CzeSL-man v1 searchable – korpus češtiny nerodilých mluvčích s ruční chybovou anotací podle zjednodušeného víceúrovňového schématu

CzeSL-man v1 searchable obsahuje přepisy textů vytvořených nerodilými mluvčími češtiny. Je to ručně anotovaná část textů z korpusu *CzeSL-SGT*, která odpovídá textům z *CzeSL-man v0* od zahraničních studentů češtiny. Obsahově se *CzeSL-man v1 searchable* kryje s korpusem *CzeSL-man v1 downloadable*.

Ruční chybová anotace je zjednodušená verze dvoustupňového (2T) anotačního schématu, vytvořeného pro projekt *CzeSL* (viz dále §2). Anotace obsahuje opravy zdrojového textu – cílovou hypotézu (ruční), typy chyb (ruční), morfosyntaktické kategorie a lemmata pro opravený text (automatické) a závislostní syntaktickou strukturu a funkce opraveného textu (automatické). Většina textů je vybavena metadaty o autorovi a textu.

Tento korpus lze prohledávat on-line pomocí *KonText*u, vyhledávacího rozhraní Českého národního korpusu.[1] Korpus se liší od *CzeSL-man v0* a *CzeSL-man v1 downloadable* ve dvou aspektech: (i) neexistují žádné texty s alternativní chybovou anotací, každý text je anotován jen jedním anotátorem, a (ii) dvoustupňové anotační schéma je zjednodušeno tak, aby konvenovalo vyhledávacímu nástroji, který je orientován na anotaci po tokenech (slovech). Jinak jsou obsah i metadata shodné s korpusem *CzeSL-man v1 downloadable* a vyhledávací možnosti jsou podobné jako u *CzeSL-SGT*.

Další informace o projektu žákovských korpusů *CzeSL*, včetně přehledu všech verzí žákovského korpusu *CzeSL* s odkazy na možnosti vyhledávání nebo stahování, viz http://utkl.ff.cuni.cz/learncorp/ a Rosen et al. (2020).

1 Výběr textů

Korpus obsahuje přepisy esejů nerodilých mluvčích češtiny, napsaných v letech 2009–2013. Jde celkem o 645 textů od rodilých mluvčích 32 různých jazyků. Texty obsahují 128 tisíc tokenů. [2]

Počet textů podle prvního jazyka a úrovně znalostí češtiny uvádí následující tabulka (IE = neslovanský indoevropský, nIE = neindoevropský, S = slovanský,? = neznámý).

Texty jsou anonymizovány nahrazením osobních jmen vhodnými tvary jmen *Adam* a *Eva*. Místní názvy (ulice, vesnice, menší města) a další potenciálně citlivá data jsou nahrazena řetězcem `QQQ`. Nečitelné znaky nebo slova jsou přepsána jako `XXX`.

2 Anotace

2.1 Dvoustupňové anotační schéma

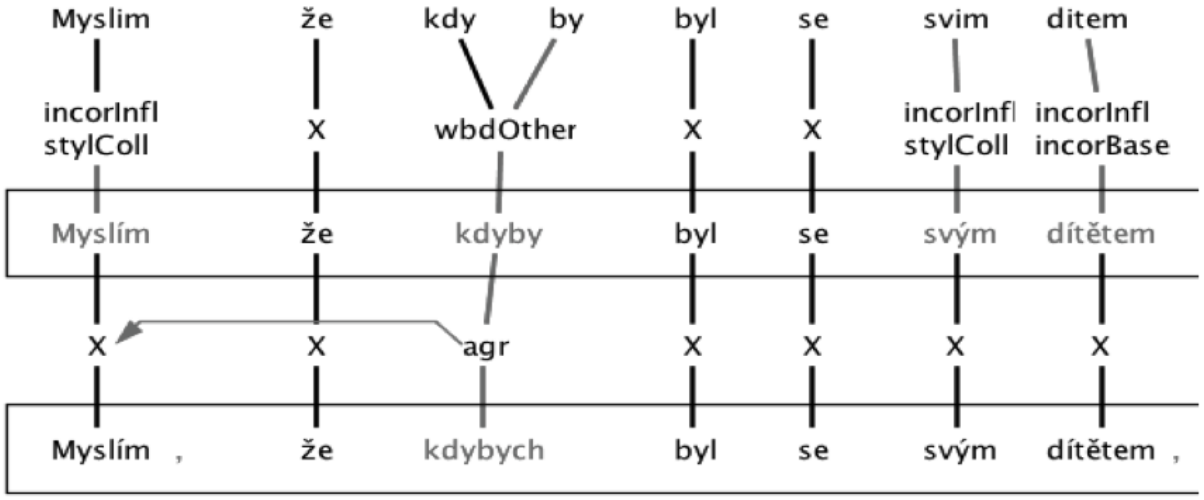
Původní dvoustupňové (2T) anotační schéma se skládá ze tří vzájemně propojených úrovní - viz obr. 1:

[1] https://kontext.korpus.cz/first_form?corpname=czesl-man

[2] Počet tokenů v korpusu udávaný nástrojem *KonText* je o něco nižší (124 tisíc), protože se počítají tokeny v opravené verzi korpusu.

	IE	nIE	S	?	celkem
A1	6	4	49		59
A1+		3			3
A2	26	67	18		111
A2+	9	59	81		149
B1	26	30	123		179
B2	11	15	102		128
C1		2	10		12
nezjištěno				4	4
celkem	78	180	383	4	645

Tabulka 1: Počet textů podle prvního jazyka (rodiny jazyků) a úrovně znalostí



Obrázek 1: Příklad dvoustupňového anotačního schématu

- Rovina 0 (T0) – anonymizovaný přepis ručně psaného originálu, včetně některých vlastností rukopisu (varianty, nečitelné řetězce)
- Rovina 1 (T1) – tvary, které jsou nesprávné samy o sobě, mimo kontext, jsou opraveny. Výsledkem je posloupnost správných tvarů, i když věta jako celek správně být nemusí. Tato rovina v *CzeSLman v1 searchable* chybí.
- Rovina 2 (T2) – řeší všechny ostatní typy chyb (valence, shoda, slovosled atd.)

Opravy na T2 se týkají chyb ve shodě, valenci, analytických tvarech, zájmenném odkazování, záporové shodě, vidu, slovesném čase, lexiku nebo idiomech a také ve slovosledu. V tomto korpusu jsou opravy na T1 už zahrnury pod T2. Tabulka 2 uvádí seznam typů chyb ručně anotovaných na T2. Automaticky určené chyby se týkají slovosledu a analytických tvarů *vbx*.

Vztah mezi původním tvarem a jeho (postupnou) opravou je vyjádřen explicitně. Tvary na sousedních úrovních mají obvykle vztah 1:1, ale slova se mohou také spojovat (*kdy by* jako na obr. 1), rodělovat, odstraňovat nebo přidávat. Tyto vztahy mohou propojit i libovolný počet slov, která netvoří spojitou řadu.

2.2 Zjednodušené dvoustupňové schéma

Hlavním rysem anotace této verze korpusu je obrácení zdrojového textu a jeho anotace. S opraveným textem na T2 se nakládá jako s výchozím textem pro anotaci. Tokeny tohoto korpusu tedy jsou slova na T2. Původní text je pak anotací těchto tokenů. Každý token opraveného textu je označen atributy podle

Tyo chyby	Popis	Příklad
agr	chyba shody	to jsou *hezké* chlapci; Jana *čtu*
dep	chyba valence	bojí se *pes*; otázka *čas*
ref	chyba v zájmenném odkazu	dal jsem to jemu i *jejího* bratrovi
vbx	chyba v analytickém tvaru slovesa nebo složeném přísudku	musíš *přijdeš*; kluci *jsou* běhali
rflx	chyba ve zvratném výrazu	dívá na televizi; Pavel *si* raduje
neg	chyba v negaci	žádný to *ví*; *půjdu ne* do školy
lex	lexikální nebo frazeologická chyba	jsem *ruská*; dopadlo to *přírodně*
use	chyba v užití gramatické kategorie	pošta je *nejvíc blízko*
sec	sekundární chyba	stará se o *nemocných dětech*
stylColl	hovorový výraz	viděli jsme *hezký* holky
stylOther	knižní, nářeční, slangový, hyperkorektní výraz	zvedl se mi *kufr*
stylMark	redundant discourse marker	*no*; *teda*; *jo*
disr	rozvrácená konstrukce	*kratka jakost vyborné ženy*
problem	problematický případ	

Tabulka 2: Ručně určené chyby na rovině T2

odpovídající tvaru ze zdrojového textu a chybovou značku na T2. V této anotaci se ztratí všechny opravy a chybové značky z T1, a vztahy mezi tokeny T0 a T2 se zjednodušují na 1:1. Důsledkem je i to, že se v anotaci nezachovají všechna slova z T0. Také jejich pořadí může ovlivněno slovosledem na T2.

2.3 Lingvistická anotace

Opravené tvary jsou označkovány morfosyntaktickými kategoriemi a lemmaty pomocí standardních nástrojů. Každému slovu jsou přiřazeny lemma a tag ze standardní morfologické sady značek Hajič (2004), viz https://wiki.korpus.cz/doku.php/seznamy:tagy.

Radikální zjednodušení dvoustupňového schématu anotace chyb umožnilo analýzu cílové hypotézy na T2 podobným způsobem jako v jiných českých korpusech, které lze prohledávat v *KonText*u, například *SYN2015*. Seznam atributů souvisejících se syntaxí přiřazených každému tokenu viz tab. 4, podrobnější popis najdete zde: https://wiki.korpus.cz/doku.php/pojmy:syntakticka_analyza.

2.4 Jak anotaci využít při hledání

V tab. 3 jsou uvedeny atributy představující základní chybovou a morfosyntaktickou anotaci v této verzi. Atributy související se syntaxí jsou uvedeny v tab. 4. Stejně jako v *CzeSL-SGT* lze v dotazech a vizualizaci výsledků využít dynamické atributy, odvozené z morfosyntaktické značky, viz tab. 5.

`err`	chybová značka slova na T2 (`word`)
`word0`	tvar na T0 (zdrojový text)
`lemma0`	lemma `word0`; rovno `word0`, pokud slovo nebylo rozpoznáno
`tag0`	morfologická značka `word0`; pokud slovo nebylo rozpoznáno: `X@----------`
`word`	opravený tvar na T2; je-li slovo rozpoznáno jako správné, rovno `word0`
`lemma`	lemma `word`
`tag`	morfologická značka `word`

Tabulka 3: Morfosyntaktická a chybová anotace korpusu *CzeSL-man v1 searchable* vyjádřená atributy tokenů v *KonText*u

3 Metadata

Metadata jsou uvedena jako v *CzeSL-SGT* a jsou v angličtině i v *KonText*u. 15 položek se týká autora textu a 15 položek je o textu samotném. Seznam všech položek a hodnoty česky a anglicky najdete zde:

proc	krok v disambiguaci, který rozhodl o dané analýze
afun	syntaktická funkce
parent	relativní odkaz na rodiče
eparent	relativní odkaz na efektivního rodiče
prep	předložka jako rodič
p_tag	značka rodiče
p_lemma	lemma rodiče
p_afun	syntaktická funkce rodiče
ep_tag	značka efektivního rodiče
ep_lemma	lemma efektivního rodiče
ep_afun	syntaktická funkce efektivního rodiče
lc	slovo T2 malými písmeny
lemma_lc	lemma T2 malými písmeny
p_k	kategorie rodiče (POS)
p_c	pád rodiče

Tabulka 4: Atributy související se syntaxí použité v *KonText*u pro korpus *CzeSL-man v1 searchable*

k0, k	slovní druh (pozice 1 značky)
s0, s	slovní poddruh (pozice 2 značky)
g0, g	rod (pozice 3 značky)
n0, n	číslo (pozice 4 značky)
c0, c	pád (pozice 5 značky)
p0, p	osoba (pozice 8 značky)

Tabulka 5: Dynamické atributy použité v *KonText*u pro *CzeSL-man v1 searchable*

http://utkl.ff.cuni.cz/~rosen/public/meta_attr_vals.html.

4 Poděkování

Práce na tomto projektu podpořily v letech 2009–2012 grant Evropských strukturálních fondů *Inovace vzdělávání v oboru čeština jako druhý jazyk*, reg. č. CZ.1.07/2.2.00/07.0119 a *PRVOUK*, program podpory výzkumu Univerzity Karlovy také projekt *Český národní korpus*, v rámci programu MŠMT *Projekty velkých infrastruktur pro vědu, výzkum a inovace* (2012-2015, projekt č. LM2011023).

Reference

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague.

Rosen, A., Hana, J., Hladká, B., Jelínek, T., Škodová, S., & Štindlová, B. (2020). *Compiling and annotating a learner corpus for a morphologically rich language – CzeSL, a corpus of non-native Czech*. Karolinum, Charles University Press, Praha.

Alexandr Rosen, 11 November 2020