MERLIN: Multilingvální platforma pro evropské referenční úrovně

Barbora Štindlová, Veronika Čurdová

Mgr. Barbora Štindlová, Ph.D.
barbora.stindlova@ujop.cuni.cz

Mgr. Veronika Čurdová
veru.curdova@gmail.com

Adresa:
ÚJOP UK
Vratislavova 29/10
128 00 Praha 2

ABSTRACT:

Merlin: Multilingual platform for common reference levels

The article provides an overview of the motivation, evolution and major principles of the international project Merlin. The main output of this project is a unique trilingual learner corpus consisting of German,Italian and Czech. The corpus will be available as an online platform illustrating the Common European Framework of Reference for Languages (CEFR) with authentic learner data and enabling users to explore authentic written learner productions and related metadata (e.g. age, first language of the learner etc.). Each text in the corpus is linguistically analysed during the multiphase error annotation. This process brings some problematic issues concerning the particularly specific character of Czech as a Slavic language. The article summarizes some of these problems and their possible solutions.

Klíčová slova: SERRJ, žákovský korpus, referenční úrovně, jazykové testování, čeština, italština, němčina

Key words: CEFR, learner corpus, reference levels, language testing, Czech, Italian, German

1. Projekt Merlin

Autoritativním dokumentem v oblasti výuky a testování evropských jazyků jako jazyků cizích je od roku 2001 tzv. *Společný evropský referenční rámec pro jazyky* (SERRJ, angl. *CEFR*). Tento dokument definuje 6 úrovní jazyka a vymezuje deskriptory reflektující míru osvojení cizího jazyka v souvislosti s jazykovými dovednostmi, tj. porozuměním psanému či mluvenému projevu a schopností aktivního vyjadřování. V návaznosti na úrovně definované podle SERRJ se v současnosti vytváří téměř všechny učební materiály, sylaby, kurikula i certifikované zkoušky hodnotící jazykové schopnosti studenta. Z toho důvodu je nutné co nejpřesněji jednotlivé úrovně charakterizovat, odlišit a také poskytnout zainteresovaným odborníkům, ale i učitelům a studentům konkrétní jazyková data odpovídající požadavkům, které by měl student na jednotlivých úrovních osvojení jazyka splňovat.

Nutnost dostatečně ilustrovat úrovně vymezené v SERRJ byla i podstatnou motivací vzniku mezinárodního projektu *Merlin: Multilingvální platforma pro evropské referenční úrovně: Výzkum jazyka studentů v kontextu* (2012–2014).[1] Tento projekt je specifický svým zacílením na tři různé národní jazyky (češtinu, němčinu a italštinu), jež jsou zpracovávány na základě společných, jednotných kritérií a pravidel tak, aby bylo možné vymezit takové aspekty jazyka, které budou dostatečně charakterizovat dané referenční úrovně a které nebudou ukotveny pouze teoreticky, ale budou vycházet z reálných jazykových dat. Srov. Wisniewski a kol. (2013). Konkrétním vyústěním projektu je volně dostupná online platforma, která přispívá k

[1] Projekt Merlin je financován Evropskou komisí v rámci Programu celoživotního učení (518989-LLP-1-2011-DE-KA2-KA2MP). Koordinátorem projektu je Technische Universität Dresden (Německo), partnery projektu jsou EURAC (Itálie), Univerzita Karlova (Česká republika), Eberhard-Karls-Universität Tübingen (Německo), telc GmbH (Německo), Berufsförderungsinstitut Oberösterreich (Rakousko) a European Center of Modern Languages – Council of Europe (Rakousko).

určení a ověření úrovní definovaných podle SERRJ, nabízí autorům zkoušek a učebnic, vyučujícím i studentům vhled do písemných projevů nerodilých mluvčích a umožňuje vyhledávat konkrétní jazykové rysy typické pro danou jazykovou úroveň (A1–C1).[2] Jádrem celého projektu a základem této platformy je trojjazyčný korpus němčiny, italštiny a češtiny jako cizích jazyků budovaný ve spolupráci institucí v Německu, Rakousku, Itálii a České republice, jeden z mála žákovských korpusů, který podstatným způsobem reflektuje SERRJ.

Česká část korpusu Merlin je vedle žákovského korpusu CzeSL v současné době jediným relevantním zdrojem projevů nerodilých mluvčích češtiny. Svým charakterem se zároveň řadí k souboru akvizičních korpusů AKCES.[3]

2. Sběr a zpracování dat

Zdrojem dat pro žákovský korpus Merlin jsou písemné projevy nerodilých mluvčích, které vznikly v průběhu standardizovaných testů v rámci certifikovaných zkoušek z daných jazyků jako jazyků cizích,[4] jež se přísně vztahují k úrovním popisu podle SERRJ. Počet shromážděných textů a celkový počet slov v rámci jednotlivých subkorpusů je uveden v tabulce 1. Pro češtinu jsou v současnosti k dispozici texty reprezentující referenční úrovně A2, B1 a B2.

Tabulka 1: Přehled dat

Úroveň podle SERRJ	**čeština**	**němčina**	**italština**	**Celkem**
A1	1	57	30	88
A2	49	199	294	542

[2] Platforma projektu Merlin je dostupná na www.merlin-platform.eu.
[3] K žákovskému korpusu CzeSL viz blíže Štindlová (2013). Soubor akvizičních korpusů AKCES viz http://akces.ff.cuni.cz.
[4] telc, UNIcert, CCE

A2+	112	107	94	313
B1	89	219	343	651
B1+	75	115	53	243
B2	72	219	2	293
B2+	9	73		82
C1	4	43		46
C2		4		4
Celkem (texty)	411	1,035	816	2,262
Celkem (slova)	64,488	125,927	92,359	282,774

Spolu s písemnými projevy byla shromážděna i standardní metadata, tj. informace o mateřském jazyce kandidáta, pohlaví, věku, ale také o testující instituci a typu zpracovávaného úkolu. Zároveň je každý text klasifikován vzhledem k reálné jazykové zdatnosti autora v oblasti ortografie, gramatiky, slovní zásoby, koherence textu a sociolingvistické adekvátnosti.

2.1 Přepis rukopisů

V první fázi zpracování dat byly oskenované rukopisy přepisovány do elektronické podoby v editoru XMLmind.[5] Pro zachování textu v podobě co nejbližší originálu a pro maximální zredukování možných interpretací ze strany přepisujícího byla specifikována podrobná transkripční pravidla. Pomocí vkládaných anotačních značek lze označit nečitelné pasáže textů, je možné zachytit více možných variant interpretace u sporných grafémů, příp. věrně přepsat diakritiku, kterou čeština nedisponuje. Značky lze využít také k zaznamenání oprav provedených v textu samotným autorem či k vyznačení pasáží, které zdůraznil např. podtržením. Při přepisu se zachovávají i užité emotikony, obrázky a symboly. Během transkripce

[5] http://www.xmlmind.com/xmleditor

dochází též k anonymizaci osobních údajů. Některé z transkripčních značek uvádíme v tabulce 2.[6] Podstatnou součástí zpracování dat je také kontrola tokenizace a kontrola kvality transkripce, která má za úkol rozlišit původní chyby studenta od chyb vzniklých při přepisu a také ověřit náležité užití základních značek.
Po ukončení procesu transkripce je přepsaný text konvertován do podoby umožňující jeho následnou lingvistickou (a chybovou) anotaci.[7]

Tabulka 2: Příklady značek pro přepis

Značka	Popis
<ambiguous> <alternative>	pokud si přepisující není jist variantou slova / písmena v textu
<citation>	v případě, že autor cituje vstupní text
<comment>	komentář přepisujícího
<correction> <deletion> <insertion>	oprava autora v textu (dvě varianty – škrt, vsuvka)
<emoticon>	emotikon
<greeting>	pozdrav
<image>	obrázek
<unreadable>	nečitelný text, aj.

3. Anotace

[6] Pravidla pro přepis a kompletní seznam transkripčních značek jsou k dispozici na www.merlin-platform.eu.
[7] Pomocí nástroje PAULA (viz https://www.sfb632.uni-potsdam.de/en/paula.html).

Anotační schéma odráží zásadní koncept projektu Merlin, a to chápání žákovského jazyka jako samostatného dynamického jazykového systému, tzv. *interlanguage* (Corder, 1981). Zároveň se vztahuje ke dvěma perspektivám popisu žákovského jazyka: k učení/vyučování cizímu jazyku (angl. *FLL/FLT*), jež akcentuje odlišnosti mezi žákovou produkcí a standardem rodilého mluvčího cílového jazyka (srov. např. Lüdeling, 2008); a také k výzkumům v nabývání druhého jazyka (angl. *SLA*) kladoucím důraz na lingvistické charakteristiky žákovského jazyka, jenž není chápán jako arbitrární a jako svébytný jazykový systém by mohl, resp. měl být podroben standardní lingvistické anotaci.

Základem anotace korpusu Merlin je široká škála indikátorů, které umožňují popsat komplexní charakter žákovského jazyka, jeho standardní i nestandardní aspekty. Tyto indikátory jsou relevantní pro všechny tři jazyky a berou v úvahu jak společné rysy (např. použití spojek, verbonominální shoda, kolokace apod.), tak jevy jazykově specifické (např. Č: dvojí negace, posesivní reflexíva; I: lexikalizovaná klitika, N: modální částice, I/N: členy, apod.). Viz i Abel a kol. (2013).

Množina ortograficky, gramaticky, lexikálně a sociolingvisticky orientovaných indikátorů byla vymezena na základě charakteristik úrovní uváděných v SERR i v sekundární literatuře, podle předběžné analýzy písemných projevů studentů, ale také v souvislosti s dotazníkovým šetřením mezi budoucími uživateli korpusu, tj. učiteli, hodnotiteli, studenty.[8] Tato vícezdrojová analýza se stala základem pro stanovení více než sedmdesáti indikátorů a jim odpovídajících tagů shrnutých v anotačním schématu (příklady anotačních značek jsou uvedeny v tabulce 3).

Tabulka 3: Příklady anotačních značek (názvy specifikací ponecháváme v angličtině).[9]

EA1 = první rovina anotace / EA2 = druhá rovina anotace

[8] K indikátorům pro analýzu žákovského jazyka blíže viz Štindlová a kol. (2014: 141n.).

[9] Kompletní seznam anotačních značek užívaných v korpusu Merlin viz www.merlin-platform.eu.

JAZYKOVÁ SPECIFIKACE (ÚROVEŇ 1)	JAZYKOVÁ SPECIFIKACE (ÚROVEŇ 2)	JAZYKOVÁ SPECIFIKACE (ÚROVEŇ 3)	TAG
EA1			
GRAMMAR	word order	word order in main clause	G_Wo_womaincl
GRAMMAR	negation	negation general	G_Neg_neggen_Pos G_Neg_neggen_O G_Neg_neggen_Ch G_Neg_neggen_Ad
GRAMMAR	verb valency	complement number	G_Valency_complnumb_O G_Valency_complnumb_Ad
GRAMMAR	reflexivity	reflexive pronoun	G_Refl_pronrefl_O G_Refl_pronrefl_Ad G_Refl_pronrefl_Ch G_Refl_pronrefl_Pos
GRAMMAR	inexistent inflection	verb inflection	G_Inflect_verb_inexist
GRAMMAR	wrong inflection	case	G_Morphol_case_wrong
GRAMMAR	verb	tense	G_Verb_tns
ORTHOGRAPHY	grapheme	transposition	O_Graph_trans
ORTHOGRAPHY	grapheme		O_Graph_act_O O_Graph_act_Ad O_Graph_act_Ch
ORTHOGRAPHY	word boundary		O_Wordbd_Split O_Wordbd_Merge
ORTHOGRAPHY	punctuation		O_Punct_O O_Punct_Ad O_Punct_Ch

			O_Punct_Pos
EA2			
VOCABULARY	Formulaic sequence (FS)	collocation	V_FS_colloc
VOCABULARY	General	non-existing form	V_form_word_fs_nonexist
VOCABULARY	Semantic error	connotation (attitude)	V_semcon_att_word_fs
VOCABULARY	Form error	word formation	V_Word_form_deriv V_Word_form_comp
COHERENCE	connectors	connector accuracy	C_Con_accur_O_0 C_Con_accur_Ad_0 … C_Con_accur_O_1 C_Con_accur_Ad_1 …
COHERENCE/ COHESION	coherence	reference	C_Coh_ref
SOCIOLINGUISTIC APPROPRIATENESS	text-type specific	opening and closing formulae	S_Txt_opcl
SOCIOLINGUISTIC APPROPRIATENESS	formality	general: inadequate formality	S_Form_gen
atd.			

Anotace korpusu Merlin zároveň kombinuje, jak je u anotovaných žákovských korpusů obvyklé, značkování založené na formálních typech alternace zdrojového textu (chybějící element, přebývající element, chybně spojené elementy ap.) a hierarchicky strukturovanou lingvistickou klasifikaci. Viz Štindlová (2013: 79n.).

3.1 Anotační proces

Anotace žákovského jazyka se obvykle zakládá na rekonstrukci promluvy studenta v cílovém jazyce s minimálními zásahy (Ellis, 1994: 54), tedy v jistém smyslu na interpretaci žákova sdělení, která je formulována v tzv. cílové hypotéze (*target hypothesis, TH*). Chybu, resp. odchylku od standardu je pak možné rozpoznat a kategorizovat na základě rozdílu mezi žákovským projevem a touto rekonstruovanou podobou, která je v souladu s pravidly cílového jazyka. Srov. i Hana a kol. (2013).

Proces anotace korpusu Merlin je rozdělen do několika fází. V první řadě se projevy nerodilých mluvčích analyzují a značkují na rovině ortografické a gramatické, následně pak na rovinách vyšších (lexikologická rovina, rovina sociolingvistických aspektů a rovina koheze a koherence textu).[10] Dále jsou texty nerodilých mluvčích podrobeny automatické lemmatizaci a morfologické i syntaktické anotaci.[11]

3.2 Cílová hypotéza a chybová anotace

Jak jsme již zmínili, během anotace dat v projektu Merlin formulujeme vždy dvě cílové hypotézy.[12] První cílová hypotéza (TH1) je minimální rekonstrukcí původního textu a vzniká v souvislosti s opravou ortografických a gramatických chyb. Srovnání zdrojového textu s TH1 se odráží v chybové anotaci (*error annotation 1,* EA1) tím, že jsou konkrétním jevům přiřazovány značky, jež vystihují dané porušení ortografických a gramatických pravidel jazyka, tento typ anotace se tedy soustředí na správnost jazykového vyjádření. Chyby plynoucí z nesouladu se zásadami jazyka týkajícími se vyšších rovin, tj. slovní zásoby, koheze a koherence textu či sociolingvistického a pragmatického aspektu, se anotují v rámci následné

[10] Tato manuální dvoukolová anotace se provádí pomocí nástrojů MMAX2 a Falko Excel Addin MMAX2 (viz http://mmax2.net) a Falko Excel Addin (viz https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/marc).

[11] Automatická anotace se realizuje v rámci modulární architektury UIMA, která umožňuje integraci řady nástrojů určených pro zpracování přirozeného jazyka (viz http://uima.apache.org).

[12] V tomto směru je anotační schéma korpusu Merlin významně inspirováno podobou německého korpusu FALKO (viz Reznicek, Lüdeling a kol., 2012).

chybové anotace (*error annotation 2*, EA2), která zohledňuje druhou cílovou hypotézu (TH2), jež je výraznou interpretační rekonstrukcí a akcentuje přiměřenost jazykového projevu. Blíže viz i Štindlová a kol. (2014).
Konkrétní příklad rozdílů mezi TH1/EA1 a TH2/EA2 je uveden v tabulce 4. Tabulková forma, využívaná zejména v pilotních fázích anotace, je založena na rozdělení žákovského projevu na jednotlivé tokeny (1 token = 1 buňka) a na zpracování každého z tokenů postupně v rámci všech anotačních úrovní.

Tabulka 4: Ukázka značkování

tok	TH1	EA1	EA1	TH2	EA2
Tibor	**Tibor**			**Tibor**	
je	**je**			**je**	
z	**z**			**z**	
Madarsku	**Maďarska**	O_Graph_graphgen_O	G_Morphol_case_wrong	**Maďarska**	
a	**a**			**a**	
studuje	**studuje**			**studuje**	
v	**v**		O_Graph_act_Ad	**v**	V_Wordform_deriv
praze	**Praze**	O_Capit		**Praze**	
na	**na**			**na**	
filozofske	**filozofské**	O_Graph_graphgen_O		**filozofické**	V_semdenot_word/fs_1
fakultě	**fakultě**			**fakultě**	
.	**.**			.	

Manuální anotace korpusu Merlin je tedy dvousložková, zachycuje charakter povrchové alternace a zároveň v hierarchické struktuře mapuje lingvisticky klasifikované rysy žákovského jazyka. Posloupnost lingvistické anotace představuje

značkování na třech rovinách: vymezení kategorie (tj. ortografie, gramatika, slovní zásoba, koherence/koheze, sociolingvistická přiměřenost, pragmatika), vymezení typu (např. chybná flexe jako subkategorie u gramatiky) a v některých případech i podrobnější klasifikace odchylky (např. rod u typu chybná flexe). Srov. příklad 1.

Příklad 1:

Přepsaný žákovský text:

Přišti semestru budu psát diplomovou prace.

Cílová hypotéza s doplněním tagů:

Příští (O_Graph_act_O, O_Graph_act_O) *semestr* (G_Morphol_case_wrong) *budu psát diplomovou práci* (G_Morphol_case_wrong, O_Graph_act_O).

O_Graph_act_O = ortografie : grafém : chybí diakritika
G_Morphol_case_wrong = gramatika : flexe : chyba v pádu
G_Prep_Ch = gramatika : prepozice : záměna prepozice

4. Problémy anotace

V následujícím oddíle shrnujeme některé konkrétní problémy, na něž anotátoři českého subkorpusu naráželi.[13] Uvedené příklady vymezují vlastně v širším pojetí dva typy komplikací, resp. jejich rozlišování: jednak jde o jevy skutečně gramaticky a ortograficky chybné, jednak o odchylky od úzu, které nejsou agramatické, ale pro rodilého mluvčího nejsou v daném kontextu adekvátně formulovány.

4.1 Pravidlo minimálních zásahů do textu

[13] Na anotaci české části korpusu Merlin se v první fázi (TH1) podílelo 8 anotátorů, ve druhé fázi (TH2) 5 anotátorů. Kontrola anotace byla kromě jiného zajištěna také za prvé jednotným konsenzem anotátorů v případě neshodných cílových hypotéz či anotací u vybraného vzorku 10 textů, za druhé dvojí anotací dalších 10 textů, jež umožnila detailní kontrolu mezianotátorské shody (IAA).

Vzhledem k tomu, že každé definování aspektů žákovského jazyka jako chybných je vždy nutně hypotetické a často se nabízejí různé interpretace projevu nerodilých mluvčích, a protože zároveň další analýza žákovského jazyka závisí z velké části právě na této interpretaci, platí při zpracovávání žákovských textů, že cílová hypotéza je formulována na základě přísně vymezených pravidel, z nichž zásadní je pravidlo minimálního zásahu do podoby zdrojového textu (srov. Lüdeling, 2008). Anotátoři jako rodilí mluvčí intuitivně vnímají nedostatky textu, obtížněji se však volí cílová hypotéza takového charakteru, aby splňovala kritérium minimálního zásahu. To se ukázalo při zpětné revizi cílových hypotéz během fáze značkování chyb. Nejčastějšími problémovými případy byla spojení, která mají rodilí mluvčí v běžném úzu. Anotátoři opravovali jako ortograficky, resp. gramaticky chybné takové případy, jež byly v jistém smyslu neobvykle (příznakově) formulovány. Odlišná podoba těchto frází není však gramaticky nesprávná, nýbrž spíše v daném kontextu netypická. Proto jsou formulace tohoto typu (příklad 2) ponechány na TH1 beze změny a jejich reformulace se objevuje až na TH2.

Příklad 2:

*Vzala si své **oblíbenější** botičky.*

*Potřebuju vědět, kdy to začíná, **abych mohl vědět,** kdy mám odejít.*

Požadavek minimálního zásahu do textu úzce souvisí s pravidlem, aby anotátor nedovozoval intenci autora textu, tj. nesnažil se odhadovat, co chtěl nerodilý mluvčí říct. To však může být vzhledem k faktu, že cílová hypotéza je do jisté míry anotátorskou interpretací, docela problematické. V následujících příkladech si ukážeme různé typy výchozího studentova textu, které mohou působit komplikace.

1. Sdělení je obsahově jasné (a gramaticky korektní), zásah anotátora není nutný.

Příklad 3:

Myslíš, že budeš končit sraz v 5 hodin?

2. Sdělení je obsahově srozumitelné, ale gramaticky nestandardní, zásah anotátora je nutný.

Příklad 4:

> ***Zvadríš všechny a hlavně Petra.***
>
> Anotátor původně formuloval TH1 jako *Pozdravuj všechny a hlavně Petra*. Revize upřednostnila formulaci s menším zásahem do originální podoby *Zdravíš všechny a hlavně Petra*.

3. Zcela nesrozumitelné úseky, zásah anotátora je spekulací.

Příklad 5:

> ***Sle fotky?***
>
> Navrhované TH1 (*Pošle fotky? / Pošleš fotky? / Šli fotky?*) byly odmítnuty jako spekulativní, žádné řešení nelze hodnotit jako uspokojivé. Pro takové nejasné případy je určen tag pro neexistující formu.
>
> ***Když to jde, ubytování bylo fajn.***
>
> Tato zdánlivě nekomplikovaná věta (všechna slova jsou srozumitelná a pravopisně i gramaticky bezproblémová) je interpretačně nesrozumitelná. Anotátoři nedošli ke shodě ohledně TH1 (*Když to šlo, ubytování by bylo fajn / Když to jde, ubytování by bylo fajn* / *Když o to jde, ubytování bylo fajn*.). V tomto případě, kdy by formulace cílové hypotézy byla podobně jako v příkladu výše postavena na spekulativním vyvozování intence nerodilého mluvčího, ponecháváme řešení až na TH2.

Uvedené příklady zastupují řadu těch, u kterých byla provedena oprava původně formulované TH1. I přes tuto revizi musíme konstatovat, že výsledek některých případů je ve značné míře interpretační.

4.2 Chyby na hranici ortografie a gramatiky

Vzhledem k tomu, že cílem TH1 je provést anotaci všech chyb gramatických a ortografických, spektrum užívaných značek bylo skutečně široké. Anotační schéma je velmi podrobně rozpracované, reflektuje velké množství možných chyb a obsahuje řadu vzorových příkladů toho, pro kterou situaci je vhodná která značka. Avšak jako každý uměle vytvořený konstrukt snažící se popsat jazyk (v tomto případě jazyky tři) není ideální. V průběhu anotace se objevily případy, kdy bylo velmi obtížné konkrétní chybě adekvátní značku přiřadit, protože se chyba buď pohybovala na hranici dvou možných lingvistických domén, nebo stála mimo všechny ve schématu zohledněné gramatické kategorie. V takových případech bylo nutné po společné diskusi anotátorů zvolit co nejlepší řešení, resp. řešení s nejvyšší shodou. Toto řešení pak bylo zahrnuto do anotačního schématu a aplikováno u všech obdobných jevů.

Jedním z hlavních problémů bylo rozhodování, zda chybu v koncovce ohebných slovních druhů značit jako chybu ortografickou, nebo gramatickou, resp. v jakých případech preferovat jaké řešení. Společné anotační schéma požadovalo, aby byla jakákoli chyba v koncovce flektivních jmen značkována jako gramatická. V případě českého subkorpusu se však rozvinula velkým množstvím reálného materiálu podložená diskuse o tom, zda je žádoucí problematiku takto zjednodušovat. V řadě příkladů nestandardních koncovek je totiž zcela proti intuici rodilého mluvčího přiřazovat takové chybě morfologický tag. Diskutované případy se týkaly především problémů s diakritikou (chybějící, nadbytečná, záměna) a záměn vokálů, např. *tve > tvé, kratke > krátké, vašim > vašem*, *dětí > děti, mužů > můžu,* aj. Pro tyto konkrétní příklady byl pro češtinu sestaven algoritmus, který lze úspěšně aplikovat na řadu – i když ne na všechny – z problémových jevů (viz Štindlová a kol., 2014).

4.3 Problémy s rozšířenou cílovou hypotézou (TH2)

Nejasnosti v procesu manuální anotace se netýkaly pouze značkování na rovině TH1. Při formulaci TH2 se objevila řada komplikací při rozhodování, zda daný jev spadá do oblasti nižší jako problém gramatický, nebo vyšší jako otázka lexikální. K

hlavním otázkám značkování na TH2 patří problematika vidu, uzuálních kolokací a stylové adekvátnosti. Příklady uvádíme u příkladu 6.

Příklad 6:

Chtěl bych tě pozvat ke mně doma.

V procesu anotace jsme se museli rozhodnout, zda bude tento typ chyby značkován jako problém valenční, tedy problém TH1, nebo jako lexikální záměna, tedy TH2.

Docházím na nádraží, samozřejmě.

Zde se jedná o vidovou záměnu *(dojdu).* Diskuse se týkala zařazení chyb vidových záměn na TH2.

Dobrý den, dámy a pánové, máte volný pokoj?

V původním anotačním schématu nebyla k dispozici značka odpovídající stylové neadekvátnosti tohoto typu (oslovení neodpovídající kontextu promluvy), bylo tedy nutné ji doplnit.

Závěr

Projekt Merlin přispívá v oblasti SERRJ k detekci jazykových rysů, které odpovídají úrovni znalosti jazyka na příslušné referenční úrovni. Korpus, jenž tvoří základ tohoto projektu, zahrnuje tři rozdílné jazyky a je anotován na základě robustního výběru anotovaných jevů. Vzhledem k cíli, který si projekt vytyčil a v souvislosti s novou zkušeností anotování tří jazyků podle jednotných pravidel se jako zcela zásadní jeví spolehlivost anotace. Ta byla zajišťována podrobným proškolováním jednotlivých anotátorů, jimž jako podpora sloužila přehledná a podrobná dokumentace včetně anotačního manuálu s příklady. Velmi přínosnou byla kromě

supervize anotací také dvojí anotace vzorku korpusu a ověření její spolehlivosti pomocí výpočtu mezianotátorské shody (IAA).

Literatura:

ABEL, A. – WISNIEWSKI, K. – NICOLAS, L. – BOYD, A. – HANA, J. – MEURERS, D. (*v tisku)*: A Trilingual Learner Corpus illustrating European Reference Levels. In: *Proceeding of the Learner Corpus Research 2013 Conference (LCR2013)*, Bergen 27. – 29. 09. 2013.

CORDER, S. P. (1981): *Error Analysis and Interlanguage.* Oxford: Oxford University Press.

ELLIS, R. (1994). *The study of Second Language Acquisition.* Oxford: Oxford University Press, 1994.

HANA, J. – ROSEN, A. – ŠTINDLOVÁ, B. – FELDMAN, A. (2013). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, s. 1–28.

LÜDELING, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Walter, M. – Grommes, P. (eds.): *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitsprachenerwerbsforschung.* Tübingen: Niemeyer, 119–140.

REZNICEK, M. – LÜDELING, A. – KRUMMES, C. – SCHWANTUSCHKE, F. – WALTER, M. – SCHMIDT, K. – HIRSCHMANN, H. – ANDREAS,T. (2012). *Das Falko-Handbuch. Korpusaufbau und Annotatione*n. Version 2.01. HU Berlin.

ŠTINDLOVÁ, B. (2013). *Žákovský korpus češtiny a evaluace jeho chybové anotace.* Varia, FF UK: Praha.

ŠTINDLOVÁ, B. – ČURDOVÁ, V. – KLIMEŠOVÁ, P. – LEVOROVÁ, E. (2014). Žákovský korpus Merlin: jazykové úrovně a trojjazyčná chybová anotace. In: Práce s chybou ve výuce cizích jazyků (včetně češtiny pro cizince). Sborník z mezinárodní konference, 17.-18.6.2014. Praha: ÚJOP UK, s. 140–148.

WISNIEWSKI, K. – SCHÖNE, K. – NICOLAS, L. – VETTORI, C. – BOYD, A. – MEURERS, D. – HANA, J. – ABEL, A. (2013). MERLIN: An online trilingual learner corpus empirically grounding the CEFR Reference Levels in authentic data. In: *ICT for Language Learning*, *Conference Proceedings* 2013. Firenze: Libreriauniversitaria.