

CzeSL-SGT – korpus češtiny nerodilých mluvčích s automaticky provedenou anotací

Žákovský korpus CzeSL-SGT (*Czech as a Second Language with Spelling, Grammar and Tags*) obsahuje přepisy písemných prací nerodilých mluvčích češtiny. Navazuje tak na část *ciz* korpusu CzeSL-plain (<http://wiki.korpus.cz/doku.php/cnk:czesl-plain>): obsahuje její jazyková data, obsahuje však navíc další texty, sebrané v roce 2013. Podrobněji viz část 1.

Slovní tvary jsou označeny slovním druhem, morfologickými kategoriemi a základním tvarem (lemmatem). Některé tvary jsou opraveny a výsledná podoba textu znovu slovnědruhově a morfologicky označena. Na základě porovnání původní a opravené podoby tvarů je stanoven druh chyby. Všechny tyto údaje jsou určeny automaticky, je tedy třeba počítat s nepřesnostmi a omyly. Podrobněji viz část 2.

Nové texty jsou opatřeny údaji (metadaty) o autorovi a typu textu, ta byla nově doplněna i u velké většiny textů původních. Podrobněji viz část 3.

Korpus lze prohledávat on-line přes vyhledávací rozhraní Českého národního korpusu (<https://kontext.korpus.cz/run.cgi/first?corpname=czesl-sgt>), viz §3.3, nebo ho lze získat celý jako data ze serveru LINDAT (<http://hdl.handle.net/11234/1-162>), viz §3.4. Aktuální verze korpusu má pořadové číslo 2 a liší se od původní verze odstraněním některých formálních chyb a ve verzi na serveru LINDAT také formátem – celý korpus má nyní podobu XML dokumentu, ve kterém je veškerá anotace vyjádřena pomocí XML atributů.

Další informace o žákovském korpusu *CzeSL* a o projektu akvizičních korpusů *AKCES*, k nimž se řadí, najdete na stránce <http://utkl.ff.cuni.cz/learncorp/> a <http://akces.ff.cuni.cz/>, včetně seznamu literatury, viz např. Šebesta & Škodová (2012) nebo Štindlová (2013).

1 Výběr textů

- přepisy písemných prací nerodilých mluvčích z kursů češtiny v letech 2009 až 2013
- rozšiřuje část *ciz* korpusu CzeSL-plain (<http://ucnk.ff.cuni.cz/czesl-plain.php>) o texty sebrané v roce 2013
- 8 617 textů od 1965 různých autorů s 54 různými prvními jazyky
- 111 tisíc vět, 958 tisíc slov, 1 148 tisíc pozic (tokenů)
- bez transkripčních značek, s rekonstrukcí konečné autorské podoby textu

2 Anotace

Každá pozice v korpusu (slovní tvar i interpunkční znaménko) je označena těmito atributy:

- `word` – původní slovní tvar
- `lemma` – lemma původního tvaru, pokud tvar není rozpoznán, je lemma totožné s původním tvarem
- `tag` – slovnědruhová a morfologická značka původního tvaru, pokud tvar není rozpoznán, je uvedena značka pro neznámé slovo: `X@-----`
- `word1` – opravený tvar, pokud byl původní tvar vyhodnocen jako správný, je zde uveden tvar původní
- `lemma1` – lemma určené na základě slovního tvaru a jeho kontextu v opraveném textu
- `tag1` – slovnědruhová a morfologická značka, určená na základě slovního tvaru a jeho kontextu v opraveném textu

- `gs` – údaj o tom, zda případná chyba byla vyhodnocena jako pravopisná (S) nebo gramatická (G); gramatická chyba se obvykle vyznačuje tím, že původní slovní tvar byl rozpoznán
- `err` – typ chyby, určený na základě porovnání původního a opraveného tvaru, podrobný popis viz <http://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1.html>.

word	lemma	tag	word1	lemmal	tag1	gs	err
Tén	Tén	X@-----	Ten	ten	PDYS1-----	S	Quant1
pes	pes	NNMS1-----A----	pes	pes	NNMS1-----A----		
míluje	míluje	X@-----	miluje	milovat	VB-S---3P-AA---	S	Quant1
svého	svého	X@-----	svého	svůj	P8MS4-----	S	Voiced
kamarada	kamarada	X@-----	kamaráda	kamarád	NNMS4-----A----	S	Quant0
-	-	Z:-----	-	-	Z:-----		
člověka	člověk	NNMS2-----A----	člověka	člověk	NNMS4-----A----		
.	.	Z:-----	.	.	Z:-----		

Tabulka 1: Ukázka anotace jedné věty

Při hledání v korpusu přes rozhraní Českého národního korpusu lze kromě těchto atributů lze využít i tzv. dynamické atributy, odvozené z některých pozic značek `tag` a `tag1`. Lze jimi v dotazu např. specifikovat hodnoty jednotlivých morfologických kategorií bez použití regulárních výrazů, požadovat identitu těchto hodnot u dvou a více slovních tvarů pro vyhledání gramatické shody nebo porovnávat hodnoty stejné kategorie u původního a opraveného tvaru. Tyto atributy jsou k dispozici pro následující kategorie původního a opraveného tvaru:

- `k`, `k1` – slovní druh (1. pozice značky)
- `s`, `s1` – detailní určení slovního druhu (2. pozice značky)
- `g`, `g1` – jmenný rod (3. pozice značky)
- `n`, `n1` – gramatické číslo (4. pozice značky)
- `c`, `c1` – jmenný pád (5. pozice značky)
- `p`, `p1` – gramatická osoba (8. pozice značky)

Popis použitých slovnědruhových a morfologických značek je uveden na stránce <http://wiki.korpus.cz/doku.php/seznamy:tagy>, pro CzeSL-SGT ve verzi z korpusu syn2000, tedy bez pozice 16 (vid) a s tzv. proměnnými za některé hodnoty.

Texty jsou anonymizovány záměnou osobních jmen za příslušné tvary jmen *Adam* a *Eva*. Názvy menších míst (ulic, vesnic, menších měst) a jiné potenciálně zneužitelné údaje jsou zaměněny za řetězec `QQQ`. Nečitelné znaky nebo slova jsou nahrazeny řetězcem `XXX`.

3 Metadata

Texty nové a velká většina starých jsou vybaveny metadaty, celkem 15 údaji o autorovi textu a 15 údaji o textu samotném. Přehled všech atributů a hodnot česky i anglicky je uveden zde: http://utkl.ff.cuni.cz/~rosen/public/meta_attr_vals.html. Počty dokumentů podle hodnot atributů najdete zde: http://utkl.ff.cuni.cz/~rosen/public/sgt_counts_by_meta.html. Význam jednotlivých položek je vysvětlen níže v §3.1 a §3.2.

Ne všechny texty jsou vybaveny všemi položkami. Např. autorství je zjištěno u 96,7 % textů, první jazyk u 86,3 % textů. Chybějící položky jsou uvedeny jako prázdná hodnota. Více hodnot u jedné položky je odděleno znakem „|“.

Metadata jsou v české a anglické verzi. Na stránkách ČNK, kde je korpus přístupný přes vyhledávací rozhraní, jsou metadata česká. Na serveru LINDAT (<http://hdl.handle.net/11858/00-097C-0000-0023-95B1-E>), odkud je možné korpus získat celý, jsou metadata anglicky.

Všechny položky se týkají příslušného textu, tedy elementu jménem `doc`.

3.1 Údaje o autorovi textu (studentovi, žákovi)

- `s_id` – identifikace; jedna hodnota: znakový řetězec, např. `TOU_H305`
- `s_pohlavi` – pohlaví; jedna z hodnot:
 - `m` – muž
 - `z` – žena
- `s_vek` – věk; jedna hodnota: celé číslo
- `s_vek_kat` – věková kategorie; jedna z hodnot:
 - `6-11`; `12-15`; `16-`
- `s_jazyk1` – první jazyk; jedna z hodnot: dvouznakový kód podle normy ISO 639-1, např. `sq` (albánština); pro jazyk chybějící ve dvouznakovém kódu je použit trojznakový kód ISO 639-3, např. `xal` (kalmyčtina) nebo `bem` (bembština), viz http://cs.wikipedia.org/wiki/Seznam_kódů_ISO_639-1 a http://en.wikipedia.org/wiki/List_of_ISO_639-3_codes.
- `s_jazyk1_skupina` – jazyková skupina prvního jazyka; jedna z hodnot:
 - `IE` – indoevropský neslovanský
 - `nIE` – neindoevropský
 - `S` – slovanský
- `s_dalsi_jazyky` – znalost dalších jazyků; jedna nebo více z hodnot: kód ISO (viz výše `s_jazyk1`)
- `s_cj_SERR` – úroveň znalosti češtiny v době záznamu; jedna z hodnot:
 - `A1`; `A1+`; `A2`; `A2+`; `B1`; `B2`; `C1`; `C2`
- `s_cj_v_rodine` – znalost češtiny v rodině; jedna nebo více z hodnot:
 - `matka`; `otec`; `partner`; `sourozenec`; `3` (3 rodinní příslušníci); `jine`; `nikdo`
- `s_roky_v_CR` – délka pobytu v ČR v letech; jedna z hodnot:
 - `-1`; `1`; `-2`; `2-`
- `s_studium_cj` – absolvované nebo právě probíhající studium; jedna nebo více z hodnot:
 - `individualni`
 - `komercni`
 - `samouk`
 - `VS` – vysoká škola
 - `zahranicni`
 - `ZS_SS` – základní a/nebo střední škola
 - `jine`
- `s_studium_cj_mesice` – délka studia češtiny v měsících; jedna z hodnot:
 - `-3`; `3-6`; `6-12`; `12-24`; `24-36`; `36-48`; `48-60`; `60-`
- `s_studium_cj_hod_tyden` – intenzita studia češtiny v hodinách za týden; jedna z hodnot:
 - `-3`; `5-15`; `15-`
- `s_ucebnice` – učebnice, ze kterých žák studoval nebo studuje; jedna nebo více z hodnot:
 - `BC` – Basic Czech
 - `CC` – Communicative Czech

- CE – Čeština pro ekonomy
- CMC – Chcete mluvit česky?
- CpC – Čeština pro cizince
- ECE – Easy Czech Elementary
- NCSS – New Czech Step by Step
- jiné
- s_bilingvni – bilingvní; jedna z hodnot:
 - ano; ne

3.2 Údaje o textu

- t_id – identifikace; jedna hodnota: znakový řetězec, např. TOU_H305_442
- t_datum – datum sběru textu; jedna hodnota: datum ve formátu RRRR-MM-DD
- t_medium – médium textu; jedna z hodnot:
 - rukopis; pc
- t_limit_minut – časový limit na napsání práce v minutách; jedna z hodnot:
 - 10; 15; 20; 30; 40; 45; 60; jiné; ne
- t_pomucka – povolená pomůcka; jedna nebo více z hodnot:
 - ano; slovník; učebnice; jina; ne
- t_zkouska – byl text napsán kvůli zkoušce?; jedna nebo více z hodnot:
 - ano; průběžná; závěrečná; ne
- t_limit_slova – rozsah textu ve slovech podle zadání; jedna z hodnot:
 - 20; 20–; 25; 30; 35–; 40; 40–; 50; 50–; 60; 60–; 70; 70–; 80; 80–; 90; 90–; 100; 100–; 120; 120–; 150; 150–; 170; 180; 200; 200–
- t_nazev – název; jedna nebo více hodnot: znakový řetězec, např. Událost, která změnila můj život
- t_tema_typ – typ tématu; jedna z hodnot:
 - obecně; speciální
- t_aktivita – aktivita před psaním práce; jedna z hodnot:
 - cvičení; diskuse; obrázek/film; slovní zásoba; jiné; ne
- t_tema_zadane – téma podle zadání; jedna z hodnot:
 - vyber z nabídky; zadane; volne; jiné
- t_postup_zadany – slohový postup podle zadání; jedna z hodnot:
 - volne; zadane
- t_postup_prevazujici – skutečný převažující slohový postup; jedna z hodnot:
 - informace; popis; úvaha; vyprávění
- t_pocet_slov – skutečný počet slov; jedna hodnota: celé číslo
- t_rozsah_slov – rozsah podle skutečného počtu slov; jedna z hodnot:
 - -50; 100–149; 150–199; 200–; 50–99

3.3 Hledání v korpusu

K hledání v korpusu lze využít jednotné vyhledávací rozhraní Českého národního korpusu (korpus.cz). Korpus czech-sgt je jedním ze Synchronních psaných korpusů a je zařazen v kategorii specializované. Ponecháme-li Typ dotazu nastavený na základní a nezměníme-li žádné další nastavení, řetězec zadaný do rámečku Dotaz vyhledá všechny věty, kde se tento řetězec objeví v původním, neopraveném textu. U pokročilejších dotazů, které obsahují odkazy na morfologické značky, lemmata, typy chyb, opravené formy a metadatové atributy, je třeba Typ dotazu změnit na CQL a/nebo modifikovat nastavení v části Specifikovat dotaz podle metainformací. Uživatelskou příručku k vyhledávacímu rozhraní najdete zde <http://wiki.korpus.cz/doku.php/manual:uvod>.

Chceme-li např. najít všechny tvary, které lze interpretovat jako tvary substantiva *Čech*, zadáme do rámečku Dotaz výraz [lemma1="Čech"]. Mezi výsledky pak najdeme i tvary jako *češy*, *Češy*, *česi*, *Česi*, *češím*, *češly*, *čehy*, *čeha*, *Češhů*, *Čechů*, *Čechá* nebo *Cechy*.

3.4 Formát textů

Tato část se týká úplných textů, dostupných na serveru LINDAT.

Údaje o celém textu a větě jsou uvedeny jako elementy XML s příslušnými atributy. Ve verzi 1 (<http://hdl.handle.net/11858/00-097C-0000-0023-95B1-E>) jsou vlastní data uvedena ve sloupcích, oddělených tabelátory, v pořadí podle tabulky 1. Ukázka viz níže:

```
<doc t_id="UJA2_PH_003" t_date="2010-12-21" t_medium="manuscript" t_limit_minutes="45" t_aid="none"
t_exam="yes|interim" t_limit_words="25" t_title="E-mail kamarádce/kamarádovi" t_topic_type="general"
t_activity="" t_topic_assigned="specified" t_genre_assigned="specified"
t_genre_predominant="informative" t_words_count="30" t_words_range="-50" s_id="UJA2_PH" s_sex="m"
s_age="17" s_age_cat="16-" s_L1="vi" s_L1_group="nIE" s_other_langs="" s_cz_SER="A1"
s_cz_in_family="" s_years_in_CzR="" s_study_cz="university"
s_study_cz_mesice="" s_study_cz_hrs_week="15-" s_textbook="NCSS" s_bilingual="no">
<s id="1">
mám mít VB-S---1P-AA--- mám mít VB-S---1P-AA---
dobře dobře Dg-----1A---- dobře dobře Dg-----1A----
. . Z:----- . . Z:-----
</s>
<s id="2">
V v RR--4----- V v RR--4-----
neděli neděle NNFS4----A---- neděli neděle NNFS4----A----
dival dival X@----- díval dívat VpYS---XR-AA--- S Quant0
jsem být VB-S---1P-AA--- jsem být VB-S---1P-AA---
se se P7-X4----- se se P7-X4-----
na na RR--6----- na na RR--6-----
televizi televize NNFS6----A---- televizi televize NNFS6----A----
a a J^----- a a J^-----
uklízěl uklízel X@----- uklízel uklízet VpYS---XR-AA--- S Quant0|Caron1
jsem být VB-S---1P-AA--- jsem být VB-S---1P-AA---
. . Z:----- . . Z:-----
</s>
<s id="3">
Ano ano TT----- Ano ano TT-----
přijdu přijít VB-S---1P-AA--- přijdu přijít VB-S---1P-AA---
se se P7-X4----- se se P7-X4-----
tebe ty PP-S2--2----- tebe ty PP-S2--2-----
do do RR--2----- do do RR--2-----
kina kino NNNS2----A---- kina kino NNNS2----A----
a a J^----- a a J^-----
taky taky Db----- taky taky Db-----
mám mít VB-S---1P-AA--- mám mít VB-S---1P-AA---
čas čas NNIS4----A---- čas čas NNIS4----A----
jen jen TT----- jen jen TT-----
večer večer Db----- večer večer Db-----
, , Z:----- , , Z:-----
večer večer Db----- večer večer Db-----
půjdeme jít VB-P---1F-AA--- půjdeme jít VB-P---1F-AA---
do do RR--2----- do do RR--2-----
kina kino NNNS2----A---- kina kino NNNS2----A----
. . Z:----- . . Z:-----
</s>
<s id="4">
Tvoje tvůj PSHS1-S2----- Tvoje tvůj PSHS1-S2-----
kamarád kamarád NNMS1----A---- kamarád kamarád NNMS1----A----
. . Z:----- . . Z:-----
```

Ve verzi 2 (<http://hdl.handle.net/11234/1-162>) je celý korpus jako jeden XML dokument. Jednotlivé texty jsou elementy div. Anotace každého slova je vyjádřena XML atributy elementu word:

```
<word lemma="dival" tag="X@-----" word1="dival" lemma="divat"
tag1="VpYS---XR-AA---" gs="S" err="Quant0">dival</word>
```

4 Poděkování

Tento korpus by nikdy nevznikl bez úsilí mnoha nerodilých mluvčích a obětavých sběračů a přepisovačů textů. Sběr a prepisy nových dat a metadat a rozsáhlé dohledávání metadat k původním textům zajišťovala a všechny práce koordinovala Kateřina Lundáková; významně k sběru nových dat přispěla Dagmar Toufarová.

Za nástroje na značkování a lemmatizaci vdčíme všem, kdo se podíleli na vývoji popisu české morfologie, slovní zásoby a taggeru (Hajič (2001), Votrubec (2006)), za nástroj pro opravu českých textů Michalu Richterovi a Milanu Strakovi (Richter (2010), Richter et al. (2012)) a za nástroj pro určení typu chyby Tomáši Jelínkovi (Jelínek et al. (2012)).

Velký dík patří také Pavlu Procházkovi z ÚČNK, který se ochotně ujal finálního technického zpracování, a Haně Skoumalové, která trpělivě řešila všechny technické problémy.

Lví podíl zásluh na koncepci tohoto korpusu má Karel Šebesta. Výsledek by se nikdy nedostavil ani bez podpory a motivace Barbory Štindlové, Svatavy Škodové a Jirky Hany.

Práce byly v letech 2009-2012 podporovány z grantu Evropských strukturálních fondů *Inovace vzdělávání v oboru čeština jako druhý jazyk*, reg. číslo CZ.1.07/2.2.00/07.0119, hlavním výstupem byl zveřejněný korpus CZESL-PLAIN; nové sběry, prepisy a doplňování metadat k textům a další práce spojené se vznikem CZeSL-SGT z Programu rozvoje vědních oblastí na Univerzitě Karlově P10 – *Lingvistika, tematický modul Osvojování a vývoj jazykové a komunikační kompetence u vybraných komunit České republiky* a z projektu *Český národní korpus*, podporovaného Ministerstvem školství České republiky v rámci programu *Projekty velkých infrastruktur pro vědu, výzkum a inovace* (2012-2015, číslo projektu LM2011023).

Reference

- Hajič, J. (2001). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Praha. 334 pp.
- Jelínek, T., Štindlová, B., Rosen, A., & Hana, J. (2012). Combining manual and automatic annotation of a learner corpus. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue – Proceedings of the 15th International Conference TSD 2012*, number 7499 in Lecture Notes in Computer Science, pages 127–134. Springer.
- Richter, M. (2010). *Pokročilý korektor češtiny [An Advanced Spell Checker of Czech]*. Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague.
- Richter, M., Straňák, P., & Rosen, A. (2012). Korektor – a system for contextual spell-checking and diacritics completion. In *Proceedings of COLING 2012: Posters*, pages 1019–1028, Mumbai, India. The COLING 2012 Organizing Committee.
- Votrubec, J. (2006). Morphological tagging based on averaged perceptron. In *WDS'06 Proceedings of Contributed Papers*, pages 191–195, Praha, Czechia. Matfyzpress, Charles University.
- Šebesta, K. & Škodová, S., editors (2012). *Čeština – cílový jazyk a korpusy*. Technická univerzita v Liberci, Liberec.
- Štindlová, B. (2013). *Žákovský korpus češtiny a evaluace jeho chybové anotace [A Learner Corpus of Czech and Evaluation of its Error Annotation]*. Univerzita Karlova v Praze, Filozofická fakulta, Praha.