, s. 103–115.

ECKMAN, F. Markedness and the contrastive analysis hypothesis. *Language Learning.* 1977, vol. 27, s. 315–330.

EEG–OLOFSSON, J.; KNUTSSON, O. Automatic Grammar Checking for Second Language Learners – the Use of Prepositions. In *Nodalida,* 2003. Dostupný z WWW: http://www.csc.kth.se/tcs/projects/xcheck/rapporter/eegolofsson_knutsson03.pdf

ELLIS, R. The Effects of Linguistic Environment on the Second Language Acquisition of Grammatical Rules. *Applied Linguistics.* 1988, vol. 9, no. 3, s. 257–274.

ELLIS, R. *The Study of Second Language Acquisition*. Oxford : OUP, 1994.

ELLIS, R.; BARKHUIZEN, G. *Analysing learner language.* Oxford : OUP, 2009.

Encyklopedický slovník češtiny. NLN, 2002.

FAERCH, C. Performance analysis of learner's language. In GREGERSEN, K (ed.) *Papers from the Fourth Scandinavian Conference of Linguistics.* Odense : Odense UP, 1978, s. 87–95.

FAERCH, C.; KASPER, C. Perspectives on Language Transfer. *Applied Linguistics.* 1987, vol. 8, no. 2, s. 111–136.

FELDMAN, A.; ABUHAKEMA, G.; FITZPATRICK, E. ARIDA: An Arabic Interlanguage Database and Its Applications: A Pilot Study. In *Proceedings of the 21th International Florida Articial Intelligence Research Society Conference (FLAIRS–08).* Coconut Grove, FL : AAAI Press, 2008. Dostupný z WWW: http://chss.montclair.edu/~feldmana/publications/flairs–2008.pdf

FISIAK, J. Contrastive linguistics and foreign/second language acquisition. In GÖBEL, W., SEEBER, V. (eds.) *Anglistentag 1992 Stuttgart*. Tübingen : Niemeyer, 1993, s. 315–326.

FITZGERALD, E. C. *Reconstructing Spontaneous Speech*. Ph.D. Dissertation. Johns Hopkins University : Baltimore, 2009.

FITZPATRICK, E.; SEEGMILLER, M. S. The Montclair electronic language database project. In CONNOR, U., UPTON, T. A. (eds.) *Applied Corpus Linguistics: A Multidimensional Perspective.* Rodopi, 2004, s. 223–238. Dostupný z WWW: http://chss.montclair.edu/linguistics/MELD

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin.* 1971, vol. 76, s. 378–382.

FOSTER, P.; OHTA, A. S. Negotiation for meaning and peer assistance in second language classrooms. *Applied Linguistics.* 2005, vol. 26, s. 402–430.

GASS, S. M.; SELINKER, L. *Second Language Acquisition: An Introductory Course.* Routledge, 2008.

GILQUIN, G.; GRANGER, S.; PAQUOT, M. Learner corpora: the missing link in EAP pedagogy. In THOMPSON, P. (ed.) *Corpus–based EAP Pedagogy. Special issue of Journal of English for Academic Purposes*. 2007, vol. 6, no. 4, s. 319–335.

GRADMAN, H. The Limitations of Contrastive Analysis Prediction. *PCCLLU Papers.* 1971, vol. 3, no. 4.

GRANGER, S. (a) Error–tagged Learner Corpora and CALL: A Promising Synergy. *CALICO journal.* 2003, vol. 20, no. 3, s. 465–480.

GRANGER, S. (b) The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly.* 2003, vol. 37, no. 3, s. 538–545.

GRANGER, S. A Birds–eye view of learner corpus research. In GRANGER, S.; HUNG, J.; PETCH–TYSON, S. (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Université Catholique de Louvain, 2002, s. 3–35.

GRANGER, S. Computer Learner Corpus Research: Current Status and Future Prospects. In CONNOR, U.; UPTON, T. A. (eds.) *Applied Corpus Linguistics. A Multidimensional Perspective.* Amsterdam : Rodopi, 2004, s. 123–146.

GRANGER, S. Learner Corpora. In LÜDELING, A.; KYTÖ, M. (eds.) *Corpus Linguistics. An International Handbook.* Eds. HSK 29. 1. VOL. 1. Berlin/New York : Mouton De Gruyter, 2008, s. 259–274.

GRANGER, S. The computer learner corpus: a versatile new source of data for SLA research. In GRANGER, S. (ed.) *Learner English on Computer.* London: Longman, 1998, s. 3–19.

GRANGER, S. The International Corpus of Learner English. In AARTS, J. P.; DE HANN, P.; OOSTDIJK, N. (eds.) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam : Rodopi, 1993, s. 57–69.

GRANGER, S.; HUNG, J.; PETCH–TYSON, S. (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Université Catholique de Louvain, 2002.

HÁDKOVÁ, M. *Čeština z druhé strany*. Ústí n. L.: UJEP, 2008.

HÁDKOVÁ, M. Gramatika a čeština pro cizince. In *Usta ad Albim, Bohemica*, 2011, č. 1, s. 104–117.

HAMMARBERG, B. *Introduction to the ASU Corpus, a Longitudinal Oral and Written Text Corpus of Adult Learners' Swedish with a Corresponding Part from Native Swedes.* Stockholm University, Department of Linguistics, 2010. Dostupný z WWW: http://www–test.ling.su.se/polopoly_fs/1.13705.1302078209!/Introduction_to_the_ASU_Corpus.pdf

HAMMERLY, H. *Fluency and accuracy*. Philadelphia : Multilingual Matters, 1991.

HAN, N. R.; CHODOROW, M.; LEACOCK, C. Detecting errors in English article usage by non–native speakers. *Natural Language Engineering*.2006, vol. 12, no. 2, s. 115–129. Dostupný z WWW: http://www.cs.pitt.edu/~litman/courses/slate/pdf/nle06–HCL.pdf

HAN, N. R.; TETREAULT, J.; LEE, S. H.; HA, J. Y. Using an Error–Annotated Learner Corpus to Develop an ESL/EFL Error Correction System. In *Proceedings of the LREC 2010.* Malta : ELRA, 2010. Dostupný z WWW: http://www.lrec–conf.org/proceedings/lrec2010/summaries/821.html

HANA, J.; ROSEN, A.; ŠKODOVÁ, S.; ŠTINDLOVÁ, B. Error–tagged Learner Corpus of Czech. In *Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV).* Uppsala : Association for Computational Linguistics&Uppsala University, 2010, s. 11–19. Dostupný z WWW: www.aclweb.org/anthology/W/W10/W10–18.pdf

HASSELGREN**,** A. The EVA Corpus of Norwegian School English. *ICAME Journal*. 1997, vol. 21, s. 123–124.

HAYES, A. F.; KRIPPENDORFF, K. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures 1*. 2007, vol. 1, s. 77–89. Dostupný z WWW: http://www.unc.edu/courses/2007fall/jomc/801/001/HayesAndKrippendorff.pdf

HENDRICH, J. et al. *Didaktika cizích jazyků.* Praha : SPN, 1988.

HLADKÁ, Z. Zkušenosti s tvorbou korpusů češtiny v ÚČJ FF MU v Brně. In *SPFFBU, A.* r. 53, 2005, s. 115–124.

HOVY, E. *Annotation – A Tutorial.* Presented at the 48th Annual Meeting of the Association for Computational Linguistics, July 11. Uppsala : Uppsala Universitet, 2010.

HUGHES, A.; LASCARATOU, Ch. Competing criteria for error gravity. *ELT J*, 1982, vol. 36, no. 3, s. 175–182.

CHODĚRA, R. *Didaktika cizích jazyků.* Praha : Academia, 2006.

CHOMSKY, N. A Review of B. F. Skinner's ´Verbal Behavior´. *Language.* 1959, vol. 35, no. 1, s. 26–58.

CHOMSKY, N. *Language and Mind.* New York : Harcourt Brace Jovanovich, 1972.

CHRÁSTKA, M. *Metody pedagogického výzkumu.* Praha : GRADA Publishing, 2007.

IZUMI, E; ISAHARA, H. (a) Investigation into Language Learners' Acquisition Order Based on an Error Analysis of a Learner Corpus. In *IWLeL 2004: An Interactive Workshop on Language e–Learning*. 2004, s. 63–71. Dostupný z WWW: http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/1396/1/07.pdf

IZUMI, E.; UCHIMOTO, K.; ISAHARA, H. (b) The NICT JLE Corpus Exploiting the language learners' speech diabase for research and education. *International Journal of The Computer, the Internet and Management*. 2004, vol. 12, no. 2, s. 119 –125.

IZUMI, E.; UCHIMOTO, K.; ISAHARA, H. (c) SST *Speech Corpus* of *Japanese Learners*' *English and Automatic* Detection of Learners' Errors. *ICAME Journal*. 2004, no. 28, s. 31–48. Dostupný z WWW: http://icame.uib.no/ij28/index.html

IZUMI, E.; UCHIMOTO, K.; ISAHARA, H. Error Annotation for Corpus of Japanese Learner English. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005).* Korea, 2005, s. 71–80. Dostupný z WWW http://acl.ldc.upenn.edu/I/I05/I05–6009.pdf

IZUMI, E.; UCHIMOTO, K.; SAIGA, T.; SUPNITHI, T.; ISAHARA, H. Automatic Error Detection in the Japanese Learners' English Spoken Data. In *Proceedings of the ACL.* Sapporo, Japan, 2003. Dostupný z WWW: http://www.aclweb.org/anthology/P/P03/P03–2026.pdf

JAMES, C. *Errors in Language Learning and Use.* Longman, 1998.

JAMES, C. Judgements of error gravities. *ELT Journal.* 1977, vol. 31, no. 2, s. 116–124.

JAMES, C. The Exculpation of Contrastive Linguistics. In NICKEL, G. (ed.) *Papers in Contrastive Linguistics.* Cambridge : CUP, 1971, s. 53–68.

JANTUNEN, J. H. ICLFI: Overview of the project and corpus. *ASKeladden Network Meeting*. Bergen, 2010. Dostupný z WWW: http://www.oulu.fi/hutk/sutvi/oppijankieli/tutkimus/Bergen%20ASKelladden.pdf

JELÍNEK, T. Post–processing aneb automatické zpracování manuálně anotovaného chybového korpusu. Přednáška pronesená na workshopu *Moderní technologie a didaktika jazyka*, 8.10.2010. Dostupný z WWW: www.c2j.cz

JEŽKOVÁ, P. *Práce s chybou ve výuce cizích jazyků na příkladech ruského jazyka.* 2008. Bakalářská práce, MU, Brno.

JOHNSON, K. Mistake correction. *ELT Journal.* 1998, vol. 42, no. 2, 1988, s. 89–97.

JOHNSON, M. *A Philosophy of Second Language Acquisition.* New Haven : Yale University Press, 2004.

JORDAN, G. *Theory construction in second language acquisition*. Amsterdam : Benjamins, 2004.

KELLERMAN, E. Transfer and Non–Transfer: Where are we now? *Studies in Second Language Acquisition.* 1979, vol. 2, s. 37–57.

KORMOS, J. A new psycholinguistic taxonomy of self–repairs in L2: A qualitative analysis with retrospection. *Even Yearbook, ELTE SEAS Working Papers in Linguistics*. 1998, vol. 3, s. 43–68.

KOWAL, I. What do false–starts and self repairs tell us about narrative structure? *Psychology of Language and Communication.* 1999, vol. 3, no. 1, s. 75–82. Dostupný z WWW: http://plc.psychologia.pl/plc/contents/fulltext/03–1_6.pdf

KRASHEN, S. D. *Second Language Acquisition and Second Language Learning.* Pergamon Press, 1981.

KRASHEN, S. *Principles and Practice in Second Language Acquisition.* University of Southern California, 2009. Dostupný z WWW: http://www.sdkrashen.com/Principles_and_Practice/index.html

KRASHEN, S. Some issues relating to the Monitor Model. In BROWN, H. D.; YORIO, C.; CRYMES, R. (eds.) *On TESOL '77: Teaching and Learning English as a Second Language: Trends in Research and Practice.* Washington : TESOL, 1977, s. 144–158.

KRIPPENDORFF, K. Reliability in content analysis. some common misconceptions and recommendations. *Human Communication Research*. 2004, vol. 30, no. 3, s. 411–433. Dostupný z WWW: http://dx.doi.org/10.1111/j.1468–2958.2004.tb00738.x

KRIPPENDORFF, K. Reliability. In *Content Analysis: An Introduction to its Methodology.* Beverly Hills / London : Sage Publications, 1980, s. 211–251.

KULIČ, V. *Chyba a učení: funkce chybného výkonu v učení a v jeho řízení.* Praha : SPN, 1971.

KWON, E. Y. The "Natural Order" of Morpheme Acquisition: A Historical Survey and Discussion of Three Putative Determinants. *Columbia University Working Papers in TESOL & Applied Linguistics.* 2005, vol. 5, no. 1, s. 1–21. Dostupný z WWW: http://journals.tc–library.org/index.php/tesol/article/download/112/110

KWON, H. The SNU Korean Learner Corpus of English: Compilation and Application. *English Language and Linguistics.* 2009, r. 28, s. 203–228.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*. 1977, vol. 33, s. 159–74.

LARSEN–FREEMAN, D. Language acquisition and language use from a chaos / complexity theory perspective. In KRAMSCH, C. (ed.). *Language acquisition and socialization.* London : Continuum International Publishing Group, 2002, s. 33–46.

LARSEN–FREEMAN, D.; LONG, M. H. *An Introduction to Second Language Acquisition Research.* London/New York : Longman, 1992.

LEACOCK, C.; CHODOROW, M.; GAMON, M.; TETREAULT, J. Automated Grammatical Error Detection for Language Learners. Morgan & Claypool , 2010.

LEE, J.; SENEFF, S. Automatic Grammar Correction for Second–Language Learners. In *INTERSPEECH 2006 – ICSLP*. 2006, s. 1978–1981. Dostupný z WWW: http://groups.csail.mit.edu/sls/publications/2006/IS061299.pdf.

LEECH, G. Adding Linguistic Annotation. In WYNNE, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford : Oxbrow Books, 2005, s. 17–29. Dostupný z WWW: http://www.ahds.ac.uk/guides/linguistic–corpora/chapter2.htm

LEECH, G. Introducing corpus annotation. In GARSIDE, R.; LEECH, G.; MCENERY, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London : Longman, 1997, s. 1–18.

LEECH, G. Preface. In GRANGER, S. (ed.) *Learner English on Computer*. Longman, 1998, s. xiv–xx.

LENNON, P. Error: Some Problems of Definition, Identification, and Distinction. *Applied Linguistics.* 1991, vol. 12, no. 2, s. 180–196.

LEVELT, W. Monitoring and self–repair in speech. *Cognition*. 1983, vol. 14, no. 1. s. 41–104.

LIGHTBOWN, P.; SPADA, N. *How Languages are Learned.* Oxford, New York etc. : Oxford UP, 2007.

LITTLEWOOD, W. *Foreign and Second Language Learning: Language Acquisition Research and its Implications for the Classroom.* Cambridge : CUP, 1984.

LOMBARD, M.; SNYDER–DUCH, J.; BRACKEN, C. C. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*. 2002, vol. 28, s. 587–604. Dostupný z WWW: http://public.univie.ac.at/fileadmin/user_upload/lehrstuhl_marketing/Lehre/Lehrinhalte/09SS/Lombard_et_al__2002_–1.pdf

LOMBARD, M.; SNYDER–DUCH, J.; BRACKEN, C. C. *Practical resources for assessing and reporting intercoder reliability in content analysis research projects.* 2003. Dostupný z WWW: http://www.temple.edu/mmc/reliability.

LONG, M. H.; SATO, C. J. Methodological issues in interlanguage studies: an interactionist perspective. In DAVIES, A.; CRIPER, C.; HOWATT, A. P. R. (eds.) *Interlanguage.* Edinburgh : Edinburgh University Press, 1984, s. 253–280.

LU, X. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*. 2010, vol. 15, no. 4, s. 474–496. Dostupný z WWW: http://www.personal.psu.edu/xxl13/papers/Lu_inpress_ijcl.pdf

LÜDELING, A. et al. Multi–level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005 Conference, 14–17 July.* Birmingham, 2005. Dostupný z WWW: http://www.corpus.bham.ac.uk/pclc/#corpora

LÜDELING, A. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In GROMMES, P.; WALTER, M. *Fortgeschrittene Lernervarietäten*. Niemeyer : Tübingen, 2008, s. 119–140.

LÜDELING, A.; DOOLITTLE, S.; HIRSCHMANN, H.; SCHMIDT, K.; WALTER, M. Das Lernerkorpus Falko. *Deutsch als Fremdsprache.* 2008, no. 2, s. 67–73.

MAICUSI, T.; MAICUSI, P.; CARRILLO LÓPEZ, M. J. The Error in the Second Language Acquisition. *Encuentro.* 1999–2000, vol. 11. Dostupný z WWW: http://dspace.uah.es/jspui/bitstream/10017/953/1/The%20Error%20in%20the%20Second%20Language%20Acquisition.pdf

MAJID HAYATI, A. Contrastive linguistics: Re–evaluation and re–formulation. *Papers and Studies in Contrastive Linguistics.* 1997, vol. 32, s. 21–28. Dostupný z WWW: http://ifa.amu.edu.pl/psicl/

MATSUOKA, R., EVANS, R. Socio–Cognitive Approach in Second Language Acquisition Research. *J Nurs Studies.* 2004, vol. 3, no. 1, s.2–10.

McCRETTON, E.; RIDER, N. Error gravity and error hierarchies. *IRAL.* 1993, vol. 31, no. 3, s. 177–188.

MENEZES, V. L. Fractal Model of Language Acquisition. Dostupný z WWW : http://www.veramenezes.com/model.htm

MEUNIER, F. Computer Tools for Interlanguage Analysis: A Critical Approach. In GRANGER, S. (ed.) *Learner English on Computer.* London / New York : Addison Wesley Longman, 1998, s. 19–37.

MEURERS, D. On Automatically Analyzing Learner Language: Interpreting Form and Meaning in Context. Invited talk at *Colloquium of the Research Center for English and Applied Linguistics (RCEAL),* 8.2.2011. University of Cambridge, 2011. Dostupný z WWW: http://www.sfs.uni–tuebingen.de/~dm/presentations.html

MEURERS, D. On the Automatic Analysis of Learner Language. Introduction to the Special Issue. *CALICO Journal*. 2009, vol. 26, no. 3, s. 469–473. Dostupný z WWW http://www.sfs.uni–tuebingen.de/~dm/papers/meurers–09.pdf

MEURERS, D. On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua,* 2005, r. 115, no. 11, s. 1619–1639.

MILTON, J.; CHOWDHURY, N. Tagging the interlanguage of Chinese learners of English. In FLOWERDEW, L.; TONG, K. K. (eds.) *Entering Text.* Hong Kong University of Science and Technology : Hong Kong, 1994, s. 127–143.

MUEHLEISEN, V. Introducing the SILS Learners' Corpus: A Tool for Writing Curriculum Development. *Waseda Global Forum* , 2006, č 3. 119–125. Dostupný z WWW: http://dspace.wul.waseda.ac.jp/dspace/handle/2065/11346

NAGATA, R.; MORIHIRO, K.; KAWAI, A.; ISU, N. A feedback–augmented method for detecting errors in the writing of learners of English. In *ACL–44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.* Stroudsburg, 2006, s. 241–248. Dostupný z WWW: http://portal.acm.org/citation.cfm?id=1220175&picked=prox&cfid=18259668&cftoken=35491816

NESSELHAUF, N. *Collocations in a Learner Corpus.* Amsterdam : John Benjamins, 2005.

NESSELHAUF, N. Learner corpora and their potential in language teaching. In SINCLAIR, J. (ed.) *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia :Benjamins, 2004, s. 125–152.

NICKEL, G. Some controversies in present day error analysis. *IRAL.* 1989, vol. 27, s. 293–305.

NICHOLLS, D. The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference, 28–31 March.* Lancaster, 2003, s. 572–581. Dostupný z WWW: ucrel.lancs.ac.uk/publications/cl2003/papers/nicholls.pdf

OLLER, J.; ZIAHOSSEINY, S. The contrastive analysis hypothesis and spelling errors. *Language Learning*. 1970, vol. 20, s. 183–189.

OTT, N.; ZIAI, R. Evaluating Dependency Parsing Performance on German Learner Language. In *Proceedings of TLT9*, NEALT Proceeding Series. 2010, s. 175–186. Dostupný z WWW: http://dspace.utlib.ee/dspace/bitstream/10062/15960/1/tlt9_submission_28.pdf

PALA, K. Informační technologie a korpusová lingvistika (1). *Zpravodaj ÚVT MU.* 1996, sv. 4, č. 3, s. 8–11. Dostupný z WWW: http://www.ics.muni.cz/zpravodaj/articles/58.html

PASSONNEAU, R. J.; HABASH, N.; RAMBOW, O. Inter–annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. 2006. Dostupný z WWW: http://www1.ccls.columbia.edu/~beck/pubs/iamtc–lrec06.pdf

PÉRY–WOODLEY, M. P. Contrasting discourses: contrastive analysis and a discourse approach to writing. *Language Teaching.* 1990, vol. 23, s. 143–151.

PIEPHO, H. E. *Kommunikative Kompetenz als ubergeordnetes Lernziel im Englischunterricht.* Limburg : Frankonius, 1973.

POLIO, CH. Mesure of Linguistic Accuracy in Second Language Writing. *Language Learning.* 1997, vol. 47, s. 101–143. Dostupný z WWW: http://people.ucsc.edu/~ktellez/poliowritingmeasures.pdf

POULISSE, N. *The Use of Compensatory Strategies by Dutch Learners of English.* Berlin : Mouton de Gruijter, 1990.

POWELL, G. What is the Role of Transfer in Interlanguage? *CRILE Working Papers*. 1998, vol. 33, Lancaster University. Dostupný z WWW: http://www.ling.lancs.ac.uk/groups/crile/workingpapers.htm

PRAVEC, N. A. Survey of learner corpora. *ICAME Journal*. 2002, no. 26, s. 81–114. Dostupný z WWW: http://icame.uib.no/ij26/pravec.pdf

RANDALL, M.; GROOM, N. The BUiD Arab Learner Corpus: a resource for studying the acquisition of L2 English spelling. In MAHLBERG, M.; GONZÁLEZ–DÍAZ, V.; SMITH, C. (eds.) *Proceedings of the Corpus Linguistics Conference, 20–23 July 2009.* University of Liverpool, 2009. Dostupný z WWW: http://ucrel.lancs.ac.uk/publications/cl2009/54_FullPaper.doc

RASTELLI, S. Going beyond errors: position and tendency tags in a learner corpus. In SANSÒ, A. (ed.) *Language Resources and Linguistic Theories*. Milano : Franco Angeli, 2007, s. 96–109.

RASTELLI, S. Learner Corpora without Error Tagging. *Linguistic online*. 2009, vol. 38, no. 2. Dostupný z WWW: http://www.linguistik–online.com/38_09/rastelli.html

RASTELLI, S.; FRONTINI, F. SLA meets FLT research: the form/fiction split in the annotation of Learner Corpora. In *Proceedings of TaLC 8*. Lisabon, 2008, s. 446–451.

REIDSMA, D. *Annotations and Subjective Machines of Annotators, Embodied Agents, Users, and Other Humans.* PhD thesis series No. 08–121. Univ. of Twente, 2008.

REIDSMA, D.; CARLETTA, J. Reliability Measurement without Limits. *Computational Linguistics*. 2008, vol. 34, no. 3, s. 319–326. Dostupný z WWW: http://homepages.inf.ed.ac.uk/jeanc/reidsma–and–carletta.CL2008.pdf

REZNICEK, M.; KRUMMES, C.; HIRSCHMANN, H.; LÜDELING, A.; ENSSLIN, A.; CHAN, J. W.; ZELDES, A.; KRAUSE, T.; ZIPSER, F. (b) 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. Dostupný 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. (a) *Das Falko–Handbuch. Korpusaufbau und Annotationen Version 1.0*. Humboldt–Universität zu Berlin, 2010.

RICHARDS, J. (ed.) *Error Analysis: Perspective on Second Language Acquisition.* London : Longman, 1974.

RICHARDS, J. C. Non–Contrastive Approach to Error Analysis. In *Error Analysis: Perspectives on Second Language Acquisition.* Ed. J. C. Richards. London : Longman, 1974, s. 172–188.

RINGBOM, H. *The role of the first language in foreign language learning*. Clevedon & Philadelphia : Multilingual Matters, 1987.

ROSEN, A.; ŠKODOVÁ, S.; ŠTINDLOVÁ, B.; HANA, J. Annotating Foreign Learners' Czech. In *Proceedings of Formal Description of Slavic Languages 8.5.* 25.–27.11.2010, Brno. – v tisku

ROSÉN, V.; SMEDT, K. D. Syntactic Annotation of Learner Corpora. In JOHANSEN, H.; GOLDEN, A.; HAGEN, J. A.; HELLAND, A. K. (eds.) *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60–årsdag* [Systematic, varied, but not arbitrary. Anthology about Norwegian as a second language on the occasion of Kari Tenfjord's 60th birthday]. Oslo : Novus forlag, 2010, s. 120–132. Dostupný z WWW: http://folk.uib.no/hfosm/papers/salc.pdf

ROZOVSKAYA, A.; ROTH, D. Annotating ESL Errors: Challenges and Rewards. In *Proceedings of NAACL'10Workshop on Innovative Use of NLP for Building Educational Applications* University of Illinois at Urbana–Champ, 2010. Dostupný z WWW: http://www.cs.rochester.edu/~tetreaul/naacl–bea5.html

SAVKOV, A.; BECKER, R. ProjectX – A learner Language Annotation Project. *ISCL Hauptsrminar: Exploring the Automatic Analysis of Learner Language*, 2009. Dostupný z WWW: http://www.midapd.com:8085/DisplayX/index.html

SCOTT, W. A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quart.* 1955, vol. 19, s. 321–325.

SEEGMILLER, M. S.; FITZPATRICK, E. Practical aspects of corpus tagging. In LEWANDOWSKA–TOMASZCZYK, B. (ed.) *Palc 2001: Practical Applications in Language Corpora.* Peter Lang Pub Inc, 2003. Dostupný z WWW: http://chss.montclair.edu/linguistics/MELD/Lodzpaper.pdf

SEGALOWITZ, N.; LIGHTBOWN, P. M. Psycholinguistic approaches to SLA. *Annual Review of Applied Linguistics.* 1999, vol. 19, s. 43–63.

SEIDLHOFER, B. Giving VOICE to English as a Lingua Franca. In FACCHINETTI, R.; CRYSTAL, D.; SEIDLHOFER, B. (eds.) *From International to Local English – and Back Again.* Frankfurt : Peter Lang, 2010, s. 147–163.

SELINKER, L. Interlanguage. *IRAL*. 1972, vol. 10, no. 3, s. 209–231.

SEOK, B. J.; SUN–HEE, L.; DICKINSON, M. Annotation of Spelling Errors for Korean Learner. Paper at *Automatic Analysis of Learner Language (AALL'09),* 10.–11.3.2009, Arizona State University. Dostupný z WWW: http://jones.ling.indiana.edu/~mdickinson/presentations/jang–et–al09–2x3.pdf

SHAUGHNESSY, M. P. *Errors and Expectations: A Guide for the Teacher of Basic Writing.* New York: Oxford UP, 1977.

SHIH, H. Compiling Taiwanese Learner Corpus of English. In *Computational Linguistics and Chinese Language Processing 5.2*, s. 89–102. 2000. Dostupný z WWW: http://www.aclclp.org.tw/clclp/v5n2/v5n2a4.pdf

SCHACHTER, J. An Error in Error Analysis. *Language Learning.* 1974, vol. 24, no. 2, s. 205–214.

SCHIFTNER, B. Learner Corpora of English and German: What is their status quo and where are they headed? *Vienna English Working PaperS*, vol. 17, no. 2, s. 47–78.

SCHMID, H. Probabilistic Part–of–Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing.* 1994. Dostupný z WWW: http://www.ims.uni–stuttgart.de/ftp/pub/corpora/tree–tagger1.pdf

SCHMIDT, T. 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, 2001, s. 219–227. Dostupný z WWW: http://www.exmaralda.org/files/IRCS_Paper.pdf.

SCHWARTZ, B.; SPROUSE, R. A. When syntactic theories evolve: Consequences for L2 acquisition research. In MALDEN, J. A. (ed.) *Second Language Acquisition and Linguistic Theory*., MA: Blackwell, 2000, s. 156–186.

SIEMEN, P.; LÜDELING, A.; MÜLLER, F. H. Falko – ein fehlerannotiertes Lernerkorpus des Deutschen. In *Proceedings of Konvens 2006*, Konstanz, 2006. Dostupný z WWW: http://www.linguistik.hu–berlin.de/institut/professuren/korpuslinguistik/forschung/falko

SINCLAIR, J. (a) *Trust the Text: Language, Corpus and Discourse.* London : Routledge, 2004.

SINCLAIR, J. (b) Intuition and annotation – the discussion continues. In AIJMER, K., ALTENBERG, B. (eds.) *Advances in corpus linguistics.* Papers from the 23rd International Conference on English Language Research on Computerized corproa (ICAME 23). Göteborg 22–26 May 2002. Amsterdam/New York : Rodopi, 2004, s. 39–59. Dostupný z WWW: http://books.google.cz/books?id=xAPdODj3i3wC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

SINCLAIR, J. *EAGLES. Preliminary recommendations on Corpus Typology.* EAG–TCWG–CTYP/P. 1996. Dostupný z WWW: http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html

SMITH, B. Methodological Hurdles in Capturing CMC Data: The Case of the Missing Self Repair. *Language Learning & Technology*. 2008, vol. 12, no. 1., s. 85–103.

SOMERS, H. Learner corpora and handwriting. In COOK, V.; BASSETTI, B. (eds.) *Second Language Writing Systems*. Clevedon: Multilingual Matters, 2005, s. 147–163.

Společný evropský referenční rámec pro jazyky. Jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme. Council of Europe, 2001. Dostupný z WWW:

STENSON, N. Induced errors. In SCHUMANN, J. H.; STENSON, N. *New Frontiers of Second Language Learning*. Rowley : Newbury HP, 1974.

STEVENSON, M., GAIZAUSKAS, R. Experiments on Sentence Boundary Detection. In *Proceedings of the Sixth Conference on Applied Natural Language Processing and First Conference of the North American Chapter of the Association for Computational Linguistics*, s. 84–89, 2000. Dostupný z WWW: ftp://ftp.dcs.shef.ac.uk/home/robertg/papers/anlp00–sbd.pdf

STRITAR, M. Slovene as a Foreign Language: The Pilot Learner Corpus Perspective. *Slovenski jezik – Slovene Linguistic Studies.* 2009, 7, s. 135–152. Dostupný z WWW: kuscholarworks.ku.edu/dspace/bitstream/1808/5274/1/8Stritar.pdf

STÜHRENBERG, M. et al. 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, 2006. Dostupný z WWW: http://www.aclweb.org/anthology/W/W06/W06–2715.pdf

STÜHRENBERG, M.; JETTKA, D. A toolkit for multi–dimensional markup: The development of SGF to XStandoff. In . *Proceedings of Balisage : The Markup Conference 2009*. Montreal : Balisage Series on Markup Technologies, 2009. Dostupný z WWW: http://www.balisage.net/Proceedings/vol3/html/Stuhrenberg01/BalisageVol3–Stuhrenberg01.html#d10128e1102

SWAIN, M. Three functions of output in second language learning. In *Principle and Practice in Applied Linguistics.* Cambridge University Press, 1995, s. 125–144.

ŠEBESTA, K. (a) Akviziční korpusy. In *Minulost, přítomnost a budoucnost v jazyce a v literatuře. Ústí nad Labem 1.–3. 9.2010. PF* UJEP : Ústí nad Labem, 2011. – v tisku

ŠEBESTA, K. (b). Čeština cizinců v korpusu. In *Přednášky z 54. běhu LŠSS.* Praha : Filozofická fakulta UK v Praze. – v tisku

ŠEBESTA, K. Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky/Studies in Applied Linguistics*. 2010, roč. 1, č. 2, s. 11–34.

ŠEBESTA, K.; ŠKODOVÁ, S. Žákovský korpus a jeho využití pro češtinu jako druhý jazyk, 2011. – v tisku

ŠEVARJOV, A. P. *Obobščennyje associacii v učebnoj rabote školnika.* Moskva, 1959.

ŠKODOVÁ, S. Možnosti zachycení chyb v tzv. žákovských korpusech. In *Eurolingua&Eurolitteraria 2009.* Ed. O. Uličný. KČL TU : Liberec. 2009, s. 197–204. Dostupný z WWW: http://kcl.fp.tul.cz/cs/sbeevlevo

ŠTINDLOVÁ, B. Žákovský korpus. Budoucnost pro poznávání akvizice cizího jazyka. In *Minulost, přítomnost a budoucnost v jazyce a v literatuře. Ústí nad Labem 1.–3. 9.2010.* PF UJEP : Ústí nad Labem, 2011. – v tisku

ŠTINDLOVÁ, B., LÁBUS, V. Sběr jazykového materiálu pro připravovaný korpus C2J. Přednáška pronesená na workshopu *Žákovské korpusy a možnosti jejich využití ve výuce C2J,* 9.9. 2009. Dostupný z WWW: www.c2j.cz

ŠTINDLOVÁ, B.; ROSEN, A.; PETKEVIČ, V.; JELÍNEK, T. Emendace a chybová anotace žákovských korpusů. Přednáška pronesená na workshopu *Moderní technologie a didaktika jazyka*, 8.10.2010. Dostupný z WWW: www.c2j.cz

ŠTINDLOVÁ, B.; ŠKODOVÁ, S. Žákovské korpusy, CzeSL a čeština jako druhý jazyk. In *Recenzovaný sborník příspěvků vědecké konference s mezinárodní účastí – Sapere Aude 2011.* Hradec Králové: MAGNANIMITAS, 2011.

TAGNIN, S. E. O. A multilingual learner corpus in Brazil. In WILSON, A.; ARCHER, D.; RAYSON, P. (eds.) *Corpus Linguistics Around the World*. Amsterdam/New York : Rodopi, 2006, s. 195–202. Dostupný z WWW: http://www.fflch.usp.br/dlm/comet/artigos/A%20multilingual%20learner%20corpus%20in%20Brazil.pdf

TARONE, E. Some thoughts on the notion of communication strategy. *TESOL Quarterly*. 1981, vol. 15, no. 3, s. 285–295.

TARONE, E.; SWAIN, M. A sociolinguistic perspective on second–language use inimmersion classrooms. *Modern Language Journal.* 1995, vol. 79, s. 166–178.

TAYLOR, G. Errors and explanations. *Applied Linguistics.* 1986, no. 7, s. 144–166.

TENFJORD, K.; MEURER, P.; HOFLAND, K. The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language. [Poster]. *Proceedings from 5th International Conference on Language Resources and Evaluation (LREC).* Genova, 2006.

TETREAULT, J.; CHODOROW, M. Native Judgements of Non– Native Usage: Experiments in Preposition Error Detection. In *COLING Workshop on Human Judgements in Computational Linguistics.* Manchester, 2008, s. Dostupný z WWW: http://portal.acm.org/citation.cfm?id=1611633

THOMAS, J. Using Corpora in Language Teaching and Learning. *Teaching English with Technology, A Journal for Teachers of English.* 2005, vol. 6, no. 1. Dostupný z WWW http://www.iatefl.org.pl/call/j_soft23.htm

TOGNINI–BONELLI, E. *Corpus Linguistics at Work.* Amsterdam/Philadelphia : Benjamins, 2001.

TONO Y. A corpus–based analysis of interlanguage development: Analysing part–of–speech sequences of EFL learner corpora. In LEWANDOWSKA–TOMASZCZYK, B., MELIA, P. J. *PALC'99: Practical Applications in Language Corpora.* Papers from the International Conference at the University of Lódz, 15–18 April 1999. Frankfurt am Main : Peter Lang, 2000, s. 323–340.

TONO, Y. et al. Developing a one–million–word spoken EFL learner corpus. *Japan Association of Language Teaching (JALT) 2001.* Japan, 2001.

TONO, Y. Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 Conference.* Lancaster : United Kingdom, 2003, s. 800–809. Dostupný z WWW http://www.scribd.com/doc/8254550/Learner–Corpora

VALIŠOVÁ, P. *Korpus jako zdroj systémového popisu české konjugace v učebnicích češtiny jako cizího jazyka.* DP, FF MU, 2009. Dostupný z WWW: https://is.muni.cz/auth/th/75420/ff_m_bl/?fakulta=1421;obdobi=4703;studium=499045

VAN ELS, T.; BONGAERTS, T.; EXTRA, G.; VAN OS, C.; JANSSEN–VAN DIETEN, A. *Applied linguistics and the learning and teaching of foreign languages.* London : Edward Arnold, 1984.

VAN PATTEN, B.; WILLIAMS, J. (eds.) *Theories in Second Language Acquisition: An Introduction.* Lawrence Erlbaum Associates, 2007.

VAN ROOY, B.; SCHÄFER, L. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In ARCHER, D.; RAYSON, R.; WILSON, A.; MCENERY, T. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), 28–31 March 2003.* Lancaster : UCREL, Lancaster University. 2003, s. 835–844. Dostupný z WWW: www.corpus4u.org/upload/forum/2005092023174960.pdf

VERSPOOR, M.H.; LOWIE, W.M.; DE BOT, C.L.J., Input and Second Language Development from a Dynamic Perspective" In PISKE, T., YOUNG–SCHOLTEN (eds.) *Input Matters*. 2007. Dostupný z WWW: http://www.rug.nl/staff/c.l.j.de.bot/VerspoorLowiedeBot2007–inputmatters.pdf

WALLACE ROBINETT, B.; SCHACHTER, J. (eds.) *Second Language Learning: Contrastive Analysis, Error Analysis, and Related Aspects.* Ann Arbor, MI : University of Michigan Press, 1983.

WARDHAUGH, R. The Contrastive Analysis Hypothesis. *TESOL Quarterly.* 1970, r. 4, no. 2, s. 123–130.

WEBER, R. P. Measurement models for content analysis. *Quality and Quantity*. 1983, vol. 17, no. 2, s. 127–149.

WEINBERGER, U. *Error Analysis with Computer Learner Corpora : A corpus–based study of errors in the written German of British University Students.* 2002. Diplomová práce. Lancaster University.

WHITE, L. On the Nature of Interlanguage Representation: Universal Grammar in the Second Language. In DOUGHTY, C. J.; LONG, M. (eds.) *The Handbook of Second Language Acquisition.* Oxford : Blackwell, 2004, s. 19–42.

WHITMAN, R.; JACKSON, K. The Unpredictability of Contrastive Analysis. *Language Learning.* 1972, vol. 22, s. 29–41.

XIAO, R. Well–known and influential corpora. In LÜDELING, A.; KYTÖ, M. (eds.) *Corpus Linguistics. An International Handbook.* Eds. HSK 29. 1. VOL. 1. Berlin/New York : Mouton De Gruyter. 2008, s. 383–457.

YAER, J. E. Instances of the Comparative Fallacy. *Columbia University Working Papers in TESOL & Applied Linguistics.* 2004, vol. 4, no. 1. Dostupný z WWW: http://www.tc.columbia.edu/academic/tesol/WJFiles/pdf/JungEun2004.pdf

YANG, W. A Tentative Analysis of Errors in Language Learning and Use. *Journal of Language Teaching and Research.* 2010, vol. 1, no. 3, s. 266–268.

ZELDES, A.; HIRSCHMANN, H.; LÜDELING, A. Multilevel Learner Corpora. *Workshop on Automatic Analysis of Learner Language, 10–14 March 2009 (AALL'09), CALICO '09*. Arizona State University, 2009.

ZHANG, Z.; CHAPMAN, S.; CIRAVEGNA, F. A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality. In CIMIANO, P.; PINTO, H. *Knowledge Engineering and Management by the Masses.* Springer Berlin : Heidelberg, 2010, s. 301–315. Dostupný z WWW: http://www.dcs.shef.ac.uk/~fabio/paperi/EKAW_2010_ver7.pdf

ZINSMEISTER, H.; BRECKLE, M. ALeSKo – an annotated learner corpus. Poster presented at the poster session of the 32. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS), Humboldt–Universität zu Berlin, 2010. Dostupný z WWW: http://www.linguistik.huberlin.de/institut/professuren/korpuslinguistik/mitarbeiterinnen/amir/Posters/Zinsmeister_Breckle.pdf