



FILOZOFICKÁ FAKULTA  
Univerzita Karlova

Ústav teoretické a počítačnické lingvistiky |  
Ústav bohemistických studií

# *Zpracování žákovských textů v nástroji TEITOK*

Současný stav a výhled do  
budoucnosti

Alexandr Rosen  
Adrian Jan Zasina

Ostrava, 25. ledna 2023

# Obsah

---

- Terminologické poznámky
- Žákovské korpusy v ČR
- CzeSL-SGT
- TEITOK
- Otázky k diskusi

# Terminologické poznámky

---

- akviziční korpus
- žákovský korpus
- atriční korpus (krajanský)

Šormová & Šebesta (Eds.). (2019, s. 12–13)

- pedagogický korpus

Baker, Hardie, & McEnery (2006, s. 128)

# Terminologické poznámky

---

- žákovské texty  $\approx$  žákovská data
- žák (ang. *learner*) - "označení jinojazyčného mluvčího učícího se cizí/druhý jazyk (L2)"

Škodová (2012, s. 127)

# Žákovské korpusy v ČR

---

- CzeSL
  - ✓ CzeSL-man, CzeSL-plain, CzeSL-SGT, CzeSL-SGT-basic, CzeSL-multimodální korpus
- Korpus arabských studentů češtiny (Univerzita Ajn Šams, Káhira)
- PoLKo (UK a Varšavská univerzita)
- STANAG (Univerzita obrany)
- VESPA\_CZ (UK)
- LINDSEI\_CZ (UK)

# CzeSL-SGT

---

- typ: psaný
- velikost: 1,2 mil. pozic
- lemmatizace: ano
- morfologické značky: ano
- rok vydání: 2014

# CzeSL-SGT – problematické otázky

## ■ Jazyková vyváženost (L1)

|     | Jazyk        | Frekvence |
|-----|--------------|-----------|
| 1.  | ruština      | 684 248   |
| 2.  | čínština     | 61 974    |
| 3.  | polština     | 59 388    |
| 4.  | ukrajinština | 49 436    |
| 5.  | korejština   | 27 198    |
| 6.  | japonština   | 26 170    |
| 7.  | angličtina   | 26 058    |
| 8.  | němčina      | 25 524    |
| 9.  | kazachština  | 25 166    |
| 10. | vietnamština | 16 859    |

## ■ Popis jazykové úrovně

|    | Úroveň | Frekvence |
|----|--------|-----------|
| 1. | A1     | 684 248   |
| 2. | A2     | 61 974    |
| 3. | B1     | 59 388    |
| 4. | B2     | 49 436    |
| 5. | A2+    | 27 198    |
| 6. | A1+    | 26 170    |
| 7. | C1     | 26 058    |
| 8. | C2     | 25 524    |
| 9. | b1     | 25 166    |

# CzeSL-SGT – metadata

## ■ informace o respondentech

| s_id | s_pohlavi | s_vek | s_vek_kat | s_jazyk1 | s_jazyk1_skupina | s_cj_SER | s_bilingvni | s_dalsi_jazyky     | s_cj_v_rodine | s_roky_v_CR | s_studium_cj | s_studium_cj_mesice | s_studium_cj_hod_tyden | s_ucebnice |
|------|-----------|-------|-----------|----------|------------------|----------|-------------|--------------------|---------------|-------------|--------------|---------------------|------------------------|------------|
| HB_F | f         | 31    | 16-       | en       | IE               | A2       | no          | zh fr es ru la och | partner       | none        | university   | 12-24               | 3-5                    | CKZK       |
| HB_F | f         | 31    | 16-       | en       | IE               | A2       | no          | zh fr es ru la och | partner       | none        | university   | 12-24               | 3-5                    | CKZK       |
| HB_F | f         | 31    | 16-       | en       | IE               | A2       | no          | zh fr es ru la och | partner       | none        | university   | 12-24               | 3-5                    | CKZK       |
| HB_F | f         | 31    | 16-       | en       | IE               | A2       | no          | zh fr es ru la och | partner       | none        | university   | 12-24               | 3-5                    | CKZK       |
| JP_F | f         | 40    | 16-       | en       | IE               | B2       | no          | ru ar sk           | nobody        | none        | university   | 24-36               | none                   |            |
| JP_F | f         | 40    | 16-       | en       | IE               | B2       | no          | ru ar sk           | nobody        | none        | university   | 24-36               | none                   |            |
| JP_F | f         | 40    | 16-       | en       | IE               | B2       | no          | ru ar sk           | nobody        | none        | university   | 24-36               | none                   |            |
| JP_F | f         | 40    | 16-       | en       | IE               | B2       | no          | ru ar sk           | nobody        | none        | university   | 24-36               | none                   |            |
| JP_F | f         | 40    | 16-       | en       | IE               | B2       | no          | ru ar sk           | nobody        | none        | university   | 24-36               | none                   |            |
| JP_F | f         | 40    | 16-       | en       | IE               | B2       | no          | ru ar sk           | nobody        | none        | university   | 24-36               | none                   |            |
| CK_F | f         | 42    | 16-       | en       | IE               | A2       | no          | es fr it de        | partner       | none        | university   | 12-24               | 3-5                    | CKZK       |

## ■ informace o textech

| t_id               | t_datum    | t_medium   | t_limit_minut | t_pomucka | t_zkouska | t_limit_slova | t_nazev                | t_tema_typ | t_aktivita | t_tema_zadane | t_postup_zadany | t_postup_pt | t_pocet_slov | t_rozsah_slov |
|--------------------|------------|------------|---------------|-----------|-----------|---------------|------------------------|------------|------------|---------------|-----------------|-------------|--------------|---------------|
| AZ_ENG_DOP_HB_F_01 | 2022-10-30 | typed      | none          | textbook  | n/a       | 150           | Dopis z návštěvy Prahy | general    | none       | specified     | specified       | informative | 179          | 150-199       |
| AZ_ENG_MIS_HB_F_02 | 2022-11-19 | typed      | none          | textbook  | n/a       | 150           | Nejdivnější místo      | general    | none       | specified     | specified       | descriptive | 104          | 100-149       |
| AZ_ENG_MED_HB_F_03 | 2022-11-20 | typed      | none          | textbook  | n/a       | 150           | Média                  | specific   | none       | specified     | specified       | argumentat  | 155          | 150-199       |
| AZ_ENG_300_HB_F_04 |            |            | none          |           | n/a       | 150           | Rok 3000               | general    | none       | specified     | specified       | narrative   |              |               |
| AZ_ENG_DOP_JP_F_01 | 2022-10-29 | manuscript | none          | none      | n/a       | 150           | Dopis z návštěvy Prahy | general    | none       | specified     | specified       | informative | 234          | 200-          |
| AZ_ENG_MIS_JP_F_02 | 2022-11-16 | typed      | none          | none      | n/a       | 150           | Nejdivnější místo      | general    | none       | specified     | specified       | descriptive | 125          | 100-149       |
| AZ_ENG_MED_JP_F_03 | 2022-11-02 | typed      | none          | none      | n/a       | 150           | Média                  | specific   | none       | specified     | specified       | argumentat  | 174          | 150-199       |
| AZ_ENG_300_JP_F_04 | 2022-11-16 | typed      | none          | none      | n/a       | 150           | Rok 3000               | general    | none       | specified     | specified       | narrative   | 136          | 100-149       |
| AZ_ENG_OML_JP_F_05 | 2022-12-01 | typed      | none          | none      | n/a       | 150           | E-mail s omluvou       | general    | none       | specified     | specified       | informative | 133          | 100-149       |
| AZ_ENG_MOT_JP_F_06 | 2022-12-01 | typed      | none          | none      | n/a       | 150           | Motivační dopis        | general    | none       | specified     | specified       | informative | 107          | 100-149       |
| AZ_ENG_DOP_CK_F_01 | 2022-11-03 | typed      | none          | ?         | n/a       | 150           | Dopis z návštěvy Prahy | general    | none       | specified     | specified       | informative | 150          | 150-199       |

# CzeSL-SGT – systematizace

---

## CONSENT

I ..... hereby acknowledge consent and agree to hand over my texts for the purpose of 1) their use in teaching and research, 2) preparation of teaching materials, and 3) compilation of a learner corpus. At the same time, I agree to make my texts publicly available within the projects managed by Charles University and The University of Sheffield.

In order to preserve the author's anonymity, Charles University and the University of Sheffield undertake that the author's personal data will not be included in the texts with the exception of general information given in the questionnaire below, which will not be sufficient to identify the author personally; the author also agrees with making this information publicly available. The topics of the written tasks has been designed in such a way that no other data revealing the author's identity could be found in the text. Any sensitive data will be anonymised.

The author of the texts has been informed of all planned uses of the texts for the purpose of academic projects and academic teaching, and of the possibility of withdrawing from the consent at any moment until the data is entered in the database and learner corpus.

Author's signature ..... Date .....  
Place: .....

## QUESTIONNAIRE

- |                                                         |                           |
|---------------------------------------------------------|---------------------------|
| 1. Sex:                                                 | 9. Level of Czech (CEFR): |
| 2. Age:                                                 | 10. Mother tongue:        |
| 3. Year in which you are currently studying:            |                           |
| 4. How many lessons of Czech do you have per week?      |                           |
| 5. Total duration of studying Czech (in months):        |                           |
| 6. Other languages studied and duration of their study: |                           |
| 7. Do you have Czech relatives? Whom?                   |                           |
| 8. Have you lived in the Czech Republic? How long?      |                           |

1. Napište kamarádovi dopis z návštěvy Prahy (150 slov). V dopise zmiňte:

- a) Co v Praze děláte a jak dlouho tady budete?
- b) Co jste zajímavého navštívil/-a?
- c) Jaká česká jídla jste ochutnal/-a?

2. Popište nejdivnější místo, které jste ve svém životě viděl (150 slov). V textu zmiňte:

- a) Kde a kdy to bylo?
- b) Co bylo na tom divného?
- c) Vyděsila Vás ta situace nebo spíše pobavila?

3. Média (tisk, rádio, televize) odehrávají v současné době podstatnou roli. Zprostředkovávají informace z celého světa. Jsou však vždy důvěryhodná? Napište esej v rozsahu cca 150 slov. V textu:

- a) Vyjádřete svůj kladný nebo záporný vztah k uvedené otázce.
- b) Která média sami používáte a proč?
- c) Mohou být současná média ohrožením pro lidskost?

4. Mysleli jste někdy na budoucnost? Dokážete si představit život v roce 3000? Napište esej v rozsahu cca 150 slov, kde napíšete:

- a) Jaká bude politická situace na světě,
- b) Jak se změní příroda,
- c) Jestli bude možný let na Mars.

# CzeSL-SGT – anonymizace

---

- anonymizace citlivých údajů  
→ pseudonymizace
- nahrazování křestních jmen:  
→ Adam, → Eva
- co jména genderově neutrální? (Saša, Míša, Nikola atd.)
- PoLKo – systematizovaný návrh pseudonymizace

- A. Męski imię zakończone spółgłoską: Jan  
Męski imię zakończone -a: Kuba  
Męski imię zakończone -o: Mieszko
- B. Męski nazwisko odprzymiotnikowe: Kowalski  
Męski nazwisko zakończone spółgłoską: Nowak  
Męski nazwisko zakończone samogłoską: Narta
- C. Żeńskie imię zakończone -a: Anna  
Żeńskie imię zakończone spółgłoską: Nikol  
Żeńskie imię z odmianą miękką: ?
- D. Żeńskie nazwisko odprzymiotnikowe: Kowalska  
Żeńskie nazwisko zakończone spółgłoską: Nowak  
Żeńskie nazwisko zakończone samogłoską: Narta
- E. Nazwy ulic: Słoneczna
- F. Nazwy miast r.n.: Pogodno  
Nazwy miast r.m.: Deszczyk  
Nazwy miast r.ż.: Tęcza  
Nazwy miast l.mn.: Grady

# TEITOK

---

- Nástroj na prohlížení, tvorbu a editaci korpusů
- **TEI** + **TOK** =  
**TEI** – textová anotace podle *Text Encoding Initiative*  
**TOK** – tokenizace textu pro lingvistickou anotaci a hledání v korpusu
- Vše přes webový prohlížeč
- Korpusy: referenční, diachronní, mluvené, akviziční, paralelní, slovníky
- Autor: Maarten Jansen, ÚFAL MFF UK
- <http://www.teitok.org> (Janssen 2016)
- <http://lindat.mff.cuni.cz/services/teitok/>
- <http://utkl.ff.cuni.cz/teitok/>
- <https://ufal.mff.cuni.cz/~sevcikova/2019/docs/MaartenJanssen.pdf>

- TEITOK spravuje všechny podoby textu:
  - zdroj (nahrávku, faksimile)
  - přepis (anotovaný text, včetně stand-off anotace) – TEI XML
  - metadata – v záhlaví dokumentu s anotovaným textem
  - „vertikálu“ – tabulka co token to řádek, token a jeho atributy ve sloupcích
  - prohlédávatelný korpus – databáze Corpus Workbench / Manatee
- Z jednoho webového rozhraní lze:
  - všechny tyto podoby prohlédávat, prohlížet, vytvářet a editovat
  - skoro vše nastavit, parametrizovat, přizpůsobit (anotaci, zobrazení)
  - spouštět nástroje na zpracování textu (typicky anotaci)
- Hotový korpus lze prohlédávat v TEITOKu i v KonTextu

- Vždy zachován vztah: zdroj – anotovaný text – konkordance
  - z konkordance lze přejít k anotovanému textu a faksimile
- Korpus může být „živý“:
  - lze ho rozšiřovat o nové texty
  - texty lze editovat, doplňovat o další anotaci
  - opravy v textech se promítají do korpusu

## TEITOK pro informatika

- PHP skripty operují na daném XML souboru nebo CQP korpusu
- Vizualizace XML pomocí CSS
- Místo Xquery kompilace do CWB, odkazy do XML na znakové pozice
- Výsledky hledání v CWB přes odkazy ve fragmentech XML

## Typický postup:

1. Načtení existujících textů, skenů nebo nahrávek
2. Ruční nebo automatický přepis
3. Ruční opravy a anotace, doplnění metadat
4. Automatická anotace pomocí nástrojů dostupných on-line:
  - lingvistická (hranice slov a vět, morfologie, syntax)
  - chybová (cílová hypotéza, klasifikace chyb)
  - ...
5. Vytvoření korpusu
6. Hledání v korpusu podle nejrůznějších kritérií
7. Výsledky hledání i v původním zdroji (skenu, nahrávce)

## TEITOK

Login

Available Corpora

TEITOK@LINDAT

Latest Additions

Powered by <TEI:TOK>  
Maarten Janssen, 2014-

## Corpora

Below is the list of corpora in the TEITOK/Kontext hybrid set-up, hosted at [ÚFAL](#). To get a larger list of TEITOK projects, see the [TEITOK project page](#). A larger list of Kontext corpora at the UFAL institute can be found in the [KonText corpus list](#), or in the [repository](#). For corpora that have multiple versions in TEITOK, only the most recent version is displayed, but you can click on the version number to see all versions of the corpus. The corpora are listed by corpus type, a description of which can be found [here](#).

|                      | Acronym                                | Latest                     | Token size | Corpus Type        | Corpus Status | Corpus Content         | Corpus Language(s)                      |
|----------------------|----------------------------------------|----------------------------|------------|--------------------|---------------|------------------------|-----------------------------------------|
| <a href="#">info</a> | <a href="#">CzechVerse</a>             |                            | 13M        | Specialized Corpus | live          | Poetry                 | Czech                                   |
| <a href="#">info</a> | <a href="#">DeltaCorpus</a>            | <a href="#">1.1</a>        | 94M        | LRL Corpus         | stable        |                        | Many                                    |
| <a href="#">info</a> | <a href="#">HaCzech</a>                |                            | 18k        | Facsimile Corpus   | stable        | Handwritten texts      | Czech                                   |
| <a href="#">info</a> | <a href="#">Makoň</a>                  | <a href="#">2020-11-16</a> | 4.2M       | Spoken Corpus      | stable        | Transcribed talks      | Czech                                   |
| <a href="#">info</a> | <a href="#">Mazon</a>                  |                            | 9k         | Facsimile Corpus   | live          | Letters                | Czech, German, English, French, Russian |
| <a href="#">info</a> | <a href="#">MuNeCo</a>                 |                            | 840M       | LRL Corpus         | live          | Newspaper articles     | Many                                    |
| <a href="#">info</a> | <a href="#">OCRCZ</a>                  |                            | 27M        | Facsimile Corpus   | stable        | Printed material       | Czech                                   |
| <a href="#">info</a> | <a href="#">PDT-C</a>                  | <a href="#">1.0</a>        | 4M         | Treebank           | stable        |                        | Czech                                   |
| <a href="#">info</a> | <a href="#">ParCzech</a>               | <a href="#">3.0</a>        | 27M        | Spoken Corpus      | stable        | Parliamentary sessions | Czech                                   |
| <a href="#">info</a> | <a href="#">SIR</a>                    | <a href="#">1.0</a>        | 250k       | Specialized Corpus | stable        |                        | Czech                                   |
| <a href="#">info</a> | <a href="#">Skript 2015</a>            |                            | 400k       | Learner Corpus     | live          |                        | Czech                                   |
| <a href="#">info</a> | <a href="#">Universal Dependencies</a> | <a href="#">2.7</a>        | 26M        | Treebank           | stable        |                        | Many                                    |

# Corpus Search

CQL [word="růže"]  
Query:

Search query builder | visualize | options

5141 results • ipm: 406.83 • Showing 0 - 100 • next

Tags: Morphosyntax | Lemma | Phoebe | X-Sampa

context

není | a když zří kol **růže** kvěsti | schází mu to potěšení

context

jedny ručky bílé | jež ty **růže** trhávaly | a ty oči jež

context

kraje | již kvetly akáty a **růže** všude | na stromech zrály fíky

context

květ již plně zraje | A **růže** skláněly se u terassy | pár

context

bravurou když dívčí rozpustilou | ty **růže** trhala svou ručkou bílou | | V

context

zoufalá mi v cestu černé **růže** hází | jak samet černé růže

context

růže hází | jak samet černé **růže** bolesti a ztrát | v té

context

kdy pětistá vzácné krásy | tu **růže** plála do kola | | Ty staré

context

Hvězdy a bludičky

e mi | vždyť i má **růže** mládí moje | tu někde dřímá

context

Author Brabec, Adolf

m na dně rostou **růže** divoké | a bují trávník V

context

Date 1910

domem málo zlatovský | se dařily **růže** květy lásky | Zahrádka malá kousek

context

mne ve hruď žíhadv | Květ **růže** vidím dávno uvadlý | Tak nonrvé

## Když růže rozkvetou.

Stokrát zase vzpomenu si  
na to jaro naší lásky  
a tu člověk smutně musí  
hledět nazpět bez otázky

Nazpět za vším co již není  
a když zří kol **růže** kvěsti

schází mu to

uzříť jedno dítě

uzříť jedny ruč

jež ty růže trh

a ty oči jež se

všemu co ten

o čem blouzní mladá hlava

Člověk náhle vzpomene si

| růže         |                                                                               |
|--------------|-------------------------------------------------------------------------------|
|              | Noun (NNFS2----A----)                                                         |
| Morphosyntax | Noun (general); Feminine;<br>Singular; Genitive; Affirmative<br>(not negated) |
| Lemma        | růže                                                                          |
| Phoebe       | rUZe                                                                          |
| X-Sampa      | ru:ZE                                                                         |

## TEITOK

Login  
Available Corpora

## Skript 2015

Domů  
Dokumenty  
Hledat  
KonText  
Login

Powered by <TEI: TOK>  
Maarten Janssen, 2014-



EVROPSKÁ UNIE  
Evropské strukturální a investiční fondy  
Operační program Výzkum, vývoj a vzdělávání



## Skript 2015 – korpus rodilých mluvčích češtiny

- Modelový typ srovnávacího akvizičního korpusu písemných prací žáků na různých jazykových úrovních
- Současný stav:
  - 2582 textů, 380 tis. tokenů; žáci ZŠ, ZŠP, ZŠS, SOŠ a gymnázií (včetně víceletých) ve věku 10–15 let
  - Texty jsou vybaveny metadaty a faksimilií
  - U textů byla provedena ruční transkripce včetně záznamu oprav pisatele a ruční anonymizace
  - Texty byly ručně i automaticky anotovány a ručně revidovány, výsledky obsahují:
    - ruční emendaci na více rovinách: pravopis a morfematika, morfosyntax, slovník
    - automatickou morfologickou analýzu původního textu a všech oprav
    - automatickou identifikaci typu pravopisné a morfématické chyby
  - Po přihlášení uživatel může texty i anotace opravovat a doplňovat
- Využitelnost:
  - Učitelé mohou korpus využívat při přípravě hodiny, pomocných učebních materiálů i při vlastní výuce
  - Žáci mohou korpus využívat při samostatném studiu
- Aktuální využití a udržitelnost:
  - V kurzech DV PP i v kurzech v pregraduální přípravě učitelů
  - Při vývoji počítačových nástrojů ke zpracování textů nerodilých mluvčích češtiny nebo jiných textů odlišných od spisovné normy

## Instrukce k přípravě korpusu

### Přepis a anonymizace žákovských textů

- [Návod](#)
- [Transkripční značky](#)
- [Starý způsob transkripce, doplněk manuálu, transkripce a anonymizace romských textů](#)

### Oprava žákovských textů

- [Návod k anotaci v prostředí TEITOK](#)
- Viz též forum [Emendace žákovských textů v TEITOKu](#)

# To jsem ještě nikdy neudělal

Název To jsem ještě nikdy neudělal

## Možnosti zobrazení

Text: Přepis | **Psaná podoba** | Pravopisná oprava | Morfosyntaktická oprava | **Lexikální oprava** - Ukázat: Barvy - Značky: Slovní druh | Lemma | Slovní druh pravopisné opravy | Lemma pravopisné opravy | Slovní druh morfosyntaktické opravy | Slovní druh lexikální opravy | Lemma lexikální opravy | Chybová značka

Co jsem ještě nikdy neudělal

No ještě jsem nikdy nikoho nezabila. A neukradla. Nikdy jsem nebyla v kině protože nikdy na to nemám peníze. A u zubaře jsem taky nebyla protože se bojím a nechci být zubatá tak radši nejím sladkosti. A je vdycki odiče nemají. Na dovolené jsem taky nikdy nebyla ale chtěla bych to je asi všechno.

[Stáhnout XML](#) • [Stáhnout text](#) • [Zobrazit po větách](#) • [Statistiky](#) • [Vytvořit stand-off anotaci](#)

| vdycki                        |                                                 |
|-------------------------------|-------------------------------------------------|
| Pravopisná oprava             | vždycky                                         |
| Lexikální oprava              | nikdy                                           |
| Slovní druh                   | Nezjištěno (X@-----)                            |
| Lemma                         | vdycki                                          |
| Slovní druh pravopisné opravy | Příslovce (Db-----A--)<br>bez záporu/stupňování |
| Lemma pravopisné opravy       | vždycky                                         |
| Slovní druh lexikální opravy  | Db-----N---                                     |
| Lemma lexikální opravy        | nikdy                                           |

- [Home](#)
- [Documents](#)
- [Search](#)

user: [Alexandr](#)

- [Admin](#)
- [Help](#)
- [XML Files](#)

Powered by <TEI:TOK>  
Maarten Janssen, 2014-

## CzeSL – Czech as a Second Language

- Work in progress: new texts, transcription, annotation, imports of data in other formats
- This corpus is meant to include all available Czech texts written by non-native learners.
- The currently searchable texts are annotated automatically in the cze-sl style: forms identified as incorrect are normalized and assigned one or more error tags, the original and normalized forms are tagged and lemmatized.
- The corpus comes with metadata and scans of manuscripts (where available) for registered users.
- Manual annotation already available for some texts will be added, the automatic normalization for other texts will be checked and extended manually.

### Přepis žákovských textů

- [Návod](#)
- [Transkripční značky](#)

### Oprava žákovských textů

- [Jak si vytvořit korpus](#) – stránky Moodle pro účastníky stejnojmenného semináře
- [Práce se systémem TEITOK](#) – forum pro účastníky semináře, používejte prosím přednostně
- [Fórum pro anotaci v systému TEITOK](#) na stránkách Ústavu teoretické a počítačové lingvistiky – pro anotátory, kteří nejsou účastníky semináře
- [Emendace žákovských textů v TEITOKu](#) – forum na Google Groups, vyžaduje uživatelský účet na Googlu, používaný pro diskuse o anotaci textů rodilých mluvčích
- [Návod k anotaci v prostředí TEITOK](#) – původně určen pro opravy textů rodilých mluvčích v projektu Skript 2015; v projektu CzeSL opravujeme hodně podobně, proto vycházíme ze stejného návodu a na případné rozdíly zvlášť upozorňujeme

More information about how to set-up TEITOK can be found in the help pages: <http://teitok.corpuswiki.org/site/index.php?action=help>

# CzeSL - multi- modální korpus

EN | CS

Hlavní nabídka

- Domů
- Dokumenty
- Hledat
- Login

Powered by <TEI:TOK>  
Maarten Janssen, 2014-

## CzeSL – multimodální korpus projevů nerodilých mluvčích

- Modelový typ žákovského korpusu zachycujícího projevy týchž studentů češtiny jako cizího jazyka v modu psaném i mluveném
- Současný stav:
  - Ukázka možností systému TEITOK a automatického zpracování žákovských projevů psaných i mluvených (zatím s minimálním objemem dat)
  - Texty jsou vybaveny metadaty, která jsou přístupná pro vyhledávání i při prohlížení textů; jsou zkorigovány, morfologicky označovány, lemmatizovány a syntakticky analyzovány
  - V obou typech textů (psaných a mluvených) lze hledat najednou
  - Mluvený text je možné si přehrát, u psaného textu se faksimile rukopisu zobrazuje po přihlášení uživatele
  - Po přihlášení lze přepis zvukové nahrávky i rukopisu i jejich anotaci a metadata opravovat a doplňovat
- Využitelnost:
  - Příprava učitelů a žáků na práci (nejen) s multimodálními korpusy založenými na systému Teitok
  - Po rozšíření objemu dat mohou učitelé korpus využívat při přípravě hodiny, pomocných učebních materiálů i při vlastní výuce, žáci při samostatném studiu
- Aktuální využití a udržitelnost:
  - V kurzech DV PP i v kurzech v pregraduální přípravě učitelů
  - Plánuje se postupné rozšiřování dat ve spolupráci se školami

## Povídání o studiu

Název Povídání o studiu

- [edit header data](#) • [view teiHeader](#)

### Možnosti zobrazení

Text: [Přepis](#) [Psaná podoba](#) - Ukázat: [Barvy](#) [Audio](#) - Značky: [Ortografické kategorie](#) [Ortografické lemma](#)



Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML

S1 - *pousmání*

S1 - [...] tak aby [...]

T - mně se líbí jak studujete spoustu jazyků

T - *dlouhá pauza*

T - já jsem vystudoval taky latinu

S1 - *Hmm*

S2 - *Hmm* a co z ní znáte

T - *pousmání*

T - *Uh*

T - *dlouhá pauza*

T - co z ní *Uh* co z ní znám no

Možnosti zobrazení

Text: Přepis

Psaná podoba

Ortograficky opravený tvar

Morfosyntakticky opravený tvar

Lexikálně opravený tvar

- Ukázat: Barvy

<pb>

Skeny

- Značky:

Ortografické kategorie

Ortografické lemma

Lexikální kategorie

Lexikální lemma

Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML

Cestování

Cestu **plánujeme** ne moc dlouho. **Vždycky máme 2-3 dny**, **abychom vybrali**, kam **pojedeme**. Stát a město, ve kterém budeme **odpočívát**, **vyhledám** obyčejně já: čas toho místa vybírám tak ve kterém budeme byd chceme jít v **tom** městě se **podívat**. Kde to je, **ja**.  
A když se dostaneme **d** co se budeme **chtít pod**

|                                |                                                                                                          |
|--------------------------------|----------------------------------------------------------------------------------------------------------|
| budeme jet                     |                                                                                                          |
| Morfosyntakticky opravený tvar | pojedeme                                                                                                 |
| Ortografické kategorie         | Podstatné jméno (NNXXX-----A---8) rod libovolný; číslo libovolné; pád libovolný; nezáporný tvar; Zkratka |
| Ortografické lemma             | budeme jet                                                                                               |
| Lexikální kategorie            | Sloveso (VB-P---1F-AA---) tvar určitý; číslo množné; osoba 1.; čas budoucí; nezáporný tvar; rod činný    |
| Lexikální lemma                | jet                                                                                                      |

Cestování

Čerilova Kristina

Cestu planujeme ne moc dlouho.

Vždycky máme 2-3 dny, abychom vybrali, kam budeme jet. Stát a město, ve kterém budeme odpočívát vybíráji dospívá matka a otec. Hledám lístky obyčejně já : čas, čím budeme jet a jak dojet do toho místa vybírám taky já. Moje matka hledá nějaký hotel, ve kterém budeme bydět. Maminka vždycky ptá, kam chceme jít v tom městě a každý říká, kam bychom měli jít a na co se podívat, kde to je, jak do toho místa se dostat, hledám já.

A když se dostaneme do tohoto místa, ještě něco najdeme, na co se budeme chtít podívat. Moje rodina má ráda cestování.

Pozor na tvary budoucího času

# Cestování

Název Cestování

- [edit header data](#) • [view teiHeader](#)

## Možnosti zobrazení

Text: [Přepis](#) | [Psaná podoba](#) | [Ortograficky opravený tvar](#) | [Morfosyntakticky opravený tvar](#) | [Lexikálně opravený tvar](#) - Ukázat: [Barvy](#) | [<pb>](#) | [Skeny](#) - Značky:  
[Ortografické kategorie](#) | [Ortografické lemma](#) | [Lexikální kategorie](#) | [Lexikální lemma](#)

Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML

## Cestování

Cestu **plánujeme** ne moc dlouho. **Vždycky máme** 2-3 **dny**, **abychom vybrali**, kam **pojedeme**. Stát a město, ve kterém budeme **o** **vybrát**

**hledám** o

toho místa

ve kterém

chceme jít

se **podívat**

A když se

co se bude

| vybrát                         |                                                                                                                                    |
|--------------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| Ortograficky opravený tvar     | vybrat                                                                                                                             |
| Morfosyntakticky opravený tvar | vybrali                                                                                                                            |
| Ortografické kategorie         | Sloveso (Vf-----AA--<br>-)<br>infinitiv; nezáporný tvar; rod činný                                                                 |
| Ortografické lemma             | vybrat                                                                                                                             |
| Lexikální kategorie            | Sloveso (VpMP---XR-AA---)<br>přičestí -l; rod mužský životný; číslo množné; osoba libovolná; čas minulý; nezáporný tvar; rod činný |
| Lexikální lemma                | vybrat                                                                                                                             |

ka a otec). **Lístky**

jak **nejlépe** dojet do

hledá **nějaký** hotel,

**vždycky** ptá, kam

**by** chtěl jít a na co

dostat, **hledám** já.

**něco** najdeme, na

**ná** ráda cestování.

Cestování

Čerilova Kristina

Cestu plánujeme ne moc dlouho.

Vždycky máme 2-3 dny, abychom vybrali, kam budeme jet. Stát a město, ve kterém budeme odpovírat vyprávějí dospělí (matka a otec). Hledám lístky obvykle já: čas, čím budeme jet a jak se dojet do toho místa vyprávějí taty já. Moje matka hledá nějaký hotel, ve kterém budeme bytět.

Maminka vždycky ptá, kam chceme jít v tom městě a každý říká, kam bychom mohli jít a na co se podívat, kde to je, jak do toho místa se dostat, hledám já.

A když se dostaneme do tohoto místa, ještě něco najdeme, na co se budeme chtít podívat. Moje rodina má ráda cestování.

Pozor na tvary budoucího času

# Edit Token

Filename Searchable/TSN\_11\_D13-002.xml

Title Cestování

## Token value (w-13): vybrát

|       |                                    |                                      |
|-------|------------------------------------|--------------------------------------|
| pform | Transcription (Inner XML)          | <input type="text" value="vybrát"/>  |
| ort   | Orthographically corrected form    | <input type="text" value="vybrat"/>  |
| gram  | Morphosyntactically corrected form | <input type="text" value="vybrali"/> |
| lex   | Lexically corrected form           | <input type="text"/>                 |

|        |              |                                              |                    |
|--------|--------------|----------------------------------------------|--------------------|
| opos   | Ort POS tag  | <input type="text" value="Vf-----AA---"/>    | asistent na značky |
| olemma | Ort Lemma    | <input type="text" value="vybrat"/>          |                    |
| gpos   | Gram POS tag | <input type="text"/>                         | asistent na značky |
| lpos   | Lex POS tag  | <input type="text" value="VpMP---XR-AA---"/> | asistent na značky |
| llemma | Lex Lemma    | <input type="text" value="vybrat"/>          |                    |

insert tok after: **attached** / **separate** • before: **attached** / **separate** • insert elm: **note** ; **linebreak** • split in dtoks: **2** ; **3**

**edit** context XML • **merge** left to w-12 • create mtok left: **1** ; **2**

treat similar tokens

## Undefined attributes

deprel conj

head w-8

Vždycký mame 2-3 dnů, abych vybrát kam budeme jet.

Save

Cancel • Token Details

```

<s id="s-2">
  <tok id="w-1"
    opos="NNFS4-----A----" olemma="cesta"
    lpos="NNFS4-----A----" llemma="cesta"
    deprel="obj" head="w-2">
    Cestu
  </tok>
  <tok id="w-2"
    ort="plánujeme"
    opos="VB-P---1P-AA----" olemma="plánovat"
    lpos="VB-P---1P-AA----" llemma="plánovat"
    deprel="root">
    plánujeme
  </tok>
  <tok id="w-3"
    opos="TT-----" olemma="ne"
    lpos="TT-----" llemma="ne"
    deprel="advmod:emph" head="w-5">
    ne
  </tok>
  <tok id="w-4"
    opos="Db-----" olemma="moc"
    lpos="Db-----" llemma="moc"
    deprel="advmod" head="w-5">
    moc
  </tok>
  <tok id="w-5"
    opos="Dg-----1A----" olemma="dlouho"
    lpos="Dg-----1A----" llemma="dlouho"
    deprel="advmod" head="w-2">
    dlouho
  </tok>
  <tok id="w-6"
    opos="Z:-----" olemma="."
    lpos="Z:-----" llemma="."
    deprel="punct" head="w-2">
    .
  </tok>
</s>

```

# Dependency Tree

## Cestování

Název Cestování

- [edit header data](#) • [view teiHeader](#)

s-1 <

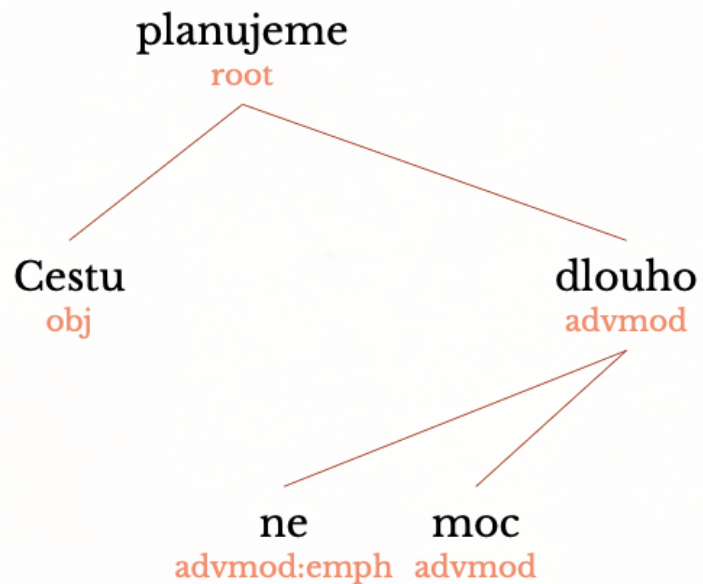
věta s-2

> s-3

Click [here](#) to edit the dependency tree

Cestu planujeme ne moc dlouho.

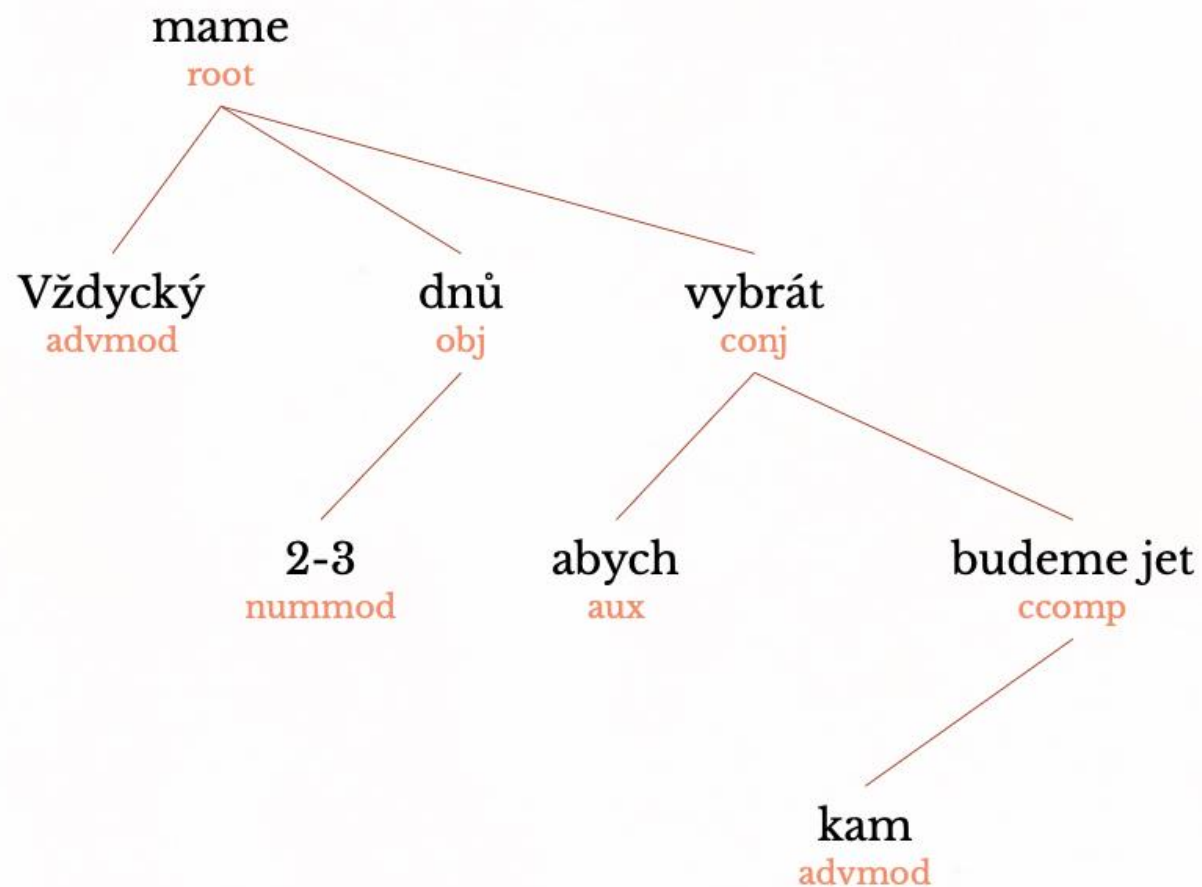
[edit metadata](#)



Click [here](#) to edit the dependency tree

Vždycký mame 2-3 dnů, abych vybrát kam budeme jet.

[edit metadata](#)



# Děkujeme za pozornost!

---

Alexandr Rosen

[alexandr.rosen@ff.cuni.cz](mailto:alexandr.rosen@ff.cuni.cz)

Adrian Jan Zasina

[adrian.zasina@ff.cuni.cz](mailto:adrian.zasina@ff.cuni.cz)

# Otázky k diskusi

---

1. Jaké jsou možnosti pokrytí dalších L1?
2. Jakým způsobem ošetřit jazykovou úroveň?
3. Jak pseudonymizovat citlivé údaje?

# Bibliografie

---

- Baker, P., Hardie, A., & McEnery, T. (2006). *Glossary of corpus linguistics*. Edinburgh University Press.
- Janssen, M. (2016). TEITOK: Text-faithful annotated corpora. In Calzolari, N. et al (ed.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, str. 4037–4043, Paris, European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2016/pdf/651\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf)
- König, A., Frey, J.-C., Stemle, E. W., Glaznieks, A., & Paquot, M. (2022, září). *Towards standardizing LCR metadata*. <http://www.maldura.unipd.it/lcr-2022/>
- 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, Praha. <https://dspace.cuni.cz/handle/20.500.11956/123103>
- Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierścieniak, P., Toufarová, D., Richter, M., Straka, M., & Rosen, A. (2014). *CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, verze 2 z 28. 7. 2014*. Ústav Českého národního korpusu FF UK. [www.korpus.cz](http://www.korpus.cz)
- Škodová, S. (2012). Nástin využití žákovských korpusů pro jazykové vyučování. In K. Šebesta & S. Škodová (Eds.). *Čeština - cílový jazyk a korpus* (pp. 125-138). Liberec.
- Šormová, K., Šebesta, K., Gráf, T., Hejhalová, V., & Kukrechtová, B. (2019). *Korpusy v jazykovém vyučování*. Univerzita Karlova, Filozofická fakulta.