Oponentský posudek doktorské disertační práce

Mgr. Barbory Štindlové

Evaluace chybové anotace v žákovském korpusu češtiny

Autorka se ve své disertační práci věnuje problematice češtiny jako cizího jazyka a v této oblasti pak zvláště tzv. *žákovským korpusům*. V obsáhlé dvousetstránkové práci (147 stránek textu + přílohy) studuje jazyk nerodilých mluvčí(ch) z hlediska teorie chyby a poté se z mnoha hledisek zabývá korpusy nerodilých mluvčích, tj. žákovskými korpusy. V první kapitole nazvané *Teorie chyby v jazyce nerodilých mluvčích* komplexně zkoumá pojem jazykové chyby. Zabývá se vůbec teorií nabývání (akvizice) druhého jazyka, kde konstatuje existenci mnoha rozmanitých přístupů. Z historického hlediska zmiňuje významná paradigmata: behaviorální, kognitivně-komputační a dialogické. Všímá si celé řady různých teoretických přístupů a jejich vývoje a projevuje v této oblasti velkou obeznámenost s relevantní literaturou (pokud mohu soudit); nastudovala toho ke své práci opravdu hodně. Poukazuje na nesnadnost vymezení chyby a rozlišuje chyby v produkci a recepci jazyka, přičemž chyby v produkci lze – jak konstatuje – zjišťovat mnohem snáze než v recepci. Dále se zaměřuje na vývoj samého pojmu chyby, zdůrazňuje interferenci prvního a druhého jazyka a především postupnou změnu v pojetí chyby: jazyk nerodilých mluvčích se totiž začíná chápat jako svébytný jazykový systém, který se postupně přibližuje cílovému jazyku. V této souvislosti se objevuje pojem *interlanguage* (= žákovský jazyk), svébytný přechodný systém mezi prvním a cílovým jazykem. Důležitá je v této souvislosti autorčina výzva, že „mezijazyk [...] by měl být analyzován jako celek, včetně nekorektních struktur".

Poté se autorka věnuje hlavním metodám analýzy žákovského jazyka, totiž kontrastivní a chybové analýze, analýze přirozené posloupnosti akvizice morfémů a analýze frekvenční. V rámci chybové analýzy si všímá mj. rozdílu mezi gramatičností a přijatelností, uvádí rozdíl mezi Corderovými *mistakes* (chyby v performanci) a *errors* (chyby v kompetenci, systémové), a upozorňuje na další anglické pojmy jemně rozlišující typy chyb (*slips*, *lapses*). Podrobně popisuje jednotlivá stadia procesu chybové analýzy (sběr dat, identifikace chyb, popis chyb a chybovou taxonomii, explanaci a evaluaci chyby). V rámci odstavce *Identifikace chyb* pěkně demonstruje klíčovou roli komunikačního kontextu při hodnocení chyby, a to na větě *Jsem student* jako odpověď na otázku: *Co děláte*?, nebo na otázku: *Co děláte dnes večer*?

Velmi důležitá je i z hlediska evaluace žákovského korpusu CZESL ve druhé polovině práce chybová taxonomie podle a) povrchové realizace a podle b) lingvistických kategorií. Zajímavá je pasáž o tzv. *interlingválních* a

počítá až po manuálním značkování a emendaci, s automatickou emendací se nepočítá.

Nejvlastnějším jádrem práce je kapitola devátá s názvem **Evaluace anotace navržené pro žákovský korpus češtiny**. Autorka se podrobně věnuje mezianotátorské shodě, kterou vyhodnocuje koeficientem shody *kappa*, a uvádí i další míry. V podkapitole 9.1.1 uvádí vzorec pro koeficient *kappa*, který ovšem patrně není zapsán správně: znak sumy by asi ve vzorci pro P(o) a P(e) měl chybět, pak teprve po mém soudu dávají smysl úvahy na s. 125 nahoře. Po zavedení míry *kappa* se pak doktorandka zabývá mezianotátorskou shodou při anotacích žákovských korpusů a uvádí různé příklady aplikace této míry. Poté se podrobně věnuje míře *kappa* použité na měření shody při anotaci korpusu CZESL a podrobně analyzuje mezianotátorskou shodu při značkování různých lingvistických jevů, a to na rovinách R1 a R2. Diskuse výsledků je logická a závěry přirozené; zajímavá je například diskuse o tagu *lex*: že při sémantické blízkosti „nastává u anotátorů vysoká neshoda v názoru na potřebu emendace a tedy i následnou anotaci". Dále autorka v souvislosti s emendací zdůrazňuje význam explicitního vyjádření cílové interpretace v anotačním schématu Lüdelingové – s tím naprosto souhlasím. V závěru pasáže o mezianotátorské shodě se zabývá příčinami mezianotátorské neshody a velmi výstižně shrnuje několik příčin neshody (na s. 142–144). V koncizním závěru autorka shrnuje obsah své práce.

Mimo kmenový text a přílohy obsahuje práce i 13 tabulek a 12 grafů přispívajících ke snazší čitelnosti a názornosti pojednávané problematiky. Autorka rovněž do práce rozumně zařadila soupis základních termínů a zkratek a jednu z příloh tvoří i cenný anglicko-český terminologický slovník

Klady práce

Na práci Mgr. Štindlové si cením zejména:

- autorčina suverénního přehledu o problematice žákovských korpusů;
- logické výstavby práce a logické návaznosti jednotlivých kapitol;
- v celé práci jasné, logické argumentace při rozboru jednotlivých témat;
- komplexnosti práce: toho, že v práci patrně neopomněla žádný podstatný zřetel v souvislosti s problematikou jazykové chyby a žákovských korpusů;
- názornosti zprostředkované tabulkami a grafy;
- evaluace žákovského korpusu češtiny, kde předvádí výbornou znalost specifik žákovského korpusu češtiny;
- vhodných přepisovacích zásad, jež se dodržují při předanotační přípravě textů českého žákovského korpusu, a příslušných pravidel pro přepis;
- pěkného jazyka, kterým je práce psána;

anglicky citovaný Sinclairův výrok v pozn. 116 na s. 72 při obhajobě objasnila a vysvětlila, jak mu rozumí; já mu totiž vůbec nerozumím. Zaznamenává však i podobný postoj Fitzpatrickové a Seemillera, že totiž chybová anotace je interpretační model ovlivňující přístup k datům. To může být pravda, v každém případě však výzkumník má možnost vždy zkoumat neanotovaná, surová data.

Autorka se dále zabývá specifickou problematikou – anotačními modely: lineárním modelem a víceúrovňovou distanční anotací (užívanou ve vlivném německém korpusu FALKO). S velkou znalostí věci tyto modely charakterizuje, přičemž si kriticky všímá jejich nedostatků. Závěr 6. kapitoly tvoří *chybová taxonomie*: autorka uvádí typologii a strukturu chybových taxonomií. Zde je velmi zajímavý sám závěr kapitoly 6: využití chybového povrchově gramatického značkování pro popis komunikační kompetence a nedostatečnost takového značkování.

V kapitole 7 autorka v přehledu analyzuje vybrané žákovské korpusy (hlavní světový korpus ICLE [Lovaň], dále NICT JLE, MELD, CLC [chybové značkování tohoto korpusu mě velmi zaujalo: obsahuje mj. sledování falešných přátel!], německý FALKO a maličký slovinský PIKUST početně maličkého slovinského národa). Obecně se ukazuje, že přístup k chybovému značkování je u různých korpusů velmi odlišný. Autorka celkově opět prokazuje výtečný přehled o této problematice.

V kapitole 8 se autorka konečně dostává k češtině jako žákovskému jazyku. (Je to společensky závažná problematika: osvojování češtiny cizinci je pro jejich integraci do většinové společnosti v České republice zcela klíčové.) Autorka nejprve podává přehled příslušné literatury české provenience týkající se problematiky žákovských korpusů obecně i korpusů češtiny jako druhého jazyka. V podkapitole 8.1 podává autorka základní informace o vznikajícím korpusu CZESL: charakterizuje nejprve metadata, jimiž se korpusové texty značkují, a pak píše podrobně o přepisu českých textů psaných nerodilými mluvčími češtiny. Je obeznámena s pravidly TEI v oblasti přepisu rukopisů, zná brněnský Korpus soukromé korespondence. Velmi podrobně se pak rozepisuje o zásadách pro přepis textů nerodilých mluvčích češtiny a odpovídajících pravidlech pro přepis: zásady i pravidla jsou samostatným dílem autorky a představují jeden z vrcholů práce. V důležité kapitole 8.4 dále Mgr. Štindlová píše o lingvistické anotaci žákovského korpusu češtiny CZESL: autorka vyzdvihuje tyto její hlavní vlastnosti:

- tři anotační roviny
- jedinou cílovou hypotézu
- dvoustupňovou (postupnou) emendaci chyb.

Popisuje i klíčovou chybovou taxonomii v korpusu CZESL, v níž se anotace chyb dělí mimo kritéria lingvistická i na chyby značkované *manuálně* a chyby přiřazované *automaticky*. S automatickou anotací chyb se v korpusu CZESL

intralingválních chybách. Zde bych se rád zeptal na pojem *exploiting redundancy* (překládaný jako *zneužití chyby*), který mi v daném kontextu není příliš jasný.

Autorka projevuje rozsáhlé znalosti vývoje myšlení o chybové analýze. Tvrdí, že bez kvalitní a podrobné typologie chyb nelze adekvátně analyzovat chyby nerodilých mluvčích a že takovou typologii nelze vytvořit bez rozsáhlého žákovského korpusu. V osmdesátých letech minulého století ovšem pro vznik korpusů ještě neexistovaly dobré podmínky (např. informatika a počítače zdaleka nebyly na dnešní úrovni).

Závěr první kapitoly tvoří stručné pojednání o analýze přirozené posloupnosti akvizice morfémů a frekvenční analýze. Celkově lze říci, že autorka prokázala vynikající přehled o problematice chyby a jejích širokých souvislostech. Tyto znalosti pak dokáže uplatnit při evaluaci koncepce žákovského korpusu nerodilých mluvčích češtiny v závěru práce.

Druhou kapitolou začíná jádro práce: komplexní pohled na žákovské korpusy. Po rozsáhlém přehledu literatury o žákovských korpusech ve světě a uvedení základní literatury české provenience autorka definuje sám pojem žákovský korpus a metody jeho zkoumání (některé z nich jsou obecně korpusové aplikované na žákovské korpusy, jiné specifické). Nejprve zdůrazňuje, k čemu takový korpus vlastně je; zabývá se kontrastivním srovnáváním jazyka rodilých a nerodilých mluvčích a srovnáváním různých mezijazyků (a v této souvislosti klíčového vlivu prvního jazyka) a také počítačem podporovanou chybovou analýzou. Poukazuje na korpusy lingvisticky *anotované* a lingvisticky *neanotované*. Dále uvádí velmi zajímavou pasáž o typologii žákovských korpusů z řady hledisek: zde mě nejvíce zaujala poslední část týkající se chybové anotace. Potěšil mě výrok zesnulého Johna Sinclaira v pozn. 92: je to jeden z mála jeho výroků, s nimiž se mohu ztotožnit (buď Sinclair přichází s truismy, nebo jsou jeho obecné výroky o korpusech po mém soudu nepravdivé).

V kapitole 3 autorka píše o zásadách výstavby žákovského korpusu a v kratičké kapitole 4 uvádí cíle budování žákovských korpusů. V kapitole 5 podrobně mapuje existující žákovské korpusy ve světě, a to z různých hledisek včetně pro mne nejzajímavějšího zřetele: chybové anotace. Přehled je nesmírně podrobný, rešerše důkladné. V podkapitole 5.3 mě mimo samozřejmé věci, jaké představuje využití žákovských korpusů při výzkumu nabývání cizího jazyka a v cizojazyčném vyučování, zaujal postřeh, že korpus jazyka nerodilých mluvčích může přispívat i poznávání cílového jazyka. Přesně tak: podobně jako studium jazykových defektů (afázie) slouží k obohacení našich znalostí o jazyce jazykově zdravých mluvčích.

V kapitole 6 se autorka podrobně zaměřuje na velmi zajímavou problematiku chybové anotace ve světě žákovských korpusů. Zdůrazňuje význam anotace obecně, citujíc G. Leeche, a (naštěstí jen) v poznámce uvádí naopak názor Sinclairův, který je k anotaci nevstřícný. Prosím autorku, aby mi

- malého počtu formálních chyb.

Zápory práce

Zápory práce shledávám pouze v několika formálních chybách, a to v pasáži o koeficientu *kappa* (zbytečný znak pro sumu, ale možná se mýlím!) a na těchto místech, kde se nacházejí následující překlepy a nejasnosti (výčet není úplný):

s. 26: *dispreffered errors*
s. 28: pozn. 53 – *grafologickou* rovinu (?)
s. 36: 4. řádek shora: *kurikulí* (?)
s. 51: dole: začínat větu slovem *Resp.* je zvláštní!
s. 84: pozn. 144: *missuse* (?)
s. 115: věta začínající *A to jak ...* uprostřed stránky vhodně nenavazuje na větu předchozí
s. 123: uprostřed: *exitujících* → *existujících*
s. 124: uprostřed: *očekává shoda* → *očekávaná shoda.*

Celkově lze konstatovat, že práce je výtečným popisem rozvíjející se oblasti žákovských korpusů obecně a českého korpusu CZESL zvláště. Myslím, že každý, kdo se toho chce o problematice žákovských korpusů hodně dozvědět, kdo se chce poučit o probíraných problémech a třeba začít budovat svůj žákovský korpus, by neměl práci Mgr. Štindlové pominout. Práci doporučuji publikovat knižně, předtím je však třeba odstranit chyby; nebylo by špatné, kdyby autorka podrobněji rozvedla architekturu anotačního systému a chybovou taxonomii českých žákovských textů.

Na závěr oficiálně konstatuji, že předložená disertační práce má velmi vysokou úroveň. Doktorandka odvedla velký kus práce a nade vši pochybnost prokázala své schopnosti samostatně vědecky pracovat. Práci rád doporučuji k obhajobě a doporučuji též, aby doktorandce byl udělen titul Ph.D.

V Praze dne 10. 5. 2011

doc. RNDr. Vladimír Petkevič, CSc.
Ústav teoretické a komputační lingvistiky FFUK
oponent