Projekt Akviziční korpusy češtiny – přepis ručně psaných textů

- Dosud se texty přepisovaly v běžném textovém editoru s využitím ručně zadávaných kódů. V takto přepsaných textech se často vyskytují formální chyby, které komplikují další zpracování textu. Proto jsme se rozhodli způsob přepisu změnit.
- Pro přepis textů používáme nově prostředí pro vytváření a prohlížení korpusů *TEITOK*, které je přístupné přes internetový prohlížeč. V tomto prostředí se formální správnost každého přepisovaného dokumentu automaticky kontroluje a případné chyby lze snadno odstranit.
- Pro snadné a spolehlivé zadávání přepisovacích kódů je vhodné využít program *AutoHotKey* s předdefinovanými klávesovými zkratkami. Program je určen pro operační systém *MS Windows*. Řešení pro případné jiné systémy je také možné.

Postup

1. V emailu na adresu alexandr.rosen@ff.cuni.cz si napište o **přístupové údaje** do stránek *TEITOK*.
2. Ze stránek https://autohotkey.com si **stáhněte a nainstalujte program *AutoHotKey***. Můžete zvolit `Express Installation`.
3. Ve webovém adresáři AHK vyberte pravým tlačítkem myši **skripty programu *AutoHotKey*** nazvané `czesl.ahk` a `czesl_offline.ahk`. Uložte je jako skripty programu *AutoHotKey* na vhodné místo (třeba na plochu). Skript `czesl.ahk` je určen pro přepis textů on-line přímo ve webovém rozhraní *TEITOK*. Skript `czesl_offline.ahk` je určen pro přepis textů s využitím lokálně nainstalovaného textového editoru.[1] Skript by měl dostat ikonku s písmenem H. Skript aktivujte dvojím kliknutím na tuto ikonku.
4. V internetovém prohlížeči otevřete http://utkl.ff.cuni.cz/teitok/czesl/ a **přihlašte se do systému *TEITOK***.[2]
5. V hlavním panelu zvolte create new XML file.
6. Do rámečku pod **XML Filename** uveďte **název souboru**. Název musí odpovídat identifikátoru rukopisu a musí být zakončen koncovkou `.xml`.
7. Do textového okna **zapište libovolný znak**.
8. Klikněte na tlačítko **`Create XML File`** pod rámečkem a červenou čárou.
9. Na nové stránce zvolte první možnost: klikněte na poslední slovo (here) v prvním odstavci: This XML has not been tokenized yet, and only the text is shown below.

[1] Doporučujeme volně dostupný editor Notepad++ (https://notepad-plus-plus.org).
[2] Dříve http://teitok.iltec.pt/czesl/. Všechna data i přístupová oprávnění by měla být z portugalského serveru přenesena. Pokud zjistíte nějaký problém, prosím o zprávu na alexandr.rosen@ff.cuni.cz.

To edit, click here. Neklikejte na here ve druhém odstavci, editovaný text se pak už začne zpracovávat pro zařazení do korpusu a nebude přístupný pro další editaci.

10. V textovém rámečku je připravena výchozí šablona textu ve formátu XML. Textové elementy jsou vyznačeny pomocí **počáteční a koncové značky**, např. `<text>` a `</text>` pro celý text, nebo `<p>` a `</p>` pro odstavec. Počáteční značka může mít jeden nebo více atributů (viz značka `<text>`).[3] Přepisovací značky obvykle zleva a zprava uzavírají označený textový řetězec (znaky nebo slova). Např. značka **add** pro označení vsuvky nebo výsledku opravy (zde tvar *jsem*) se použije takto: `<add>`jsem`</add>`.
11. Text je třeba **průběžně ukládat**, abychom o něj při případném výpadku síťového připojení nepřišli. Klikněte na tlačítko `Save` pod rámečkem a uvidíte obsah textu bez značek.[4] Zpátky se vraťte opět kliknutím na poslední slovo v prvním odstavci: This XML has not been tokenized yet, and only the text is shown below. To edit, click here. Všimněte si, že značka pro odstavec má nyní atribut `id`.
12. Místo původně zadaného libovolného znaku teď můžeme začít přepisovat text z rukopisu. Je třeba značkovat **odstavce i věty**.
13. Pokud se v textu vyskytne **chyba ve formátu XML**, *TEITOK* na ni upozorní a lze ji hned opravit. Text s chybou nelze uložit.
14. Jiné chyby lze odhalit **kontrolou textu po jeho uložení** ve formátu bez značek.
15. Na stránkách *TEITOK*u mezi XML Files najdete soubor `CODES.xml`, který obsahuje všechny **přepisovací značky**, včetně anonymizovaných. Lze si ho prohlédnout ve formátu XML i bez značek.
16. Všechny značky lze do textu vložit pomocí **klávesové zkratky**. Nejprve je třeba aktivovat skript `czesl.ahk`.
17. Přepisovací značky jsou uvedeny zvlášť v jiném dokumentu. **Přehled** obsahuje klávesovou zkratku, kód XML a příklad vizualizace kódu v textovém režimu *TEITOK*. Příslušné písmeno z klávesové zkratky je uvedeno také každou značkou v souboru `CODES.xml`.
18. Kromě anonymizačních značek se značky vkládají **současným stisknutím kláves** `Windows`, `Alt` a příslušného písmene (v kulaté závorce). U anonymizačních značek je třeba současně s `Windows`, `Alt` a příslušným písmenem (v hranaté závorce) stisknout ještě klávesu `Ctrl`.

Poznámky

1. Texty lze přepisovat i bez připojení k internetu ve vhodném editoru a pak je zkopírovat do TEITOKu. Některé klávesové zkratky však nebudou fungovat správně. V případě zájmu o verzi skriptu *AutoHotKey* pro běžné editory lze takovou verzi vytvořit.

---

[3] Některé značky jsou prázdné, neoznačují žádný text, např. značka `<gap/>` pro nečitelný řetězec.

[4] Význam některých značek však může být vyjádřen typem nebo barvou písma.

2. Skript *AutoHotKey* je otestován v prohlížeči *Chrome*. Pokud některé klávesové zkratky v jiném prohlížeči nefungují, zkuste změnit prohlížeč. V nutném případě lze pro konkrétního uživatele předefinovat skript.
3. Připomínky všeho druhu se s díky přijímají na adrese alexandr.rosen@ff.cuni.cz.

Práce se soubory

- **Stahování skenů** probíhá přes databázi *AMES*: po přihlášení do databáze – kliknout na `Nahrávky/Texty` (nahoře) – objeví se tabulka – vlevo kliknout na `texty - přehled souborů`, dále kliknout v tabulce na `přepisované`; otevře se další tabulka, z které už je možné sken stáhnout (vpravo) – kliknout na `stáhnout sken` (otevře se sken) – na liště nahoře rozkliknout `stránka` – `dát uložit jako` – poté se sken uloží do vašeho počítače, některé skeny jsou uloženy jako komprimovaná složka, kterou si lze klasicky stáhnout do PC.
- Každý text k přepisu je označen jedinečným **identifikátorem** (v tomto formátu: `Vob_AR_005`). Skeny mohou být uložené pod názvem souboru např. `UJA_AB_001_t_1.jpg`, přepis ale musí být pojmenován jen `UJA_AB_001.xml` (`t_1` do názvu přepisu už nedávejte). Přepsané soubory tedy budou mít jinou příponu: `.xml`.
- Některé texty jsou na dvou a více skenech, jednotlivé skeny jsou pak pojmenovány a, b, … – je nutné stáhnout si celý text žáka (v komprimované složce) a přepsat ho do jednoho dokumentu, v názvu přepisu už a, b nebude (např. skeny `UJA_SB_001a`, `UJA_SB_001b` – přepis `UJA_SB_001`).
- U identifikátoru se mohou vyskytnout **škrty** – toto není projev studenta. Identifikátory se nepřepisují, ani když jsou přeškrtnuté – je důležité rozeznat to od vlastního textu.
- Přepis po sobě **zkontrolujte** a uložte ho do svého adresáře s přepsanými texty na serveru *TEITOK*.

Upravené zásady anotace vycházející ze starého manuálu

- Úkolem přepisu je **ZACHOVAT AUTENTICKOU PODOBU** původního textu; přepisovač tedy v žádném případě **NEOPRAVUJE CHYBY** ani **NEUVÁDÍ SPRÁVNOU VERZI**. Dávejte pozor na to, abyste bezděčně nevkládali správnou variantu např. u diakritiky, interpunkce ap. Tzn. nepište *ve škole* místo původního znění *v škole; máma* místo *mama* atd.).
- Při přepisu je třeba zachovat **pisatelovy vlastní opravy (rektifikace): vsuvky, škrtání, přepisy, spojení a rozdělení slov**. Viz přehled přepisovacích značek a příklady.
- **Značky nelze kombinovat**, tj. řetězec označený jednou značkou nelze označit současně značkou další. Např. nelze zaznamenat škrt ve vsuvce. V takových případech je nutné zvolit jen jednu z více značek, např. značku pro vsuvku a místo dalších značek vycházet z intence autorských rektifikací.

- Veškeré **učitelovy škrty, vpisky a opravy** v textu ignorujte, nijak je při přepisu nezachycujte. Pouze v případě, že původní žákova slova jsou kvůli učitelovu zásahu nečitelná, označte příslušnou pasáž odpovídajícím způsobem (viz přehled značek).
- Někdy jde velice špatně rozeznat, co je oprava studenta a co učitele; pokud narazíte na tento problém a nebudete si jistí, je lepší tam opravu zaznamenat (kontrola to pak případně vymaže).
- Studenti často píší text nejprve **nanečisto tužkou** a teprve poté přes to propiskou. Následně text tužkou vygumují, ale stává se, že tam zbytky předlohy zůstanou. Je nutné to rozlišit a takovéto zbytky textu nepřepisovat.

- **Nadpis**, je-li v textu přítomen, se vyznačuje značkou **`title`.**
- **Předtištěný text**, na který žák navazuje, se přepisuje pomocí značky **`quote`**. Např. `<quote>Ahoj Jakube</quote>`. Předtištěným textem není zadání, slova či informace, které má žák v textu použít, atd. Předtištěný text je pouze to, na co žák bezprostředně navazuje, co je součástí jeho textu. Uvnitř předtištěného textu nepoužíváme značky pro odstavce ani pro věty.
- Začátek a konec **věty** a **odstavce** se vyznačuje pomocí značek XML **`s`** a **`p`**. V ojedinělých případech se může stát, že hranice mezi větami nejsou rozeznatelné v celém textu (pisatel například neužívá interpunkci). Takový text zachováme v jeho původní podobě, tedy bez interpunkce. Značky **`s`** a **`p`** nepoužíváme uvnitř nadpisů a předtištěných textů.
- Jednotlivá slova oddělujte **mezerou**, opravy jednoho nebo více znaků v rámci slova uvádějte bez mezer před přepisovací značkou i za ní.
- V přepsaném textu nezáleží na **řádkování**, a tedy ani na případných volných řádcích. Pro přehlednost je ale vhodné začínat každou větu na novém řádku a mezi odstavci vynechávat volný řádek.

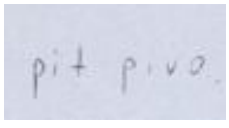
- Dodržujte **psaní velkých a malých písmen přesně podle originálu**, a to včetně všech chyb; pokud je celý text psaný velkými písmeny, přepíše se také tak.
- Při přepisu je nutné brát v potaz celý text a respektovat **způsob psaní autora**. Různé, zejména grafické odchylky (např. od psacího písma latinky) není třeba hodnotit jako chyby nebo varianty, ale projev individuálního způsobu psaní autora textu. Příklady viz níže. Není třeba hledat problémové jevy tam, kde nejsou.
- Pokud se objeví v textu „**i“ a „j“ bez tečky**, bude to přepisovač uvádět na pravou míru a přepisovat s tečkou.
- **Mezera** před interpunkčním znaménkem nebo mezera za ním se nepřepisuje, tj. přepisovač bude tuto chybu sám uvádět na pravou míru. Dvě a více interpunkčních znamének za sebou se přepíší podle rukopisu.
- **Dělení slov** na konci řádku se nezaznamenává.
- Pokud text obsahuje posloupnost **teček** (autor tím dává najevo, že přemýšlí o tématu, či jde o nedokončenou větu; může se vyskytovat i větší počet teček) – zapíše se vždy 3 tečkami.

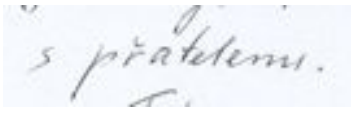
- **Diakritická znaménka** ???
- Pokud je možné přepsat úsek textu více způsoby, je třeba se rozhodnout pro ten nejpravděpodobnější, případně nejbližší očekávané podobě. **Varianty** v novém způsobu přepisu už nejsou přípustné.
- **Položky seznamu** se přepisují značkami **`item`**, ohraničené značkami **`list`.** Odrážky se nijak zvlášť nepřepisují.

- Zcela **nečitelná slova/řetězce** se zapisují jako `<gap/>`. Je-li přepisovač schopen identifikovat počet nečitelných slov, uvede počet slov např. takto: `<gap unit="word" quantity="3"/>`. Jsou-li nečitelná jen písmena, uvádí se řetězec např. 4 nečitelných písmen takto: `<gap unit="char" quantity="4"/>`. Stejně se řeší i slova nebo znaky zapsané v jiném grafickém systému (např. azbukou).
- Značka `<gap/>` se používá pouze pro slova či slovní spojení, která do textu patří (tzn. žák by je přečetl, jsou pro něj srozumitelná, ale přepisovači a kontroloři je nedokážou identifikovat). Tato značka se nepoužívá pro zaškrtaná slova či slovní spojení. Pokud se objeví **zaškrtané slovo**, které lze přečíst, zapíše se pomocí značky `<del rend="strikethrough">...</del>`, přeškrtnutého písma, pokud ho nelze přečíst, přepisuje se jako `<del rend="strikethrough"><gap/></del>`. ??? Začmárané a zcela zaškrtané, tudíž nečitelné řetězce se nepřepisují. ???
- Pokud je **slovo/řetězec nečitelný**, ale přepisovač je schopen na základě kontextu nebo jiných skutečností interpretovat nějakou možnost čtení, může ji zapsat, a to pomocí značky **unclear**, např. Pracuje jako `<unclear>dělník</unclear>`.
- Vyskytne-li se v textu **obrázek** (vlepený, kreslený, emotikon apod.), uveďte na příslušném místě značka **`image`** se stručným popisem obrázku, i víceslovným. Emotikony se přepisují obvyklým způsobem, jen je třeba se vyvarovat znaku. Např. šel tam pes `<image>pes</image>`. `<image>:-)</image>` Emotikony nesmí být ve formě obrázku. Není-li obrázek přímo v textu, umístit kód za nejbližší odstavec nebo na konec textu.
- Pro poznámky přepisovače se používá značka **note**.

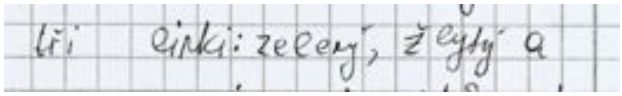
- **Anonymizace** vlastních jmen, místních názvů, adres, telefonních čísel: viz **anon** v přehledu přepisovacích značek ???

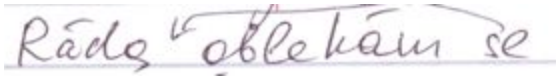
Příklady



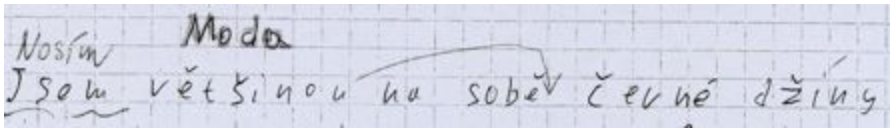
pit pivo 

s přátelemi

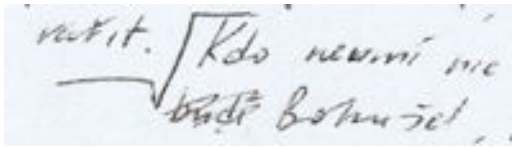
 tři linki: zelený, žlytý a



Ráda <add id="1">se</add> oblékám <del target="1">se</del>

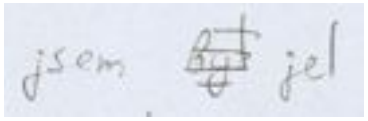


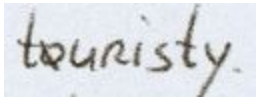
Nosím <del target="1">většinou</del> na sobě <add id="1">většinou</add> černé džíny.

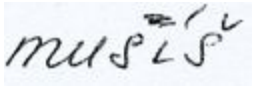


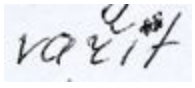
Oprava v odsazení odstavce se nijak neznačí. Bere se v úvahu jako nový odstavec.

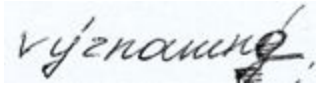
 Pak <add>jsme</add> jeli.

 jsem <del rend="strikethrough">Byl</del> jel

 t<del rend="strikethrough">o</del>uristy

 mu<del rend="strikethrough">š</del><add>s</add>íš

 vař<del rend="strikethrough">í</del><add>i</add>t

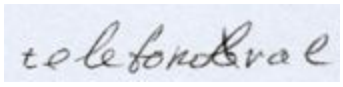
 významn<del rend="overwritten">ý</del><add>é</add>

 ???

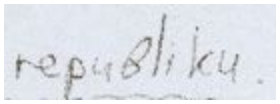
 mi Češi, přírodu

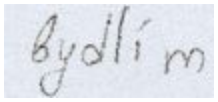
 školy.,

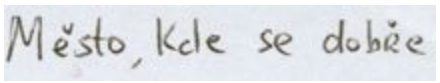
myslím ja (<gap unit="word" quantity="3"/>)

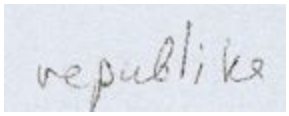
 telefonov<del rend="strikethrough"><gap/></del>val

 se<split/>my: <note>přeškrtnutá čárka nad y</note>

 repu**b**liku

 **b**ydlím

 Město, **k**de se dob**ř**e

 repu**b**lik**a**