Projekt *InterCorp* – postup přípravy textů verze 4

Alexandr Rosen a Martin Vavřín

24. února 2020

Obsah

Seznam tabulek

Seznam obrázků

Poděkování

- Děkujeme všem autorům za materiály, z nichž tento návod vychází. Dalším, zde neuvedeným účastníkům projektu, pak za všechny cenné podněty a připomínky.

    1. Eliška Boková *Tabulka problémů skenování* (`http://utkl.ff.cuni.cz/~rosen/public/tabulka-problemy_skenovani.pdf`)
    2. Jan Kocek *Zásady skenování textů* (část 3)
    `https://intercorp.korpus.cz/dokumenty/zasady_skenovani_InterCorp.pdf`
    3. Pavel Vondřička *Stručný návod k používání databáze textů InterCorp* (část 7)

- Další připomínky a podněty všeho druhu uvítáme na adrese `martin.vavrin@ff.cuni.cz`, nebo (jsou-li důležité i pro ostatní účastníky projektu): `intercorp@ff.cuni.cz`.

1 Postup ve zkratce

1. Výběr textů (2)
2. Získání textů:
   - z ÚČNK (2, bod 3a) $\Longrightarrow$ 5
   - z webu: (2, bod 3b)
     - legálně zadarmo
     - za peníze
   - z nakladatelství (2, bod 3c)
   - skenováním (3)
3. Úpravy před zarovnáním:
   (a) korektury (3.2)
   (b) export z Wordu (4.1)
   (c) segmentace po větách
4. Zarovnávání: 5

Software, který je třeba nainstalovat

- FineReader (jen pro skenování) `http://www.abbyy.com/download/?param=28619`
- Makro ICorpExport (příloha A.2)

2 Výběr a získání textu

1. Než se pustíte do práce s textem, ověřte si, zda je paralelní text skutečně přeložený. Některé překlady mohou být velmi volné. Takové překlady často neumožňují zarovnávat jednotlivé věty.
2. Pokud je to možné, vyberte si takové verze textu, kde je použito podobného (ideálně stejného) členění na odstavce. Výrazně si tím ulehčíte práci při zarovnávání.
3. Podívejte se, zda text není možné získat rovnou v elektronické podobě: ze zdrojů ÚČNK (3a), z internetu (3b) nebo z nakladatelství (3c):
   (a) Projděte databázi textů projektu InterCorp (viz část 7), pokud si nejste jisti, raději se ještě emailem přeptejte hlavního koordinátora.
   (b) Zjistěte si, zda by nebylo možné danou knihu získat v elektronické verzi volně a legálně na internetu nebo např. v podobě e-booku (Většinu e-book formátů Vám můžeme pomoct převést do formátů HTML a Word). Knížky v elektronické podobě se dají koupit za příznivou cenu v internetových obchodech.
   (c) Další možností je získat texty i se smlouvou na omezené citování přímo z nakladatelství. Možnosti konverze z různých formátů lze konzultovat emailem s hlavním koordinátorem.
4. Ke skenování z papírové předlohy přistupte až po ověření, že text není možné získat jiným způsobem a že ho bude možné zarovnat.

3 Skenování

Skenováním se zde rozumí sejmutí elektronického obrazu tištěné předlohy pomocí skeneru a „přečtení“ znaků v textu programem pro optické rozpoznávání znaků (OCR). V projektu InterCorp k tomuto účelu používáme program FineReader (`http://www.abbyy.com/`).

- Návod k programu FineReader najdete na adrese `http://finereader.abbyy.com/guide/`.

3.1 Skenování a uložení textu

1. Údaje z tiráže zapište do databáze (snažte se o skenovaném textu zjistit co nejvíce informací). Kromě jiného nezapomeňte vyplnit pole Originál a Jazyk originálu (Ty se vyplňují pouze jednou u české verze. Povinné údaje jsou vyznačeny žlutě.). Všechny údaje z databáze jsou důležité pro využívání korpusu (např. při specifikaci subkorpusu).
2. Spusťte program ABBYY FineReader.[1] Tím se automaticky otevře nová Nepojmenovaná dávka.
3. V menu Soubor zvolte Uložit dávku jako... a dávku uložte (doporučujeme vytvořit název podle konvence InterCorpu, ze kterého je hned zřejmé, o jakou knihu a verzi jde, např. `Brown-Vecere_v_horach.cs`). Dávka je vlastně složkou, do které se ukládá úplně vše, tj. naskenované obrázky stránek, konfigurace programu, rozvržení stránky i výsledný rozpoznaný text.
4. V okně Nástroje→Možnosti→2.Číst→Jazyk rozpoznávání nastavte příslušný jazyk pro rozpoznávání.
5. Vložte knihu do skeneru a klepněte na Skenovat. Objeví se dialogové okno skeneru a starý obraz naposledy skenovaných stránek. Klepněte na Preview – tím se načte aktuální obraz stránek, které budete skenovat. Zároveň se objeví i rámeček, který většinou přesahuje kontury textu. Nastavte rámeček tak, aby se zbytečně neskenovalo i „okolí“. Nastavíte-li rámeček úplně přesně, bude pak potřeba přesně zachovat polohu knihy při dalším skenování.

[1] Uvedený postup odpovídá verzi 8. U novějších verzí jsou některé ovládací prvky uspořádány jinak. Také z toho důvodu je třeba uvedený postup považovat nikoli za závazný, ale pouze za jeden z více možných. Na základě vlastních zkušeností vám může vyhovovat jiný postup.

6. **Neskenujte**: obsah, různé tabulky, obrázky a popisky obrázků, seznamy slov, předmluvu, doslov, tiráž.

7. Klepněte na Scan. Poté, co je obraz stránky naskenován, dialogové okno skeneru zmizí a naskenovaný obraz se objeví vlevo pod číslem 1, a také uprostřed jako obrázek; vpravo je pak okno textového editoru, které oznamuje *Text není rozpoznán*.

8. Klepněte na Číst. Je-li rozpoznáno i něco nežádoucího (např. čísla stránek – když jste nastavili rámeček při skenování obrazu větší, než bylo nutné), můžete nežádoucí rámce dodatečně vymazat nebo upravit. Když kurzorem vyberete nežádoucích zelené rámečky a stisknete *delete* na klávesnici, daný rámec se vymaže včetně textu v rozpoznané oblasti. Když velikost rámečku upravíte přetažením okraje, je potřeba znovu přečíst stránku nebo daný rámec. Úpravy můžete ovšem dělat i ve výsledném wordovském dokumentu, takže třeba ta čísla stran můžete smazat až tam.

9. Takto pokračujte s každou stránkou – jen s tím rozdílem, že už nebudete muset nastavovat rámeček, jestliže jste ho nastavili správně. Vždy se ujistěte se, že je naskenováno všechno – to zjistíte třeba ve spodním okně FineReaderu, kde je obraz zvětšený, takže je jasně vidět, zda např. nechybí první či poslední řádky textu.

10. Text knihy musí být ve výsledku spojen **do jednoho souboru**. Když jsou naskenovány všechny stránky, přistupte k exportu (uložení) ideálně do formátu RTF nebo MS Word. Klepněte na Uložit. Objeví se okno, ve kterém je nutné nastavit následující parametry:

    - Název souboru: Zvolte tak, abyste při další práci s textem s jistotou poznali jak `id` textu, tak jakzykovou verzi.
    - Uložit jako typ: Rich Text Format (`.rtf`).
    - Uložit strany: Všechny strany (zaškrtnout tuto volbu, nikoli jen vybrané strany!)
    - Volby souboru: Vytvořit samostatný soubor pro všechny stránky.
    - Zachování nastavení: Zachovat font a velikost fontu.
    - Uchovat obrázky: zrušit, je-li zaškrtnuto.
    - Vedle Uchovat obrázky je volba Nastavení formátů – klepněte na ni, a pak zaškrtněte Odstranit volitelné spojovníky.
    - Pak klepněte na Uložit.

11. Uložený text je třeba zkorigovat. To lze provádět na libovolném počítači s editorem, který si poradí s formátem `RTF`. Pokyny ke korekturám viz část 3.2. Konverzi do textového formátu před zarovnáním pomocí makra ICorpExport je však možné provést pouze v editoru MS Word (viz část 4.1).

3.2 Korektury naskenovaného textu

Výstupní text musí projít pečlivou korekturou. Při korektuře dodržujte tyto zásady:

- Převedený text by měl co nejvíc odpovídat originálu, a to zejména v členění na odstavce, interpunkci, uvozovkách a diakritických znaménčích.

- **Zachovávejte odstavce.** Dodržujte členění textu na odstavce podle předlohy. Odstavce je nutné oddělovat dvěma znaky konce odstavce (dvojím stisknutím klávesy *enter*, tedy jedním prázdným řádkem), jinak je konverzní program při dalším zpracování textu nerozezná. Při zobrazení netisknutelných znaků ve Wordu se znak konce odstavce ukáže jako ¶ (máte-li v nabídce Nástroje→Možnosti, v oddíle Značky formátování zaškrtnuto Konce odstavců).

- **Pozor na konce stránek**: na konci stránky udělá FineReader při rozpoznávání znaků automaticky odstavec, i když tam třeba odstavec nekončí. Znak konce odstavce na takovém místě je nutné odstranit. Podobně je to i s rozdělovníkem.

- Všechny druhy **uvozovek** musí být v souboru uloženy jako znaky s významem uvozovek, tedy nikoli jako jiná interpunkční znaménka: čárky a apostrofy (jednoduché nebo zdvojené), menšítka a většítka (jednoduchá nebo zdvojená). Zásady používání uvozovek určuje pro každý jazyk jeho koordinátor, ale obecně by měla být dodržována sazba v knize a konvence v daném jazyce. Můžete vybírat z následujících znaků, přičemž dvojice odlišných znaků pro otevírací a uzavírací uvozovky mohou být pro další zpracování a využití korpusu výhodnější. V posledních dvou sloupcích uvádíme klávesové zkratky pro vložení odpovídajícího znaku v operačním systému MS Windows. Klávesové zkratky fungují při vložení čtyřciferného kódu na numerické klávesnici se současným stiskem klávesy levý Alt.[2]

Popis	Ukázka	Otevírací	Uzavírací
anglické dvojité uvozovky	“abc”	Alt+0147	Alt+0148
rovné dvojité uvozovky	"abc"		
české dvojité uvozovky	„abc“	Alt+0132	Alt+0148
francouzské dvojité uvozovky	«abc»	Alt+0171	Alt+0187
německé dvojité uvozovky	»abc«	Alt+0187	Alt+0171
anglické jednoduché uvozovky	‘abc’	Alt+0145	Alt+0146
rovné jednoduché uvozovky	'abc'		
české jednoduché uvozovky	‚abc‘	Alt+0130	Alt+0146
francouzské jednoduché uvozovky	‹abc›	Alt+0139	Alt+0155
německé jednoduché uvozovky	›abc‹	Alt+0155	Alt+0139

  Mezi nejčastější chyby tohoto druhu patří **dvě čárky** (, ,) místo českých dvojitých otevíracích uvozovek („), **jedna čárka** (,) místo otevírací české jednoduché uvozovky (‚),[3] **dva apostrofy** (’ ’) místo znaků pro české uzavírací uvozovky (“).

  Kromě uvozovek se na začátku přímé řeči používá také dlouhá nebo krátká **pomlčka** (– a —), jen zřídka spojovník (-). Odpovídající klávesové zkratky jsou Alt+0150 Alt+0151. Nahrazovat pomlčky uvozovkami může být problematické: může jít o narušení záměru autora textu a jednoznačné určení konce uvozeného úseku může být obtížné. Proto nahrazování pomlček uvozovkami obecně nedoporučujeme.

- **Interpunkční znaménka** musí být umístěna **stejně jako v předloze**. Čárky a tečky většinou následují hned za předcházejícím slovem – **bez mezery**. Stejně tak vykřičník, otazník, dvojtečka, středník, výpustka (trojtečka - Alt+0133), závorky, uvozovky – pokud v originále nejsou oddělené od slova – musí být bez mezery spojené se slovem.[4] Za interpunkční znaménka naopak mezera patří.

- Na **mezery** je třeba dávat pozor obecně. Mezerery jsou důležité jak pro rozpoznávání konců vět, tak pro rozdělení věty na slova. Proto nesmí být v textu mezery uprostřed slova navíc nebo naopak chybět za tečkou nebo přímou řečí. Opět platí zásada, že mezery mají být umístěny **stejně jako v předloze**, není však žádoucí napodobovat grafickou podobu předlohy vkládáním dvou a více mezer.

- Podle uvážení koordinátora lze rozlišovat **spojovníky** (-) a **pomlčky**, ať už krátké (–) nebo dlouhé (—).

- Další zvláštní znaky jsou třeba výpustka (…, jinak též trojtečka, elipsa – Alt+0133) nebo tečka uprostřed (· pro katalánštinu – Alt+0183), atd. Přitom tři tečky v textu není potřeba nahrazovat výpustkou, protože tuto změnu zařídí makro (viz 4.1 níže).

- Naskenovaný text může a nemusí odpovídat předloze i co do řezů písma – **tučné**, *kurzíva*, případně ***tučná kurzíva***. Záleží na rozhodnutí koordinátora pro daný jazyk, může se týkat i konkrétního textu. Písmo jiného řezu by se však nemělo vyskytovat tam, kde je v textu řez základní (stojaté, netučné písmo) – FineReader v tom často chybuje.

[2] Váš textový editor může být nastaven tak, že při stisknutí klávesy s uvozovkou nebo apostrofem se do souboru vloží přímo správný znak. To závisí také na nastavení jazyka a kontextu – jde-li o uvozovku otevírací nebo uzavírací. Rovné uvozovky pak lze vložit tak, že v nastavení editoru vypnete automatické nahrazování odpovídajících znaků při psaní.

[3] V běžném písmu tyto dva znaky od sebe často nelze rozlišit, ale uložené v souboru mají odlišný kód. Najít lze třeba hledáním textu «mezera><čárka>").

[4] Francouzština patří mezi výjimky: některá interpunkční znaménka stojí samostatně, oddělená od sousedních slov mezerou.

V tabulce na adrese `http://utkl.ff.cuni.cz/~rosen/public/tabulka-problemy_skenovani.pdf` jsou uvedeny příklady nejčastějších chyb vzniklých při skenování a doporučení, jak je odstranit.

4 Úpravy textu před zarovnáváním

Texty, které už dříve prošly zpracováním v ÚČNK, jsou už připravené k zarovnávání ve správném formátu v editoru paralelních textů InterText (část 2, bod 3a; o InterTextu viz část 5). Můžete je zpracovávat dále podle části 5 a tuto část přeskočit. [5] Pokud takový text chcete zkontrolovat s paralelní verzí, kterou dosud zpracováním v ÚČNK neprošla (ideálně ještě před začátkem skenování), lze ho většinu stáhnout přímo z databáze textů `https://intercorp.korpus.cz/DocDatabase/`.

Texty získané jinak než přímo z ÚČNK (viz část 3 a část 2, body 3b a 3c) je nutné převést do formátu vhodného pro zarovnání. Doporučený postup je následující:

1. Otevřete text v editoru MS Word.
2. U textu, který neprošel vaší korekturou, zkontrolujte dodržení zásad z části 3.2.
3. Exportujte text z editoru MS Word pomocí makra ICorpExport (viz dále část 4.1).

Všechny úpravy textu doporučujeme provést pokud možno ještě v této fázi, před exportem pomocí makra a předáním výsledku do ÚČNK. Text exportujte pomocí makra až po tom, co se ujistíte, že jste provedli všechny požadované kroky a že v textu nejsou chyby, které by bylo možné odstranit.

V textech se před zarovnáním vyznačí značkami hranice odstavců (`<p id=...>`) a vět (`<s id=...>`), případně sekcí (`<div id=...>`). Vyznačí se také řezy písma, pokud je původní text rozlišuje, a to značkami pro tučné písmo `<b>`, kurzívu `<i>` apod.

Hranice odstavců a řezů písma zanese do textu makro ICorpExport, hranice vět pak sada programů na serveru ÚČNK. Automaticky zjištěné hranice vět se v ÚČNK pokud možno ještě kontrolují a opravují.

Takto zpracované texty nahrajeme automaticky do editoru paralelních textů InterText. Pokud jsou v InterTextu už obě verze textu (česká i cizojazyčná), provede se automaticky i zarovnání a text se zobrazí v seznamu příslušného koordinátora.

4.1 Export textů pomocí makra ICorpExport

Úpravy textu před odesláním do ÚČNK provede speciální makro v editoru MS Word.

- Návod k instalaci makra najdete v příloze A.2.
- Před použitím makra musí být text ve formátu, který umí Word zobrazit (nejčastěji `.doc` nebo `.rtf`).
- Pokud jste získali elektronické verze textu v jiných formátech a nevíte si rady s jejich konverzí, pokusíme se Vám poradit.
- Popis operací, které makro provede, najdete v části A.1.

Postup

1. Po překontrolování textu a zajištění úprav vyžadovaných při skenování (viz část 3) spusťte makro ICorpExport, které soubor uloží v požadovaném formátu s příponou `.txt`.
2. Podrobný návod k použití makra najdete v příloze A.3. Zde uvádíme jeho stručnou verzi.
   (a) Nejprve si soubor zálohujte, protože makro soubor před konverzí automaticky ukládá i s případnými změnami, které v něm během kontrol před exportem provedete.

[5] Obvykle je předem k dizpozici pouze česká verze.

(b) Makro ICorpExport spustíte kliknutím na příslušné tlačítko na nástrojové liště (pokud bylo při instalaci makra vytvořeno), nebo přes nabídku Nástroje→Makro→Makra.

(c) Pokud makro zjistí, že mezi některými dvěma odstavci není prázdný řádek, přeruší svůj běh a umožní Vám prázdný řádek vložit na zvýrazněné místo.

(d) V jednom textu se může tento problém opakovat, proto doporučujeme zbytek souboru zkontrolovat makrem CheckParagraphs a makro ICorpExport spustit znovu, teprve až dojdete na konec textu. Makro CheckParagraphs pokračuje v kontrole vždy od pozice kurzoru.

(e) Makro ICorpExport se zastaví také v případě, že text obsahuje tabulky nebo obrázky, s nimiž si neví rady. Takový objekt lze smazat nebo – je-li žádoucí, aby obsah tabulky byl v korpusu obsažen – převést na text. Můžete to zkusit třeba v Wordu funkcí Tabulka→Převést→Tabulku na text.

(f) Makro se nakonec ještě zeptá, kam se má exportovaný soubor s příponou `.txt` uložit.

3. Po exportu předejte text koordinátorovi pro příslušný jazyk, který ověří bezchybnost textu a pak ho nahraje do Databáze textů. Kromě vyexportovaného souboru ve formátu `.txt` je nutné odevzdat i původní soubor `.doc` nebo `.rtf`.

4. ÚČNK následně zajistí u textů rozdělení na věty, automatické zarovnání a nahraje text do InterTextu, kde je připraven k finální kontrole zarovnání (viz část 5).

5 Zarovnávání

Jakmile koordinátor odevzdá opravený a zkontrolovaný text do Databáze textů, ÚČNK zajistí automatické rozdělení textu na věty, poloautomatickou opravu tohoto členění a nakonec nahraje text do InterTextu, kde se provede automatické zarovnání programem Hunalign.[6]

Přes veškerou snahu a pravidelné vylepšování programů pro segmentaci vět a automatické zarovnání je stále v zarovnáních vytvořených pouze pomocí automatických nástrojů řada chyb, které mohou ztěžovat spolehlivou práci s korpusem. Proto se text po automatickém zarovnání nahrává do editoru paralelních textů InterText,[7] který umožňuje texty projít a opravit.

V současné době jsou k dispozici dvě verze InterTextu. Jedna jako internetové rozhraní, které můžete otevřít pomocí běžného webového prohlížeče na stránce `https://intercorp.korpus.cz/intertext` a druhá v podobě nativní aplikace InterText editor, kterou si můžete stáhnout na adrese `http://wanthalf.saga.cz/intertext` dole. Na stejném místě je k dispozici i podrobný návod k používání InterText editoru. Ve většině případů si vystačíte s webovou podobou InterTextu, bez nutnosti instalace a s přístupem k textům odkudkoliv, kde máte k dispozici internetové připojení. Nativní aplikaci InterText editor se vyplatí použít zejména v těchto situacích:

- Potřebujete pracovat bez možnosti připojení k internetu.
- Pracujete s jazykem, který se zapisuje zprava doleva. Taková zarovnání se dají ve webovém rozhraní prohlížet, ale webové prohlížeče nedokážou během editace s pravo-levým textem pracovat správně, takže ve webovém InterTextu je jeho editace nepohodlná až nemožná.
- Potřebujete v textu provést hromadné náhrady. Typický příklad: potřebujete v celém textu nahradit jeden znak, který se během OCR chybně rozpoznává.
- Opravujete zarovnání, kde jsou časté nepřeložené kapitoly a potřebujete spouštět automatické zarovnání na konkrétní části textu.
- Když si jen nevěříte a chcete mít k dispozici funkce UnDo a ReDo.

Detailní návod pro práci s nativním editorem InterText v angličtině najdete na adrese: `http://wanthalf.saga.cz/InterText-editor.pdf`. Pro většinu práce na projektu InterCorp vám však bude stačit internetové rozhraní. Internetovému rozhraní a pravidlům, podle kterých se texty pro InterCorp upravují, se budeme věnovat v dalších kapitolách.

[6] Viz `http://mokk.bme.hu/en/resources/hunalign/`.
[7] Viz `http://wanthalf.saga.cz/intertext`.

5.1 Ovládání InterTextu

Návod k internetové verzi editoru paralelních textů InterText najdete na stránce `https://intercorp.korpus.cz/intertext/help.php#almanager`.

5.2 Opravy zarovnání

Když už víte, jak InterText ovládat, vysvětlíme si, co je vlastně cílem oprav a jak je dělat co nejefektivněji. Během oprav zarovnání v InterTextu je třeba sledovat hlavně dva okruhy problémů: chybně vyznačné hranice vět (část 5.2.1) a chybně zarovnané věty nebo jejich části (část 5.2.2).

5.2.1 Opravy chybně rozpoznaných hranic vět

Věty v textech vyznačuje pro každý jazyk sada nástrojů. Pro většinu jazyků jde o specifickou konfiguraci programu Punkt,[8] ale používáme také program Tokenize[9] pro češtinu i naše vlastní nástroje, např. pro hindštinu. Pak text prochází kontrolou pomocí sady regulárních výrazů, která pomáhá najít obvyklá místa, kde automatické nástroje selhávají, ta pak tým InterCorpu opravuje ještě před nahráním textu do InterTextu. I poté však mohou v textech zůstat místa, kde jsou hranice vět vyznačeny nesprávně. Taková místa byste měli během kontroly zarovnání najít a opravit. Nejčastěji se podobné chyby vyskytují za zkratkou nebo když je v textu chybně naskenovaná uvozovka a jiná interpunkční znaménka. Pro určování hranice věty používáme následující pravidla:

Konec věty je vždy:

- po tečce, následované mezerou a velkým písmenem, pokud není za zkratkou, iniciálou apod. uprostřed věty;
- po středníku;
- po otazníku, vykřičníku, dvojtečce a výpustce, následuje-li mezera a velké písmeno.

Další pravidla, která je třeba uplatňovat:

- uvozovky a závorky by měli zůstat s větou, ke které náleží;
- vloženou větu bez koncové interpunkce nedělit;
- přímou řeč neoddělovat od uvozovací věty, pokud to není po dvojtečce následované velkým písmenem;
- pomlčka uvozující přímou řeč by měla zůstat na začátku věty, ke které náleží.

Takže správně rozdělené věty jsou např.:

> *John T. Unger pocházel z rodiny, kterou v Hadesu – v malém městě na řece Mississippi – znali už po několik generací:*
>
> *Johnův otec tam obhájil? titul mistra golfu! v mnoha rozčilujících soutěžích.*
>
> *Paní Ungerovou znali „od kováře až po notáře“, jak se v Hadesu říkalo: pro její politické projevy;*
>
> *a mladý John T. Unger, který právě dovršil šestnáct let, tančil všechny nejnovější tance z New Yorku dřív, než oblékl dlouhé kalhoty.*
>
> *„Pamatuj si, že tady jsi vždycky vítán,“ řekl.*
>
> *„A at' uděláš co uděláš, nic ti neuškodí.*
>
> *Jsi Unger z Hadesu.“*

[8] Viz `http://nltk.org/_modules/nltk/tokenize/punkt.html`

[9] Implementace pravidel dělení textu věty pouze pro češtinu od Pavla Květoně.

> *Otec se několikrát pokoušel prosadit, aby je vyměnili za něco trochu vervnějšího, co by víc dýchalo elánem, jako třeba „V Hadesu na vás čeká příležitost“, nebo aspoň za prostý nápis „Vítejte“, umístěný nad srdečný stisk rukou, pěkně vyšitých elektrickými žárovkami.*
>
> *Když jim řekl, odkud pochází, zeptali se ho vždycky žoviálně:*
>
> *„Tam máte hezky horko, co?“*
>
> *Byl by odpovídal srdečněji, kdyby tenhle vtip nedělali všichni – nanejvýš s obměnou jako „Máte tam dost teplo?“*
>
> *Což nenáviděl zrovna tak.*
>
> *„Četl jsem ve Světovém almanachu,“ začal John, „že v Americe je jeden člověk s příjmem nad pět milioů ročně a čtyři lidé s příjmy nad tři miliony ročně a. . . “*
>
> *„To nic není,“ Percy stáhl ústa do pohrdavého půlměsíce.*

Tato pravidla platí pro většinu jazyků a většinu situací. Nejde však o neměnné fromule. Cílem jsou správně vyznačené hranice vět,[10] takže existují výjimky, ve kterých lze ustoupit. Například když se ve Vašem jazyce chová středník jiným způsobem než v češtině. Také by konec věty za dvojtečkou být neměl, když po ní následuje pouze seznam jmen, nápis nebo jiný podobný celek, který nelze považovat za samostatnou větu,

> *Štefan si nejprve vylovil obálku, na níž byl odesílatel, který ho zaujal: Svaz německých psychologů.*

Během opravy zarovnání v InterTextu by se už texty neměly znovu číst celé. Chyby v dělení vět se asi nejsnáze najdou tak, že sledujeme segmenty, které jsou správně zarovnané, ale nejsou zarovnané 1:1. U takových segmentů je pak vhodné ověřit, že věty jsou skutečně vyznačeny správně. Další možnost je sledovat zkratky na konci věty. Spolu s opravou zarovnání, by to mělo ve většině případů stačit.

5.2.2 Opravy chyb v zarovnání

Během oprav zarovnání je cílem zajistit, že v jednotlivých segmentech[11] je proti sobě vždy odpovídající překlad. Stejně jako u kontroly vět, není ani u kontroly zarovnání žádoucí text znovu číst, protože by taková oprava zabrala neúměrně mnoho času.

Při opravách zarovnání vycházíme z předpokladu, že na obou stranách zarovnání jsou všechny věty přeložené. Pokud je tomu opravdu tak, automatické zarovnání proběhně obvykle hladce a na stránce bez chyb bude vidět nejčastěji segmenty typu 1:1, ve kterých si věty odpovídají délkou, nebo zarovnání jiného řádu (nejčastěji 2:1, 1:2, 2:2), kde si obě strany segmentu budou délkou odpovídat.[12] Takovou stránku můžete po zběžném shlédnutí označit za zkontrolovanou. Při opravách se potom zaměřujeme právě na bližší kontrolu segmentů, které si délkou vět neodpovídají. Pokud některý segment není vůbec přeložený (např. 0:1) nebo délka překladu je výrazně jiná, musíte zkontrolovat, zda jsou daný segment i jeho okolí správně zarovnané. V tom případě postupujeme následovně:

- Najdeme nejbližší segment předcházející chybě, který je na první pohled v pořádku.
- Ověříme, že následující segment začíná v obou jazycích stejně. Mělo by stačit zkontrolovat prvních pár slov ve větě, ale je třeba dát pozor na jazyky s opačným slovosledem apod.
- Pokud začátek segmentu souhlasí, přejdeme k ověření konce segmentu. Pokud nesouhlasí ani začátek segmentu, jde nejčastěji o nepřeloženou větu. Pak je třeba najít místo, kde už si překlad opět odpovídá, a před toto místo vložit prázdné segmenty na stranu, kde překlad schází.

[10] Podle pravidel konkrétního jazyka

[11] Segment je nejmenší jednotka zarovnání, která je nejčastěji složena z jedné věty na české straně a jedné věty odpovídajícího překladu na straně cizojazyčné. Takový segment nazýváme 1:1. Ne vždy však jsou segmenty zarovnané 1:1. Např. když je na české straně souvětí, které je na cizojazyčné straně přeloženo více jednoduchými větami, můžeme mít třeba zarovnání 1:3.

[12] Při porovnávání délky vět u některých jazyků je třeba brát v úvahu, že stejný obsah se vyjadřuje delšími slovy nebo více slovy. Např. stejný text přeložený z češtiny do němčiny bude na německé straně zhruba o pětinu delší.

- Pokud začátek segmentu souhlasil, ověříme, zda souhlasí i konec segmentu. Opět by většinou mělo stačit zkontrolovat několik posledních slov. Pokud souhlasí i konec, označíme segment za zkontrolovaný, a to i v případě, že např. není přeložena jedna z vedlejších vět nebo je překlad stručnější (což bývá důvod nesrovnalostí v délce vět).
- Pokud konec segmentu nesouhlasí, je třeba připojit následující segment nebo naopak oddělit některé věty ze současného segmentu tak, aby si obsah v daném segmentu na obou stranách zarovnání odpovídal.

Tímto způsobem je třeba projít celý text a zkontrolovat nebo opravit všechny nesrovnalosti, na které v textu narazíte.

6 Zpracování zarovnaných textů

1. Zarovnané texty lze využívat pomocí programu ParaConc. Návod k programu ParaConc (anglicky) najdete na adrese `http://www.athel.com/paraweb.pdf`. Texty v potřebném formátu si můžete vyexportovat z InterTextu pomocí funkce Export format: ParaConc.
2. ÚČNK zpracuje exportované texty do formátu TEI-XML tak, aby byly připraveny k využívání pomocí centrálního korpusového manažeru.
3. Dodatečně je v textech možné označit jednotlivá slova (`<w id=...>`) k podrobnějšímu značkování (určení základního tvaru a morfologických kategorií slov) nebo zarovnávání po úsecích kratších než věta. To provádí ÚČNK.

7 Evidence textů

- Texty a jejich stav se sledují v Databázi textů projektu InterCorp na adrese `https://intercorp.korpus.cz/DocDatabase/`.
- Informace o jednotlivých textech nejsou součástí textů samotných, ale jsou uloženy v databázi. Vazba mezi záznamem v databázi a vlastním textem je zajištěna pomocí identifikátoru textu.
- Koordinátoři pro jednotlivé jazyky a ÚČNK vedou na webových stránkách projektu evidenci textů a postupu jejich zpracování.
- U každého textu se uvádějí jeho bibliografické údaje, odkaz na osobu, která za text odpovídá, typ textu a příznaky aktuálního stavu zpracování. Z těchto údajů se generuje hlavička souboru v korpusu.
- Příznaky stavu textu:
  1. text naplánován koordinátorem
  2. text text je schválený pro zpracování
  3. text je v papírové podobě
  4. text je v elektronické podobě
  5. text je označkován
  6. stav zarovnání (u cizojazyčných textů):
     - text je zarovnán po odstavcích
     - text je zarovnán po větách (automaticky)
     - text je zarovnán po větách (zkontrolováno)
- Příznaky právního zajištění:
  1. žádná smlouva
  2. omezené citování
  3. otevřený text (nekomerčně)
  4. otevřený text
  5. ústně / s vědomím nakladatelství

7.1 Přístup do databáze

- Přístupové jméno a heslo vystavuje ÚČNK automaticky koordinátorům jednotlivých jazyků, ostatním zájemcům až po odsouhlasení příslušnými koordinátory.
  - Důvodem pro toto opatření je, že každý uživatel smí editovat veškeré záznamy o textech jen ve „svém“ jazyce (a případně českou kanonickou verzi, kterou sám vytvořil).
  - Heslo do databáze si můžete změnit v menu User. Pokud heslo zapomenete, musíte kontaktovat ÚČNK.
- S databází se pracuje pomocí webového rozhraní přístupného na adrese: `https://intercorp.korpus.cz/DocDatabase/`. Pro vstup je nejprve nutné zadat výše zmíněné uživatelské jméno a heslo. Po přihlášení se zobrazí nabídka s položkami: moje texty, vybrat texty, přidat nový text, správa pracovníků, výplaty, přehledy, uživatel a odhlásit se. Tlačítko odhlásit se slouží k ukončení práce s databází a návratu na domovskou stránku projektu InterCorp.

7.2 Postup při zařazování nového textu do databáze

1. Nejprve zkontrolujeme, zda daný text v české verzi v databázi už není. Z menu zvolte odkaz vybrat texty. Na nové stránce pak můžete zvolit různá vyhledávací kritéria, kterými se dá omezit výpis hledaných textů. Všechny filtry pracují současně (logické AND), tzn. že lze vyhledávat například dramata, jen překlady s „William“ ve jméně autora. Výsledkem hledání budou hlavně překlady Shakespearových her, které jsou k dispozici v ÚČNK. Nad seznamem vybraných textů je seznam filtrů, které byly při výběru aktivní. Pokud nenajdete očekávaný text, zkuste si zkontrolovat, zda nemáte zapnutý některý filtr navíc nebo zkuste zjednodušit dotaz. Pokud si např. nejsem jistý, zda je anglická autorka zapsaná v databázi česky nebo anglicky, raději zadám do položky autor pouze Rowling a databáze potom vypíše všechna díla všech autorů, kteří by měli ve jméně Rowling, tedy i Rowlingová, pokud je tato podoba jména u některých titulů použita. Pokud najdete v databázi českou verzi, můžete bod 2 preskočit.
2. Databáze je nastavena tak, že nejprve musíte zadat českou verzi textu (odkaz přidat nový text), při tomto kroku je zvolen identifikátor, který bude určovat text během budoucí práce a který už nebude možné měnit. Výběru identifikátoru proto věnujte zvýšenou pozornost. Zároveň je automaticky nastaven jazyk textu na češtinu. Vyplňte pokud možno co nejvíce údajů, které jste o daném textu schopni zjistit. Náplň jednotlivých položek je popsána přímo na stránce, pokud není zřejmá ze samotného názvu. Po uložení vyplněných údajů se text zobrazí v seznamu moje texty.
3. Pak je třeba zobrazit detail záznamu. Pokud jste záznam zadávali sami, můžete ho vybrat v přehledu moje texty. Pokud už existoval, najdete ho pomocí vybrat texty. Musíte vybrat konkrétní záznam kliknutím na odkaz tvořený identifikátorem textu. Zobrazí se Vám informace o textu nebo seznam verzí, pokud je jich už v systému víc.
4. Je-li zobrazen seznam verzí nebo detail české verze, je v menu k dispozici nový odkaz přidat novou verzi.
   - Odkaz přidat novou verzi slouží k přidání cizojazyčné verze textu. Jazyk verze je zvolen podle toho, jaký jazyk máte právo editovat. Identifikátor je použit stejný jako u verze české, ostatní údaje zadejte podle dané knihy. Předpokládá se, že např. údaje o anglické verzi textu budou v angličtině.
5. Pokud v seznamu verzí kliknete na konkrétní jazykovou verzi, např.: en-00, zobrazí se detaily dané verze. Spolu s detaily se v menu zpřístupní opět několik nových odkazů:
   - všechny verze textu – zobrazí seznam všech jazykových verzí daného textu

   Pokud jste osoba odpovědná za daný text, zobrazí se i možnosti:
   - editovat záznam – umožňuje změnit údaje o textu nebo stav
   - smazat záznam
   - předání textu – slouží k předání souborů `.doc` a `.txt` do ÚČNK

7.3 Záznam údajů o stavu textu

Celou práci by mělo průběžně provázet informování všech vašich kolegů o postupu práce na daném textu. K tomu slouží zápis textu do databáze a aktualizace jeho stavu zpracování. Následuje shrnutí průchodu textu zpracováním v projektu InterCorp spolu s tím, jak by měly být jednotlivé fáze práce zaznamenány v databázi.

1. Během prosince by si koordinátoři měli připravit zhruba 15 textů, na kterých by rádi pracovali během příštího roku. Tyto texty je potřeba na začátku ledna zadat do Databáze textů. (Stav bude nastaven automaticky na: plán)
2. V průběhu ledna dojde v ÚČNK ke schválení některých z připravených textů na základě stanovených pravidel (A.6) a v počtu odpovídajícímu finančním prostředkům na daný rok. Na schválených textech lze začít pracovat. (Stav bude nastaven automaticky na: schválený plán)
3. Ověřte si v databázi, zda je česká verze textu už k dispozici.
4. Poté, co daný text získáte v knižní nebo jiné papírově podobě, je vhodné doplnit v databázi scházející bibliografické údaje.
5. Až student odevzdá naskenovaný text, koordinátor zkontroluje cizojazyčnou verzi. Při kontrole by koordinátor měl odhadnout míru chybovosti pečlivým přečtením několika náhodně vybraných odstavců a letmou kontrolou celého textu. Při velkém počtu chyb je nutné text vrátit studentovi k opakované korektuře.
6. Koordinátor by měl zkontrolovat i českou verzi, pokud ji student skenoval sám a nebyla vydána ze zdrojů ÚČNK.
7. Pak je čas na předání textu do ÚČNK. K tomu je potřeba vybrat konkrétní záznam o cizojazyčné verzi textu – buď pomocí menu vybrat texty nebo přímo ze seznamu pod položkou moje texty. Potom se v menu objeví odkaz předání textu. Když na tento odkaz kliknete, zobrazí se formulář, jehož pomocí můžete nahrát texty do databáze. Formulář je rozdělen do několika bloků:

   - Upload souborů – cs: dva řádky, ve kterých máte vybrat `.doc` a `.txt` soubory pro českou verzi. Soubor `.doc` není k úspěšnému nahrání textu povinný, ale doporučený. Když soubor(y) vyberete, klikněte na Odeslat soubory a ty se nahrají na sever ÚČNK. Pokud je nahrání úspěšné, přejde formulář rovnou k vyplnění dalšího bloku. Pokud se nahrání nepovede, vypíše formulář přehled chyb, které uploadu zabránily. Můžete je zkusit odstranit a zkusit text nahrát znovu nebo stačí vyplnit 2. blok a oznámit problém hlavnímu koordinátorovi e-mailem. Texty se na server uloží i v případě chyby, takže je není nuté zasílat spolu s oznámením.
   - Výběr spolupracovníků – cs: tři řádky, ve kterých máte vybrat spolupracovníky odpovědné za jednotlivé kroky skenování. Skenováním se v tomto kontextu myslí pouze fyzické skenování pomocí skeneru (tedy snímání obrázků stran). Formátováním se myslí doladění textu do podoby vyžadované projektem. Tedy např. odstranění poznámek pod čarou, čísel stránek, obrázků apod. Tuto operaci je obvykle nutné dělat u všech textů.

   Pokud je česká verze už označkovaná, 1. blok se nezobrazuje vůbec a 2. blok je deaktivovaný, takže je vidět, kdo text zajistil, ale nelze nic měnit.

   - Upload souborů – en: 3. blok funguje stejně jako 1., vztahuje se ale k jazykové verzi, kterou právě editujete. Např. koordinátor angličtiny má možnost předání textu pouze u anglických verzí. Zde bude předávat anglický soubor `.doc` a `.txt`.
   - Výběr spolupracovníků – en: Výběr spolupracovníků odpovědných za jednotlivé kroky skenování cizojazyčné verze textu.

   Opět platí, že pokud je daná verze textu už označkovaná, 3. blok se nezobrazí a 4. blok je neaktivní. Pokud je označkovaná jak česká, tak cizojazyčná verze textu. Jsou vidět 2. a 4. blok jako neaktivní a zobrazí se dva nové bloky, které se vztahují k nahrávání zarovnaných souborů z programu Paraconc. Tyto bloky se už nepoužívají, protože texty se teď v dalších krocích spravují pomocí InterTextu a zůstavají v databázi jen kvůli zpětné kompatibilitě. Pokud nepracujete se starými soubory, nic zde nevyplňujte.

8. Pokud se povede soubory úspěšně nahrát, ÚČNK zajistí segmentaci (vyznačení hranic vět) a po kontrole je text nahrán do InterTextu a automaticky zarovnán. Dokončení této operace je automaticky označeno změnou stavu na: text je označkován. Stejně bude označen text, pokud je vydán z archivu ÚČNK.

9. Po opravě zarovnání v InterTextu a kontrole hlavním koordinátorem je stav cizojazyčné verze opět automaticky nastaven na: zarovnán po větách (zkontrolováno).

7.4 Správa spolupracovníků

Menu správa spolupracovníků slouží k vytváření a editaci záznamů o všech, kteří se podíleli na přípravě textů. Pokud chcete při nahrání nového textu vybrat někoho jako zodpovědnou osobu za jednotlivé operace, je třeba nejprve založit záznam o této osobě ve správě spolupracovníků. To platí i pro samotné koordinátory, takže každý koordinátor by měl zadat do databáze nejprve své vlastní údaje. Součástí registrace je i vytvoření účtu pro přístup do InterTextu. V případě koordinátorů je potom ještě potřeba přepnout tento učet do režimu KOORDINÁTOR, což musí na požádání udělat hlavní koordinátor.

Po kliknutí na odkaz správa spolupracovníků se zobrazí jednoduchá stránka s jedinou rozbalovací nabídkou. Po rozvinutí je k dispozici seznam už dříve založených záznamů a položka Nový spolupracovník – tou vytvoříte nový záznam.

Po vybrání požadované položky se zobrazí detail záznamu. Vyplňte všechna povinná políčka a stiskněte Odeslat na konci stránky.

- Heslo do InterTextu je kvůli možným překlepům nutné zadat dvakrát.
- Pokud je daná osoba cizinec, po přepnutí hodnoty v poli Státní příslušnost se změní seznam polí, které je nutné vyplnit. (**EDIT:** Protože se v databázi už nadále nespravují smlouvy DPP, není už potřeba vyplňovat osobní data spolupracovníků. U většiny údajů, prosíme, uveďte pouze hodnotu X nebo 0 podle požadavků kontroly a vyplňujte pouze e-mail.)
- Některá pole, jako třeba e-mail, vyžadují správný formát dat.
- Smazání záznamu je k dispozici po vybrání konkrétního záznamu v detailech.
- Smazáním daného záznamu se neprovede úplné odstranění osoby z databáze. Kvůli návaznosti dat (záznamy o výplatách, apod.) dojde pouze k deaktivaci záznamu a vymazání většiny osobních údajů, ale vlastní ID a s ním spojená data zůstávají. Pokud dojde k smazání nedopatřením, nezakládejte proto nový záznam, ale požádejte ÚČNK o obnovení původního záznamu.

7.5 Výplaty a přehledy

U každého spolupracovníka, který má založen záznam ve správě spolupracovníků, je v databázi veden virtuální účet, na který jako kladné položky přibývají odměny za jednotlivé operace na textech a jako záporné položky se odečítají skutečně provedené platby. V menu přehledy můžete pro každou osobu, kterou máte ve správě spolupracovníků, zobrazit přehled všech textů na kterých pracoval.

1. V části Vykonané práce jsou v řádcích abecedně seřazeny jednotlivé texty, u kterých byl dotyčný během nahrávání do databáze vybrán jako zodpovědná osoba. České a cizojazyčné verze jsou uváděny zvlášť. Ve sloupcích jsou potom všechny operace, které se s textem běžně provádějí.7
   - Hned po nahrání naskenovaného textu do databáze jsou ve všech sloupcích nuly.
   - Po kontrole a nahrání textu do InterTextu, se doplní odpovídající částky ve sloupcích: Skenování, Korektura a Formátování, pokud nebyla některá operace označena jako operace se nekonala, a pokud všechny operace prováděla osoba, jejíž záznam si právě prohlížíte.
   - Po úpravě textu v InterTextu a její kontrole hlavním koordinátorem je text uzavřen a odměna za skenování a koordinaci se zobrazí v příslušných sloupcích.
2. V části Bylo vyplaceno se úvádějí jednotlivé výplaty.

- V případě výplat pres DPP je uveden poslední den v měsíci, za který byla platba prováděna, ačkoliv na účet přichází peníze reálně až následující měsíc kolem desátého.
- V případě faktur je uváděn datum splatnosti faktury.

3. Na konci je potom celkový rozdíl předchozích částí, tedy stav účtu daného spolupracovníka u ÚČNK.

V menu výplaty potom vidí koordinátor stav účtů všech spolupracovníků, kteří jsou pod ním v Databázi textů vedeni.

A Přílohy

A.1 Co dělá makro ICorpExport

- **Úpravy textů nutné pro použití v InterTextu**:

  1. Převedení do podoby holého textu v kódování UTF-8, se kterým pracuje InterText.
  2. Nahrazení některých znaků, které je třeba vyhradit pro značkování (<,>, &).
  3. Další úpravy usnadňující práci s textem (např. nahrazení tří teček znakem „výpustka“, tedy …).

  Tyto změny jsou vratné a text v konečném formátu korpusu bude v původním nebo požadovaném stavu.

- **Změny potřebné k následnému zpracování** – explicitní vyjádření formátu textu pomocí značek jazyka HTML a úprava těchto značek do takové podoby, aby soubor vyhovoval standardu XML.

  1. označení odstavců (`<p id=...>`)
  2. označení řezů písma (kurzíva `<i>`, tučné písmo `<b>` atd.)

  Značky pro odstavce a věty budou po zarovnání použity k vytvoření linkovacích souborů, které zajistí vlastní propojení jednotlivých vět mezi různými jazykovými verzemi.

- Podrobnější popis funkcí makra je uveden v příloze A.3.

A.2 Instalace makra ICorpExport

Instalace makra pro Word (Word 2000 a pozdější; je možné, že starší verze balíku MS Office je nutné upgradovat — není ověřeno).

1. Z adresy `http://korpus.cz/intercorp/files/ICorpExport.dot` si stáhněte šablonu pro Word, která obsahuje makro.

2. Zkopírujte soubor ICorpExport.dot z adresáře, do kterého se uložil po stažení, do adresáře C:\Program Files\Microsoft Office\Office\STARTUP.[13]

3. Spusťte Word.[14] Na kartě Zobrazení klikněte na tlačítko Makra. V seznamu maker by se Vám měla zobrazit nová makra CheckParagraphs a ICorpExport. Pokud už máte nainstalovnán větší počet maker, můžete pro přehlednost vybrat v poli prohledat pouze šablonu ICorpExport.dot.

   - Před použitím maker budete možná ještě muset povolit jejich spouštění ve Wordu. Toto nastavení najdete pod tlačítkem Office → Možnosti aplikace Word → Centrum zabezpečení → Nastavení Centra zabezpečení → Povolit všechna makra.

   Tím jsou potřebná makra připravena k použití.

4. Makra lze spouštět buď pomocí tlačítka Makra, nebo si vytvořit tlačítka přímo pro konkrétní makro v pásu karet pro pravidelné a snadné použití. Tato možnost bohužel chybí zrovna v MS Word 2007, ale v pozdějších verzích je už (opět) k dispozici.

   Postup pro vytvoření tlačítek v pásu karet:

   (a) Pravým tlačítkem myši klikněte kamkoli na pás karet a v kontextovém menu zvolte Přizpůsobit pás karet.

   (b) V nabídce Zvolit příkazy z: v levém sloupci vyberte položku Makra.

[13] Adresář 'Office' v adresáři 'Microsoft Office' se může podle verze MS Office jmenovat také 'Office12' nebo 'Office14' apod. Důležité je, aby v daném adresáři byla už předem složka 'STARTUP'.

[14] Postup je platný pro MS Word 2007, ale měl by se shodovat is s verzí 2010 a 2011.

(c) Potom si v pravém sloupci vyberte kartu, na které chcete tlačítka umístit a tam vytvořte novou skupinu (tlačítko Nová skupina vpravo dole).

(d) V levém sloupci klikněte na makro ICorpExport v pravém na nově vytvořenou skupinu a potom uprostřed na tlačítko Přidat.

(e) Potom můžete v pravém sloupci kliknout na přidané tlačítko a dole na Přejmenovat a změnit název tlačítka a případně zvolit vhodnou ikonu.

A.3 Návod k použití makra ICorpExport

1. Po úspěšné instalaci by váš Word měl v pásu karet obsahovat dvě nová tlačítka: ICorpExport a CheckParagraphs (pokud byla při instalaci makra vytvořena). Jinak jsou makra dostupná přes kartu Zobrazení→Makra. K samotnému exportu slouží makro ICorpExport, makro CheckParagraphs je jeho součástí, ale dá se spustit samostatně.

2. Makro nejprve uloží otevřený dokument, proto si dejte pozor na změny, které v textu během pokusů o export uděláte. Pokud nemáte soubor uložen ještě na jiném místě, nebude možné vzít provedené změny po skončení práce s makrem zpět.

3. Pak makro zkontroluje, zda jsou za každým odstavcem dva znaky konce odstavce. Tuto operaci provádí i makro CheckParagraphs, které pouze kontroluje odstavce, a to vždy od místa, kde skončila předešlá kontrola. Makro ICorpExport musí začínat s kontrolou vždy od začátku souboru. V dlouhých textech tedy doporučujeme následující postup:

   (a) Export začněte spuštěním makra ICorpExport.

   (b) Pokud makro narazí na chybu v označení odstavců, přeruší svůj běh a umožní vám opravit chybu, kterou zároveň označí zvýrazněním.

   (c) Protože se dá předpokládat, že v delším textu budou ještě další chyby, pokračujte v kontrole pomocí makra CheckParagraphs.

   - **Protože makro CheckParagraphs kontroluje vždy od pozice kurzoru, může se stát, že když umístíte kurzor na volnou řádku před další odstavec, nahlásí makro chybu i tam, kde dva znaky konce odstavce jsou. V tomto případě stačí umístit kurzor někam do středu předešlého odstavce a pokračovat v kontrole.**

   (d) Když dojdete na konec textu, spusťte opět makro ICorpExport a proveďte export textu.

4. Makro dále kontroluje, zda soubor neobsahuje tabulky nebo obrázky, se kterými si neumí poradit při exportu do textového souboru. Texty pro korpus InterCorp by takové objekty obsahovat neměly, pokud však makro na tabulku přesto narazí, opět přeruší svou činnost a zvýrazní objekt, který nevyhovuje podmínkám. Pokud je obsah tabulky důležitý pro význam textu, můžete použít například funkci Wordu na kartě Nástroje Tabulky→Rozložení→Převést na text, nebo vyjmout důležitý obsah tabulky ručně a zbytek smazat.

5. Po prvních kontrolách se makro zeptá, kam má zkonvertovaný text uložit a jak ho má pojmenovat. Zvolte požadované umístění a jméno souboru a makro bude pokračovat ve svém běhu.

   - **S doposud uvedenými typy chyb, které makro kontroluje v první části svého chodu, se budete běžně setkávat a jejich oprava by měla být jednoduchá. V další fázi konverze už by se chyby měly vyskytovat výjimečně, a pokud si nejste jisti, proč se chyba vyskytla, raději kontaktujte hlavního koordinátora. Zkopírujte obsah okna s chybovým hlášením do emailu a zašlete ho na adresu** `martin.vavrin(zavináč)ff.cuni.cz`**.**

6. Následuje nahrazení všech znaků, které brání úspěšnému zpracování textu v InterTextu za kódy (tzv. *znakové entity*) formátu HTML nebo Unicode. Do této kategorie spadají znaky `<`, `>`, `&`, které v upraveném textu slouží k oddělení značkování textu od textu samotného atp.

7. Pro pozdější použití jsou do textu zaznamenány řezy písma pomocí značkování v textu (značky HTML — např. `<b>`slova tučně`</b>`).

A.4 ParaConc

- program pro vytváření a prohlížení paralelních korpusů
- pro systém MS Windows
- `http://www.athel.com/para.html`
- příručka (anglicky): `http://www.athel.com/paraconc.pdf`
- Předpoklady pro instalaci:
  - operační systém MS Windows 95 a vyšší
  - při instalaci ve Windows 95 je třeba minimálně 16 MB RAM, jinak 32 MB
  - pro uložení vytvořeného korpusu, zpracovaného programem ParaConc, je třeba na disku prostor 2–20 MB, případně více
- Instalace: Soubor o velikosti asi 1,4 MB zkopírujeme kamkoli na disk (pokud vám to systémová oprávnění umožňují, nejlépe do složky *Program Files*, se zástupcem na ploše).

A.5 Často kladené dotazy

A.5.1 Úpravy českého textu

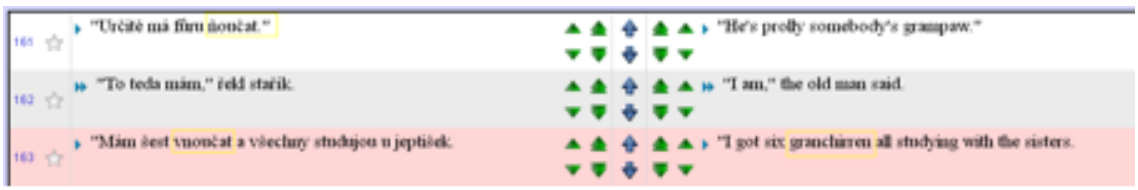
Otázka: Chápeme to tak, že zatímco v anglické verzi textu můžeme dělat opravy, česká verze je vzhledem ke svému statusu pivotového jazyka v českém textu pro úpravy uzamčena. Přesto bychom se chtěli ještě jednou ujistit, jak provádět či neprovádět úpravy v české verzi textu. Jde především o úpravy (a) pravopisu a (b) dělení vět. Můžete zásady a postup těchto úprav prosím ještě jednou vysvětlit?

Odpověď Ad (a): Pokud bych ten pravopis přepsal na korektury chyb, tak ty můžete dělat i v češtině, dokonce je to žádoucí. Jenom to samozřejmě musí být ku prospěchu věci. Např. neopravovat v hovorové řeči pravopis ;D. Ale samozřejmě běžné chyby OCR apod. určitě opravit.

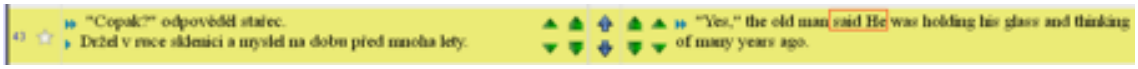
Ad (b) Pokud je chyba v dělení vět v češtině, tak je také žádoucí to opravit, ale bohužel to nejde snadno. Vzhledem k tomu zmiňovanému pivotu, je třeba opravit všechna zarovnání v ostatních jazycích tak, aby spojení nebo rozdělení věty umožňovala. Proto v případě češtiny je třeba pouze udělat v tom místě bookmark (záložku) a opravu samotnou musíme udělat v ÚČNK.

A.5.2 Srovnání se zdrojovým textem

Otázka: Existují nějaká pravidla, kdy úpravy pravopisu a dělení vět provádět? V některých textech si nejsme jistí, zda jde o chybu, nebo záměr autora / překladatele (například Medvídek Pú možná využívá schválně pozměněný pravopis, naopak v textu Stařec a moře zřejmě chybějí texty v českých větách).



Obrázek 1: Medvídek Pú



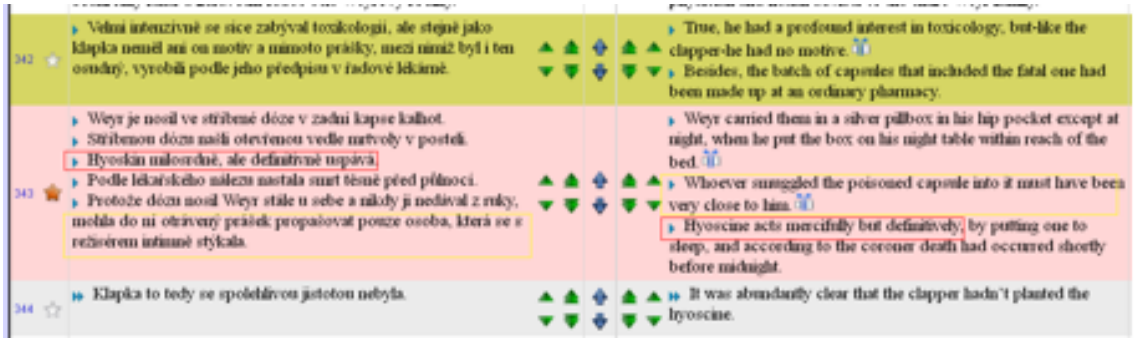
Obrázek 2: Stařec a moře

Ideální by bylo mít možnost srovnat tyto případy se zdrojovým textem. Chci se tedy zeptat, jak v takových případech postupovat?

Odpověď: V normálním případě míváte ten zdrojový text k dispozici, protože to nejprve skenujete, ale i tak by bylo dost práce to místo v knize vyhledat a zkontrolovat. Trochu lépe se to dá řešit, když máte uložený sken ještě ve FineReaderu, tam se dá prohledávat a přitom tam vidíte originál. V případě jako je ASPAC nezbývá, než se řídit citem.

A.5.3 „Překřížení“ sémantické informace

Otázka: Věty, které obsahují korespondující si sémantickou informaci, jsou ve výchozím textu a v překladu „přehozeny“. Pokud by měly být součástí jednoho segmentu, je nutné spojit všechny věty do tohoto jednoho segmentu, který pak narůstá. Je to v pořádku? Je nějaká hranice pro velikost segmentu?



Obrázek 3: Překřížení

Odpověď To se občas stát může. Zase jde hlavně o cit. Hranice, od kdy už to nedávat do jednoho segmentu, je tak asi délka středního odstavce. Příklad na obrázku bych ještě nechal pohromadě, kdyby to mělo být ještě tak asi o 2 dlouhé věty delší, tak už bych nechal vždy tu jednu větu proti prázdnému segmentu.

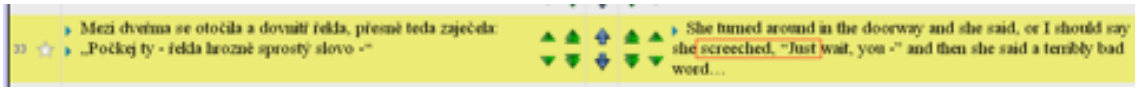
A.5.4 Jednotka textu, která se zobrazí při hledání

Otázka: Obecně jsme se chtěli zeptat, co je základní jednotka, která se zobrazí uživateli vyhledávajícímu v korpusu – je to jeden segment bez ohledu na jeho velikost?

Odpověď: To je trochu složité. Záleží hodně na okolnostech. V základě je to opravdu jeden segment, v Kontext je to v podstatě výchozí jednotka, ale lze si nastavit i zobrazování jen vět nebo jen určitého počtu slov kolem KWICu. Stejně dobře si lze nastavit větší kontext. Navíc pokud je zobrazeno víc jazyků, tak se musí ten segment občas rozšířit, v případě, kdy nejde o zarovnání 1:1 tak, aby se zobrazil odpovídající úsek ve všech jazycích.

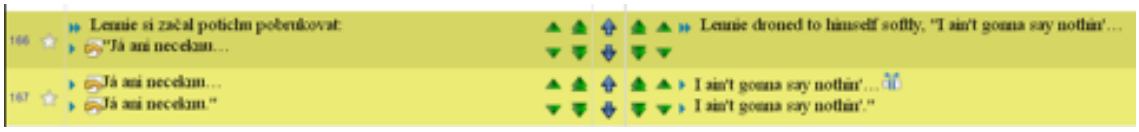
A.5.5 Pravidla pro dělení vět

Otázka: Obecná pravidla pro dělení vět se vztahují především na český úzus zápisu interpunkce. Chtěli bychom se zeptat, zda existují i pravidla pro další jazyky? Angličtina například odděluje uvozovací větu od přímé řeči čárkou, následuje velké písmeno. V češtině se používá dvojtečka, což pravidla pro dělení upravují. Pokud by se i anglická věta rozdělila na dvě, bylo by možné zarovnat kratší segmenty (viz níže). Jak se tedy má v takových případech postupovat?



Obrázek 4: Ukázka dělení vět v angličtině

Dále pravidla pro dělení vět udávají, že pokud po výpustku následuje velké písmeno, začíná nová věta. Předpokládáme, že toto neplatí, pokud je velké písmeno součástí vlastního jména. Je to tak?



Obrázek 5: Další ukázka dělení vět v angličtině

Odpověď: To je podobné jako předchozí problém, ale ještě větší guláš :D. Cílem by mělo být to, aby byly odděleny věty, tak jak jsou v daném jazyce chápány jako samostatná jednotka. Stejná pravidla aplikujeme, protože většinou dávají podobné výsledky. Samozřejmě v různých jazycích jsou různé výjimky, ale v průměru to vychází takhle celkem dobře. Třeba Váš první příklad: anglická forma se v češtině občas vyskytuje také, zvlášť kolem 80-tých prý byla naprosto běžná, jak mi tu našeptává kolega, a naopak je běžné, že v angličtině se používá dvojtečka. V okamžiku, kdy bychom začali opravovat místa v uvedených příkladech, tak už z toho nevyjdeme a můžeme všechna pravidla hodit do koše. Zrovna tyhle věci bych spíš nechal. I když je to vždycky nakonec zas jen o citu toho, kdo to dělá. Třeba když budu mít v textu:

> *„Opravdu?“ řekl Johny.*

Tak to bude jedna věta. Ale kdyby to samé bylo v angličtině, tak to podle pravidel dopadne takto:

> *“Really?”*
> *Johnny said.*

V tom případě je na místě to opravit a spojit, ačkoli to je proti pravidlům, protože ta uvozovací věta nedává bez spojení smysl. Ještě horší je to potom třeba v němčině, kde se navíc píší s velkým písmenem i podstatná jména. Obecně – kde to alespoň jakž takž dává smysl, tak bych se snažil držet pravidel, aby se v zarovnaných textech dalo alespoň na něco spolehnout. Pokud je to příliš proti rozumu a citu, tak je možné udělat výjimku tak, aby výsledek dával pokud možno smysl.

A.5.6 Spojování vět v segmentu (souvisí s předchozím dotazem)

Otázka: anglická verze segmentu je někdy rozdělena do více vět zarovnaných proti jedné české. Máme v takovém případě anglické věty spojit, nebo ponechat beze změny? (zřejmě opět problém chybějících pravidel pro dělení vět v angličtině)

Odpověď: Ano, viz. předchozí bod. Věty by měli být vyznačeny tak, aby dávaly smysl jako věty. Zarovnání je pak druhá věc, pokud na jedné straně odpovídá dlouhé souvětí 6 jednoduchým větám na druhé straně, tak je to v pořádku.

A.5.7 Věty zarovnané proti nule

Otázka: Pokud je více vět automaticky zarovnáno proti nule, je nutné rozdělit věty tak, aby byla každá věta v jednom segmentu, nebo je naopak všechny sloučit do segmentu jednoho (nebo vůbec neřešit)?

Odpověď: Neřešit. Dokud je to 0:N, tak je jedno kolik je N. Jestli bude celá stránka v jednom segmentu nebo to bude po jedné větě, je jedno. Dokonce mám dojem, že do indexu v NoSke se to ve výsledku komprimuje, aby to bylo vždy maximální N ku 0.

A.5.8 Posun většího počtu segmentů

Otázka: V textu chybí na jedné straně celá kapitola a další text je špatně zarovnán. Co s tím dělat?

Odpověď: Najděte první segment, který lze opět správně zarovnat za nepřeloženým úsekem. Posuňte první větu po chybějícím úseku na správné místo. K posunutí segmentu o více řádků můžete použít modrou šipku označující začátek věty. Po kliknutí na šipku zadáte číslo segmentu, kam se má celý následující text posunout. Když máte správně zarovnaný nepřeložený úsek N:0 a první větu po něm. Dejte znovu zarovnat zbytek textu.

A.6 Jaké texty plánovat

Pro výběr platí zavedená pravidla, seřazená podle klesající priority. Texty, u nichž lze odpovědět kladně na následující otázky, mají větší šanci:

1. Je už v korpusu originál textu?
2. Je už v korpusu více jazykových verzí textu? Pro budování společného "pevného jádra" se lze stále inspirovat také seznamem titulů s nejvyšším počtem překladů. Seznam textů setříděných podle počtu překladů najdete tady: `https://intercorp.korpus.cz/files/IntercorpTop200/`, nebo si seznam nechte vypsat v databázi textů [15]. Můžete přitom použít libovolný další filtr.
3. Pomáhá text udělat korpus reprezentativnější – z hlediska typu textu, autora, směru překladu, v rámci jazyka? Pomoci Vám může opět seznam původních titulů v různých jazycích "Jaké knížky mi v InterCorpu chybí". Překlady těchto titulů doporučují zařadit do korpusu koordinátoři pro jazyk originálu. Seznam můžete i nadále průběžně upravovat a rozšiřovat: Seznam knih pro InterCorp.
4. Je text dostupný elektronicky? Elektronickou dostupnost textu uvádějte v poznámce. Je-li text k dispozici korektně digitalizovaný přímo v nakladatelství jako e-book, uveďte do poznámky "e-book" – pak nepočítáme s odměnou za korektury a pro daný jazyk je možné schválit v plánu více textů. Pokud ale text korektně digitalizován není a korektury jsou nutné, tak s překročením limitu na počet schválených textů nelze počítat.

Výběr můžete ovlivnit v položce Priorita. Neschválené texty můžete znovu navrhnout příští rok. [16] Může se stát, že se během roku najdou další finance na zpracování nových textů a některé z neschválených titulů by tak bylo možné zpracovat ještě letos. Stane-li se tak, komise provede dodatečný výběr.

[15] Vybrat texty→Výpis (úplně dole)→Řadit dle→autora
[16] Editovat záznam→Plán na rok:→Uložit