# Víceúrovňová anotace českého žákovského korpusu

Svatava Škodová (TU v Liberci), Barbora Štindlová (TU v Liberci), Jirka Hana (MFF UK v Praze), Alexandr Rosen (FF UK v Praze)

Abstract:

The paper describes a learner corpus of Czech, compiled from short essays written by students of Czech as a second or foreign language. We discuss the project's background assumptions, the process of text acquisition, transcription and mark-up, and finally focus on the annotation scheme, consisting of multiple interlinked levels to cope with a wide range of error types present in the input. Manual annotation is complemented by automatic error identification wherever possible and morphosyntactic tags for all word forms both in the emended and the original text. The annotation schema is tested on a doubly-annotated sample of approx. 10,000 words with fair inter-annotator agreement results.

## 1. Úvod

Jedním z aktuálních témat současné korpusové lingvistiky je vytváření a anotování korpusů sestávajících z textů napsaných nerodilými mluvčími. Ve srovnání s národními korpusy, které většinou disponují morfosyntaktickým značkováním a lemmatizací, může anotace v těchto korpusech zachycovat nestandardní používání jazyka. V průběhu anotačního procesu jsou nestandardní formy jazyka postupně identifikovány, emendovány a opatřeny značkou specifikující typ dané chyby. Specifika češtiny coby jazyka s bohatou flexí a volným slovosledem kladou zvláštní nároky na vytvoření anotačního schématu, který by vyhovoval všem požadavkům na zachycení a popis jazykových chyb nerodilých mluvčích.

## 2. Žákovský korpus

Žákovské korpusy jsou zásadním inovačním prvkem v oboru vyučování druhého, resp. cizího jazyka,[1] a to jak ve výzkumu, tak v samotné výuce. Jejich badatelský význam spočívá v tom, že poskytují širokému okruhu výzkumníků relativně velké soubory jazykových dat pro zkoumání tzv. žákovského mezijazyka,[2] jeho vývoje a faktorů, které tento vývoj potenciálně ovlivňují. Dovolují identifikovat pravidelnosti v mezijazyce a jejich vztah k různým potenciálním činitelům, které mezijazyk a jeho vývoj ovlivňují, jako je věk, první jazyk, relevantní okolnosti osvojování druhého jazyka, délka a povaha formální jazykové výuky apod. Žákovské korpusy umožnily

---

[1] 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. Někdy se chápe termín *druhý jazyk* jako nadřazený a označuje se jím jakýkoli nemateřský jazyk, který se jedinec učí poté, co si osvojil jazyk mateřský. Pro potřeby tohoto textu mezi oběma termíny nerozlišujeme.

[2] *Mezijazyk, tj. interlanguage* je termín používaný pro jazyk nerodilých mluvčích, který má výrazně individuální a dynamickou povahu. Někdy je zvažován 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.

nově definovat a rekonstruovat oba základní typy analýz, které se při studiu osvojování druhého/cizího jazyka tradičně uplatňovaly: kontrastivní analýzu a analýzu chybovou.

Kontrastivní analýza opřená o žákovský korpus se od analýzy tradiční odlišuje tím, že je zaměřena nikoli na studium výchozího a cílového jazyka, ale výše zmíněného mezijazyka, a sice na základě srovnání mezijazyka skupiny žáků s definovaným standardem jazyka cílového nebo na základě srovnání aktuálních stavů mezijazyků dvou různých skupin žáků. Zjišťují se přitom nejen odchylky ve smyslu nekorektního užití, ale i nadužívání nebo nedostatečného užívání („podužívání“) jednotlivých jazykových prostředků a konstrukcí, ať už chybných či korektních.

Počítačem podporovaná chybová analýza je často založena na specifickém typu chybové anotace textů. To s sebou nese systematičnost a explicitnost v míře tradičními metodami obtížně dosažitelné. Velkou výhodou je i to, že při využití korpusu analyzujeme chybná užití na pozadí užití korektních, můžeme si systematicky všímat funkčního využití nekorektních forem, sledovat prvky cizorodosti způsobující tzv. aspekt cizosti (*foreign-soundingness*) apod.

Výsledky kontrastivních i chybových analýz opřených o korpus se už poměrně dlouho a hojně využívají při tvorbě jazykových slovníků a učebních materiálů, především v angličtině. Vedle toho jsou už také k dispozici tematicky různorodé korpusové studie o žákovském jazyce, pokrývající problematiku od jednotlivých lexikálních kategorií (modálních sloves, spojek, frázových sloves) přes lexikální chyby, jevy kolokační a morfologické až po lingvistickou pragmatiku.[3]

| **Korpus** | **Rozsah (v mil. slov)** | **První jazyk** | **Cílový jazyk** | **Úroveň znalosti** | **Médium** | **Chybová anotace** |
|---|---|---|---|---|---|---|
| ICLE – International Corpus of Learner English | 3,00 | 26 | angličtina | pokročilí | psaný | ano (1/4) |
| CLC – Cambridge Learner Corpus | 35,00 | 130 | angličtina | všechny úrovně | psaný | ano (částečně) |
| LINDSEI – Louvain International Database of Spoken English | 0,80 | 11 | angličtina | pokročilí | mluvený | ano (částečně) |
| PELCRA – Polish Learner English Corpus | 0,50 | polština | angličtina | všechny úrovně | psaný | ano (částečně) |
| USE – Uppsala Student English Corpus | 1,20 | švédština | angličtina | pokročilí | psaný | ne |
| HKUST – Hong Kong University of Science and | 25,00 | čínština | angličtina | pokročilí | psaný | ano (200 tis. slov) |

[3] Srov. např. (Belz et al. 2005, Granger 1999, Oksefjell 1999, Leńko-Szymańska 2004, Rogatcheva 2009, Waibel 2008).

| | | | | | | |
|---|---|---|---|---|---|---|
| Technology Corpus of Learner English | | | | | | |
| CHUNGDAHM – Chungdahm English Learner Corpus | 131,00 | korejština | angličtina | všechny úrovně | psaný | ano (6,6 mil. slov) |
| JEFLL – Japanese EFL Learner Corpus | 0,70 | japonština | angličtina | začáteč-níci | psaný | ano (částečně) |
| MELD – Montclair Electronic Language Learners' Database | 1,00 | 16 | angličtina | pokročilí | psaný | ne |
| MICASE – Michigan Corpus of Academic Spoken English | 1,80 | různé | angličtina | pokročilí | mluvený | ne |
| NICT JLE – NICT Japanese Learner English | 2,00 | japonština | angličtina | všechny úrovně | mluvený | ano (částečně) |
| FALKO – Fehlerannotiertes Lernerkorpus | 0,30 | 5 | němčina | pokročilí | psaný | ano |
| FRIDA – French Interlanguage Database | 0,20 | různé | francouz-ština | středně pokročilí | mluvený | ano (2/3) |
| FLLOC – French Learner Language Oral Corpora | 2,00 | angličtina | francouz-ština | všechny úrovně | mluvený | ne |
| PiKUST – Poskusni korpus usvajanja slovenščine kot tujega jezika | 0,04 | 18 | slovinština | pokročilí | psaný | ano |
| ASU – ASU Corpus | 0,50 | různé | norština | pokročilí | psaný | ne |
| TUFS – TUFS Learners' Corpus: Japanese | 0,60 znaků | různé | japonština | všechny úrovně | psaný | ne (v plánu) |

Tabulka 1: Některé dostupné žákovské korpusy (podle Štindlová 2011, 63n.)

## 3. CzeSL – žákovský korpus češtiny

Žákovský korpus češtiny nerodilých mluvčích[4] (viz též Hana et al. 2010, Štindlová 2011, Štindlová et al. 2011) je budován jako součást většího projektu, který zahrnuje tzv. akviziční korpusy češtiny. Projekt pod jménem AKCES vznikl v roce 2005 na FF UK (Šebesta 2010, Šebesta 2011). CzeSL je plánován v rozsahu cca 2 miliony slov, a bude tak patřit k největším neanglickým žákovským korpusům. Důležitým

[4] Za mnoho podnětů a cenných připomínek děkujeme dalším členům řešitelského týmu, zvláště Vladimíru Petkevičovi, Haně Skoumalové, Tomáši Jelínkovi a Mileně Hnátkové. Karlu Šebestovi pak kromě toho všeho i za iniciování a vedení projektu.

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.

kompozičním principem korpusu je sběr dat od čtyř skupin mluvčích s ohledem na jejich první jazyky:

- Mluvčí slovanských jazyků. Převažují data od mluvčích disponujících jako prvním jazykem ruštinou nebo jiným východoslovanským jazykem; rozsáhlejší zastoupení budou mít data od polsky mluvících; další slovanské jazyky jsou zastoupeny jen okrajově.
- Mluvčí neslovanských indoevropských jazyků. V této skupině není dominance jednoho jazyka tak výrazná, mírnou převahu mají texty od mluvčích s prvním jazykem němčinou.
- Mluvčí neindoevropských jazyků. Předpokládáme větší zastoupení Vietnamců a Egypťanů, jinak je složení poměrně velmi různorodé.
- Romští žáci. Tato skupina má odlišnou povahu, u mluvčích nelze vždy jednoznačně rozhodnout, zda je čeština jejich jazykem prvním, nebo druhým. Sociokulturní odlišnosti mezi českou neromskou komunitou a některými komunitami romskými jsou však takového druhu, že lze u jazykového vývoje romských dětí očekávat některé rysy připomínající osvojování češtiny jako druhého jazyka. Romský subkorpus je budován v některých bodech odlišně a jsou u něj zaznamenávány i zčásti odlišné parametry.

V dalších parametrech relevantních pro využití žákovských korpusů usiluje CzeSL o maximálně možnou úplnost:

- Je založen na sběru psaných i mluvených projevů žáků, i když data písemného charakteru výrazně převažují. Psané texty se sbírají převážně v rukopisné podobě a přepisují se podle podrobně stanovených pravidel (Štindlová 2011, 106n.), která zajišťují, aby bylo z původního textu zachováno maximum informací (včetně např. rektifikačních zásahů studenta apod.) Výjimku představují kvalifikační práce, které se sbírají v podobě elektronické.
- Pokrývá všechny úrovně znalosti jazyka podle SERR.[5] V tom se odlišuje od většiny světových žákovských korpusů, které obvykle zachycují pouze jazyk žáků jedné či dvou úrovní znalosti, zpravidla pokročilých a středně pokročilých. V tomto parametru CzeSL neusiluje o vyváženost. Podmínkami sběru je dána převaha dat pocházejících od studentů úrovně B.[6] Úrovně nižší jsou zastoupeny méně.
- Žánrově a tematicky shromažďuje CzeSL texty různorodé. Ve světových korpusech dochází podle způsobu sběru dat k omezení textů např. na argumentativní a úvahové eseje. Hlavní součástí korpusu CzeSL jsou eseje psané jako součást zkoušky (bez specifického omezení), podobně jako je tomu u většiny světových korpusů. Ale navíc obsahuje i kvalifikační práce, zvláště bakalářské, magisterské a doktorské. Protože jde o práce kvalitativně jiné než ručně psané eseje a také podmínky sběru jsou u nich poněkud odlišné, budou

---

5 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://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages.

6 Úroveň B podle SERR odpovídá úrovni samostatného uživatele jazyka (tj. středně pokročilému mluvčímu). Tato úroveň znalosti cizího jazyka je v kontextu SERR rozdělena na nižší a vyšší: B1 (*intermediate*), B2 (*upper intermediate*).

tvořit samostatný subkorpus a bude potřeba na jejich odlišnost pamatovat při analýze.

- Všechny texty jsou vybaveny podrobnými metadaty ohledně mluvčích a textů, zvláště o podmínkách jejich vzniku a sběru. Ve srovnání s běžnými korpusy synchronními je u žákovských korpusů externí anotace velmi detailní, což usnadní jejich následné využití k relevantním jazykovým analýzám. U autorů se zaznamenává:

    - věk
    - první jazyk
    - znalost dalších jazyků
    - délka pobytu v České republice
    - úroveň znalosti češtiny
    - doba a způsob osvojování češtiny (jak intenzivně, s využitím jakých učebnic)
    - další kontakty s češtinou (např. bilingvní rodinné prostředí)

    U textů se eviduje:

    - téma, žánr a rozsah
    - podmínky jejich vzniku, tj. míra řízenosti jejich tvorby učitelem (téma ne/zadáno, žánr ne/zadán, velikost ne/zadána, čas ne/zadán)
    - velikost a povaha opory (ne/možnost využít slovníku, event. přípravné aktivity při zadání apod.)
    - okolnosti sběru (psáno pro korpus, jako součást zkoušky apod.)

    S uvedenými parametry lze pracovat při vyhledávání v korpusu. Hodnota žákovského korpusu a akvizičního korpusu obecně s počtem zaznamenaných metadat roste.

## 4. Anotace

V celkovém kontextu osvojování druhého/cizího jazyka se žákovský jazyk považuje jako samostatný systém a měl by se analyzovat jako celek, tj. včetně chyb, které jsou akceptovány jako důležitá součást žákovského jazyka.

Data nerodilých mluvčích se v žákovských korpusech mohou anotovat dvěma na sobě nezávislými způsoby:

1. Lingvistické značkování (tj. slovnědruhová, morfologická, příp. syntaktická anotace, lemmatizace ap.). Ve světových žákovských korpusech se nejčastěji uplatňuje značkování slovních druhů, obvykle je však aplikováno jen na menší části korpusů. Pro tento typ anotace se využívají softwarové nástroje původně vyvinuté pro potřeby analýzy národního jazyka, srov. např. van Rooy a Schäfer (2003).
2. Chybová anotace, viz např. (Díaz-Negrillo a Fernández-Domínguez 2006). Navzdory skutečnosti, že chybovou anotaci je třeba z velké části provádět manuálně, a je tudíž značně časově náročná, počet žákovských korpusů vybavených touto anotací v současné době neustále roste. Úroveň, rozsah a

koncept chybové anotace se však ve značkovaných žákovských korpusech značně odlišují.

Jako chybově anotované se vymezuje přibližně 45 % světových žákovských korpusů, ovšem jen 7 % se pokouší o komplexně pojatou chybovou anotaci se systémovou taxonomií chyb. Zbývajících 38 % žákovských korpusů uplatňuje chybové značkování vázané na explicitně vymezenou výzkumnou hypotézu, např. žákovský korpus ISLE značkuje pouze nedostatky výslovnostní, žákovský korpus CEDEL2 se zaměřuje na zachycení problémů syntaktických ap., srov. (Štindlová 2011, 74).
Chyby v žákovských projevech je možné zachycovat dvěma základními způsoby: implicitním a explicitním. Viz níže a (Štindlová 2011, 79n.).

### 4.1. Implicitní zachycení chyb – rekonstrukce

V rámci rekonstrukčního přístupu je v průběhu emendace[7] chyba v textu detekována a nahrazena korektní formou. Tento typ korekce je pojímán jako implicitní specifikace chyby. Výhodou rekonstrukčního přístupu je primárně absence klasifikačního schématu (Fitzpatrick a Seegmiller 2004): anotátor se jej nemusí učit, tj. tento typ anotování je rychlejší, nedochází k chybnému zařazení chyby. Vlastní rekonstrukce textu bez kategorizace chyb může být následně pro uživatele neprůhledná, protože nepopisuje chybu a neobjasňuje důvody pro volbu použité opravy. Zároveň také v případě, že rekonstrukční korpus není morfologicky značkován, neumožňuje přístup bez chybové typologie snadnou aplikaci kvantifikačních a statistických metod.

### 4.2. Explicitní zachycení chyb – chybová klasifikace

Pro tento typ klasifikace je již před samotnou emendací explicitně vymezen výčet možných chyb; v průběhu emendace jsou nalezené žákovské chyby identifikovány a následně kategorizovány podle předem vymezené chybové typologie.

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 je pak nižší využitelnost pro analýzy s odlišnými badatelskými cíli. I přesto tento koncept při značkování žákovských korpusů přináší cenné informace a nabízí široké možnosti statistických analýz.

Chybově značkované korpusy používají následující taxonomie:

i. Lingvisticky zaměřené taxonomie, které se liší podrobností klasifikace, tj. od označení kategorií velmi široce pojatých (morfologie, lexikum, syntax) ke kategoriím pojatým specifickým způsobem (pomocná slovesa, pasivum, apod.).
ii. Variantou, resp. rozšiřující možností taxonomie i. jsou taxonomie hierarchické založené na kombinaci různých aspektů v náhledu na chybu. Mohou označovat tzv. chybovou doménu (např. gramatickou, lexikální, stylovou), chybovou kategorii (např. aglutinace, diakritika, derivační flexe, rod, modus, atp.), slovní druh (POS).

---

[7] Termín *emendace* používáme pro přímou opravu daného chybného výrazu.

iii. Taxonomie založené na formálních typech alternace zdrojového textu; tyto taxonomie zachycují: chybějící element, přebývající element, chybně utvořený element, chybné uspořádání. V anotovaných korpusech jazyka nerodilých mluvčích je tento typ klasifikace chyb často užíván jako komplementární k lingvisticky orientované kategorizaci.

## 5. Anotační schéma

### 5.1. Anotační schéma jako kompromis

Chybová anotace žákovského korpusu CzeSL by měla umožnit podrobné statistické zpracování jazykových dat. Vytvoření anotačního schématu a efektivní chybové taxonomie je však z důvodu flektivní povahy češtiny a jejího tzv. volného slovosledu náročným úkolem. Anotační schéma navíc musí respektovat následující požadavky:

- schéma musí být zvladatelné pro anotátory;
- taxonomie nemůže být příliš rozsáhlá, ale zároveň musí být dostatečně informativní, tj. musí umožňovat dostatečně podrobné zachycení chyb;
- taxonomie by měla umožňovat budoucí rozšiřování.

Dále jsme se při tvorbě anotačního schématu museli vyrovnat s následujícími problémy specifickými pro zachycení žákovského jazyka:

#### 5.1.1. Interference

Protože anotátoři nejsou experty v oblasti osvojování a učení L2, je třeba počítat s tím, že nemohou rozpoznat interference mezi jazyky, kterými disponují žáci, jejichž texty anotují. Z toho důvodu není možné od anotátorů požadovat, aby zachycovali interferenční chyby. Např. věta *Tokio je pěkný hrad* je gramaticky správná, ale její autor, rodilý mluvčí ruštiny, zde chybně užil slovo *hrad*, které v porovnání ruština – čeština patří mezi tzv. *false friends,* jako ekvivalent k ruskému *gorod* (tj. město).

#### 5.1.2. Interpretace

Pro některé typy chyb je obtížné stanovit meze interpretace. Věta *kdyby citila na tebe zlobna* je gramaticky chybná, avšak je alespoň zhruba srozumitelná ve smyslu *kdyby se na tebe zlobila*. V takových případech je úkolem anotátora spíše interpretace textu nežli jeho oprava.

Daná věta může být nahrazena interpretací *kdyby se na tebe cítila rozzlobená* nebo *kdyby se na tebe zlobila,* přičemž první věta není zcela přirozená, avšak více se blíží originálu. V takových případech je nesnadné poskytnout anotátorům jednoznačné pokyny, jak postupovat.

#### 5.1.3. Slovosled

Jiným typem chyb specifickým pro češtinu jsou nedostatky slovosledné. Např. ve větě *Rádio je taky na skříni* slovosled implikuje informaci, že v místnosti jsou alespoň dvě rádia, z nichž jedno je umístěno na skříni. Pravděpodobnější interpretace však je, že rádio je jednou z několika věcí umístěných na skříni. Tato druhá interpretace by pak vyžadovala slovoslednou úpravu: *Na skříni je taky rádio.*

### 5.1.4. Styl

Dichotomie spisovné a obecné češtiny představuje pro anotátory další problematickou oblast, především v případě obecněčeských morfologických zakončení. Žáci, tj. autoři textů, si nemusí být vědomi statutu těchto forem a adekvátního komunikačního kontextu, ve kterém by mohly být užity. Přesto jsou v navrženém anotačním schématu tyto tvary vždy značkovány jako stylově příznakové.

Výsledná chybová typologie je kompromisem mezi limity kladenými na anotační proces a badatelskými požadavky vztahujícími se na žákovský korpus.

Korpus může být využíván k porovnávání variet žákovské češtiny, resp. verzí mezijazyka různých nerodilých mluvčích, s ohledem na vymezený standard cílového jazyka (tj. češtiny). Podobně zajímavé je i porovnávání žákovských jazyků na různých úrovních osvojení. V pedagogické oblasti vedly analýzy založené na žákovských korpusech k nové induktivní metodologii, tzv. *data-driven learning*, která je založena na využívání nástrojů a technik z korpusové lingvistiky v cizojazyčné výuce (např. využití konkordancí pro cvičení, příp. na podporu nezávislých učebních aktivit).

## 5.2. Anotace na více rovinách

O chybové anotaci nelze předem říci, jaká by měla být její ideální podoba. Do značné míry záleží na cílech a možnostech projektu, a také na typu jazyka. Jednoúrovňové anotační schéma by stačilo pro úzce definovaný účel, např. ke zkoumání morfologických zvláštností jazyka studentů. Mohlo by zachycovat i více aspektů, pokud by se příslušné údaje daly připojit k původním formám. Pro naše účely však s sebou jednoúrovňová anotace nese řadu problémů. Především je náš korpus z hlediska budoucího využití koncipován velmi široce, takže se nelze omezit na úzký okruh jazykových jevů nebo určitou rovinu popisu. Z toho plyne nutnost zaznamenávat postupné opravy a udržovat vazby mezi původní a opravenou formou i u změn ve slovosledu, změn v hranicích mezi slovy, případně i u vypuštěných a přidaných výrazů. Dalším důvodem je pak potřeba anotovat chyby, které se týkají více forem najednou, často v nekontaktním postavení.

V ideálním případě by anotátor měl mít k dispozici právě tolik rovin, kolik je k provedení potenciálně postupné anotace třeba. To lze zajistit buď volbou z většího počtu lingvisticky motivovaných rovin, nebo možností vytvářet roviny anotace podle aktuální potřeby oprav dané formy. Vzhledem k tomu, že anotátor by neměl být příliš zatěžován teoretickými dilematy a že výsledná anotace by měla být jednotná, zdá se velký nebo flexibilní počet rovin pro naše účely jako málo vhodný. Proto jsme přijali kompromisní řešení – anotátor má pro anotaci k dispozici dvě roviny. Rozhodnutí, na jaké rovině se daná forma opravuje, je dáno do značné míry formálními kritérii, ale rozdíly mezi oběma rovinami přitom mají lingvistické opodstatnění.

Rovina 0 obsahuje původní text, přepsaný z rukopisu se zachováním některých rukopisných charakteristik (varianty, nečitelné řetězce). Na rovině 1 se opravují izolované formy bez ohledu na kontext – typicky jde o překlepy a chyby v pravopisu a morfologii. Výsledkem je řetězec správných českých tvarů, i když věta z nich složená správně být nemusí. Všechny ostatní typy chyb (valence, shoda, slovosled, atd) se opravují na rovině 2.

## 5.3. Formalismus

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

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

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

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

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

Častým jevem jsou tzv. sekundární chyby, jako třeba v příkladu *dívá se na americkém filmu.* Adjektivum *americkém* se náležitě shoduje s řídícím substantivem, ale po opravě pádu předmětu na akuzativ je třeba změnit i pád shodného přívlastku. V takových příkladech se používá více odkazů: od předmětu ke slovesu jako zdůvodnění opravy pádu řídícího substantiva a od adjektiva k substantivu jako zdůvodnění opravy pádu shodného přívlastku. U přívlastku jde přitom o opravu, která je vynucena jinou opravou, tzv. opravu sekundární. Tento atribut je při značkování chyb zaznamenáván.
Od počátku jsme si vědomi toho, že – alespoň v netriviálních případech – lze chybu identifikovat pouze na základě stanovení hypotetické cílové podoby chybného výrazu, přičemž někdy nemusí být nasnadě podoba jediná. Práce s více cílovými hypotézami zatím existuje jako teoretická možnost a bude aktuální v dalších fázích projektu.

## 5.4. Typy chyb

Typický student češtiny jako cizího jazyka chybuje na všech lingvisticky motivovaných rovinách, od grafémiky až po pragmatiku. Navržené anotační schéma se z praktických důvodů omezuje na konzervativní emendaci, jejímž výsledkem je

souvislý a gramaticky správný text, ale bez nároků na stylistickou vytříbenost. Anotátor by také neměl text příliš volně interpretovat. Pokud text není dostatečně srozumitelný, mohou být příslušné pasáže takto označeny, ale mohou zůstat bez emendace.

Východiskem pro taxonomii chyb jsou lingvistické kategorie ve spojení s formálním popisem chyby (typem modifikace). Ne všechny typy chyb je nutné určovat manuálně. Pokud je to možné, určujeme některé chyby automaticky porovnáním původní a opravené podoby tvaru a/nebo na základě výsledků automatické lemmatizace a morfologické analýzy (viz oddíl 5.3). Emendace zatím probíhá jen ručně, i když se zkoumá možnost využití automatického korektoru.

### 5.4.1. Chyby na rovině 1

Na rovině 1, kde se opravují chyby zjistitelné bez ohledu na kontext, se kromě chyb v pravopisu a hranicích slov řeší také chyby ve flektivní a derivační morfologii i chybné slovní základy, např. nově vytvořená nebo cizí slova. Tyto nedostatky se s výjimkou chyb pravopisných určují manuálně. Výsledkem opravy je nejpodobnější správný tvar, který může být dále na rovině 2 podle kontextu opraven na jiný – důvodem je například porušení morfosyntaktické shody nebo sémantická nekompatibilita lexému. Seznam chyb anotovaných manuálně na rovině 1 s příklady uvádí tabulka 2. Poslední tři chyby (*stylColl*, *stylOther* a *problem*) se používají i na rovině 2.

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

Tabulka 2: Chyby na rovině 1

Pravidlo, že na rovině 1 musí být všechny tvary správné, neplatí bez výjimky – chybu nelze opravit třeba proto, že anotátor nedokáže rozpoznat intenci autora. Na druhé straně se správný tvar nahrazuje jiným správným tvarem v případech, kdy jde evidentně o pravopisnou nebo hláskovou chybu, jejímž výsledkem bylo náhodné homonymum s existujícím tvarem.

### 5.4.2. Chyby na rovině 2

Opravy na rovině 2 se týkají chyb ve shodě, valenci, analytických tvarech, zájmenném odkazování, záporové shodě, v užití vidu, času, stupně, lexému a idiomu, a také ve slovosledu. U chyb ve shodě, valenci, analytických tvarech, zájmenném

odkazování a záporové shodě lze obvykle při opravě chybného výrazu odkázat na jiný správně utvořený nebo již opravený výraz, který určuje morfologické kategorie nebo jiné vlastnosti výrazu opravovaného. Typy manuálně určovaných chyb na rovině 2 uvádí tabulka 3. (Mezi automaticky identifikované chyby patří např. chyby slovosledu nebo podrobnější členění chyby typu *vbx*.)

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

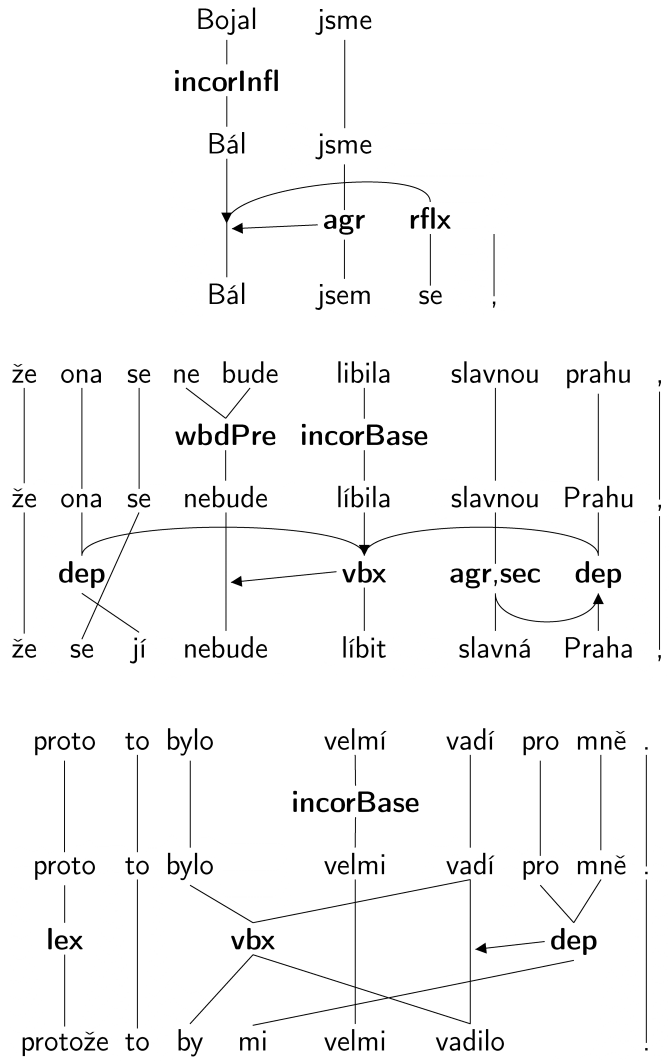
Tabulka 3: Chyby na rovině 2

### 5.4.3. Příklad

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

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

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




Obr. 1: Příklad anotace jedné věty

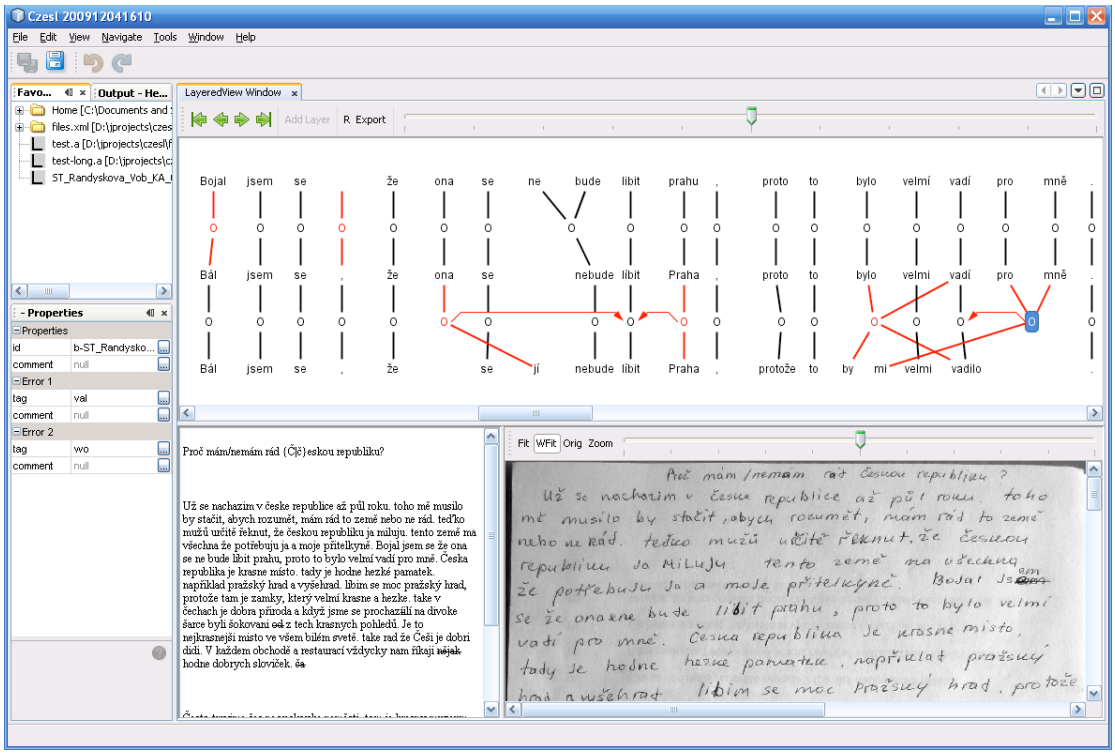
Poslední úsek věty vyžadoval na R1 jen jednu opravu (opět délka ve slovním základu). Zato bylo na R2 nutné kromě spojky (lexikální oprava) změnit celý analytický slovesný tvar, což je příklad opravy typu 2:2. S odkazem na řídící sloveso pak i předložkový pád zájmena na pád prostý (*mi*) a výsledek nakonec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

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

## 6. Postup anotace

Celá anotace probíhá v těchto krocích:

1. Rukou psaný text se pomocí běžného textového editoru přepíše do elektronické podoby ve formátu HTML rozšířeného o kódy zachycující studentovy opravy, předtištěný text, text v jiných abecedách atd.
2. Přepsaný text v elektronické podobě se zkonvertuje do formátu pro anotaci, v němž je automaticky určena rovina 0 a výchozí podoba roviny 1. Obě jsou zakódované ve formátu PML (Pajas a Štěpánek 2006; konkretizace XML pro účely strukturní lingvistické anotace).
3. Anotátor opraví chyby v dokumentu a určí jejich typ pomocí anotačního editoru *feat*.
4. Klasifikace chyb, které lze z ruční anotace odvodit automaticky, se přidá v dalším kroku.




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

## 6.1. Přepis původního textu

Vzhledem k tomu, že původní texty většinou píší studenti a žáci ve třídě při jazykových kursech nebo při zkouškách, je nutné pracovat s rukopisy.[8] Dalším důvodem je obava, že elektronické texty lze snadno korigovat nebo i vytvářet automatickými nástroji, což by podobu autentického mezijazyka výrazně zkreslilo.

I když se snažíme o maximální věrnost, někdy se při přepisu rukopisných textů neobejdeme bez jisté míry interpretace. Přepisovači si musí uvědomovat specifika rukopisu dané skupiny studentů a někdy i jednotlivců (například stejný glyf je možné interpretovat v písmu různých studentů jako písmeno *l*, *e*, nebo *a)*. Pokud je

8 Přepisy mluvených textů budou do korpusu zařazeny v dalších fázích projektu.

možné znak nebo i celý úsek textu interpretovat různě, přepisovač může uvést více variant. Tak například velikost počátečních písmen nebo hranice slov jsou často nejasné. Zvlášť se označují zcela nečitelné úseky i opravy, které provádějí sami studenti (vsuvky, škrty) a které mohou být pro výzkum akvizice jazyka také užitečné. Podrobné pokyny jsou uvedeny v přepisovacím manuálu.

### 6.2. Anotace

Ruční část anotace probíhá v prostředí anotačního editoru *feat* (http://ufal.mff.cuni.cz/~hana/feat.html), který byl vyvinut v rámci projektu. Anotátor opraví text na příslušných rovinách, upraví vztahy mezi výrazy, které si na jednotlivých rovinách vzájemně odpovídají (implicitně jsou všechny vztahy typu 1:1) a u chyb určitého typu přidá příslušnou chybovou značku. Anotovaný text je možné zobrazit v přepsané podobě i jako snímek originálu. Anotační editor je vytvořen v jazyce Java s využitím platformy Netbeans (http://platform.netbeans.org/). Na obr. 2 je ukázka anotace věty z výše uvedeného příkladu v prostředí anotačního editoru.

### 6.3. Evaluace

Použitelnost anotačního schématu a taxonomie chyb byla ověřena pomocí míry shody mezi anotátory na vzorku 67 textů v průměru po 150 slovech, celkem 9373 slov (7995 slov bez interpunkce). Autory textů byli rodilí mluvčí různých jazyků. Každý text anotovali dva anotátoři, celkem bylo anotátorů 14. Jako míra shody mezi anotátory byl použit koeficient kappa (Carletta 1996), který kromě shody nebo neshody mezi dvěma anotátory při volbě dané značky bere v úvahu i pravděpodobnost náhodné shody. Blíže o evaluaci viz (Štindlová 2011, 121n., Štindlová et al. 2011).

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

### 6.4. Následné zpracování

Po manuální anotaci následuje anotace automatická. Při ní se k textům přidávají údaje, které lze algoritmicky odvodit z originálu, provedené emendace a manuální chybové anotace. Jde o tyto údaje:

1. Rovina 1: lemma, slovní druh a morfologické kategorie pro jednotlivé tvary (tyto údaje mohou být víceznačné)
2. Rovina 2: lemma, slovní druh a morfologické kategorie pro jednotlivé tvary (jednoznačně určené)
3. Rovina 1: typ chyby (porovnáním původních a opravených řetězců) kromě lexikálních chyb, při jejichž opravě je nutné měnit lemma (např. *kadeřnička*)
4. Rovina 2: morfosyntaktické chyby způsobené narušenou shodou nebo rekcí (porovnáním morfosyntaktických značek na rovině 1 a 2)

5. Formální popis chyby na obou rovinách: typ pravopisné nebo hláskové změny, přidání/vypuštění výrazu, slovosledná chyba

V budoucnu chceme automaticky označovat chyby ve slovesných předponách, flektivních koncovkách, pravopisu, palatalizaci a chyby metateze.

## 7. Závěr

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

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

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

## Literatura

Belz J., N. Vyatkina, 2005, Learner Corpus Analysis and the Development of L2 Pragmatic Competence in Networked Intercultural Language Study: The Case of German Modal Particles. *Canadian Modern Language Review,* 62, č. 1, 17–48.

Carletta J. C., 1996, Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22, č. 2, 249–254.

Corder, S. P., 1981, *Error Analysis and Interlanguage.* Oxford University Press, Oxford.

Díaz-Negrillo A., J. Fernández-Domínguez, 2006, Error Tagging Systems for Learner Corpora. *Resla*, č. 19, 83–102.

Díaz-Negrillo A., D. Meurers, S. Valera, H. Wunsch, 2010, Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36, č. 1–2, 139–154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

Fitzpatrick E., M. S. Seegmiller, 2004, The Montclair electronic language database project. In *Applied Corpus Linguistics: A Multidimensional Perspective*, eds U. Connor, T. A. Upton. Rodopi, 223–238.

Hana J., A. Rosen, S. Škodová, B. Štindlová, 2010, Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala. Association for Computational Linguistics.

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

Lüdeling A., M. Walter, E. Kroymann, P. Adolphs, 2005, Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham.

Pajas P., J. Štěpánek, 2006, XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In *Proceedings of LREC 2006 Workshop on Merging and Layering Linguistic Information*. ELRA, Genoa, Italy.

Rogatcheva S., 2009, „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. Peter Lang, Frankfurt, 255–278.

Selinker L., 1972, Interlanguage. *IRAL*, 10, č. 3, 209–231.

Schmidt T., 2009, Creating and working with spoken language corpora in EXMARaLDA. In *LULCL II: Lesser Used Languages & Computer Linguistics II*, 151–164.

Šebesta K., 2010, Korpusy češtiny a osvojování jazyka. *Studie z aplikované lingvistiky/Studies in Applied Linguistics*. Sv. 1, č. 2, 11–34.

Šebesta K., 2011, Akviziční korpusy. In *Minulost, přítomnost a budoucnost v jazyce a v literatuře. Ústí nad Labem 1.–3. 9. 2010.* PF UJEP, Ústí nad Labem.

Štindlová B., 2011, *Evaluace chybové anotace v žákovském korpusu češtiny.* Disertační práce, Filosofická fakulta University Karlovy v Praze.

Štindlová B., S. Škodová, J. Hana, A. Rosen, 2011, CzeSL – an error tagged corpus of Czech as a second language. PALC 2011 – Practical Applications in Language and Computers, Lodž 13.–15. dubna 2011. Výběr z příspěvků vyjde v nakladatelství Peter Lang v edici Łódź Studies in Language.

Van Rooy B., L. Schäfer, 2003, An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In *Proceedings of the Corpus Linguistics 2003 Conference Lancaster University (UK), 28–31 March 2003*, eds D. Archer, R. Rayson, A. Wilson, T. McEnery, UCREL, Lancaster University, Lancaster, 835–844.

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