Projekt *InterCorp* – postup přípravy textů verze 1.0

24. února 2008

Obsah

Seznam tabulek

Seznam obrázků

Poděkování

- Děkujeme všem autorům za materiály i dalším, zde neuvedeným účastníkům projektu za všechny cenné podněty a připomínky:

  1. Eliška Boková *Zarovnávání v programu ParaConc* (část 5)
  2. Jan Kocek *Zásady skenování textů* (část 3)
     `http://korpus.cz/intercorp/dokumenty/zasady_skenovani_InterCorp.pdf`
  3. Pavel Stichauer *Skenování a export textu do Wordu (instrukce pro naprosté začátečníky)* (část 3)
     `http://www.pavel-stichauer.cz/dokumenty/sken-instrukce.rtf`
  4. Pavel Stichauer *Zarovnávání v programu ParaConc – základní instrukce* (část 5)
     `http://www.pavel-stichauer.cz/dokumenty/alignment.pdf`
  5. Martin Vavřín *Postup přípravy textů pro projekt InterCorp* (části 2, 4, 5, 7 a přílohy)
     `http://korpus.cz/intercorp/dokumenty/metodika_InterCorp.pdf`
  6. Martin Vavřín *Postup při instalaci makra potřebného pro export textů z Word-u pro použití v programu ParaConc* (příloha B)
  7. Pavel Vondřička *Stručný návod k používání databáze textů InterCorp* (část 7)
  8. Alexandr Rosen *Projekt InterCorp – jak přidávat nové texty*
     `http://korpus.cz/intercorp/dokumenty/postup3.pdf` (části 4 a 5)
  9. Alexandr Rosen *Jak na paralelní texty s programem ParaConc*
     `http://korpus.cz/intercorp/dokumenty/paraconc.pdf` (příloha C)

  Na konečné verzi se spolu s hlavním kompilátorem (Alexandr Rosen) podílel spoustou důležitých připomínek Martin Vavřín. Za všechny chyby, nepřesnosti a nedostatky čiňte zodpovědným hlavního kompilátora.

- Další připomínky a podněty všeho druhu uvítáme na adrese `alexandr.rosen@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) $\Longrightarrow$ 4
   - z webu: (2, bod 3b)
     - zadarmo
     - za peníze
   - z nakladatelství (2, bod 3c) **(jen jestli neni trochu riskantni to takhle napsat na web ;)) (píše se tam "se smlouvou na omezené citování", tak snad to nevadí?)** MV AR
   - skenováním (3)
3. Úpravy před zarovnáním:
   (a) korektury (3.2)
   (b) export z Wordu (4.1)
   (c) segmentace po větách (4.2)
4. Zarovnávání:
   (a) načtení do ParaConku (5.1)
   (b) zarovnání po odstavcích (5.2)
   (c) automatické zarovnání po větách (5.3)
   (d) manuální kontrola zarovnání po větách (5.4)
   (e) export textů z ParaConku (5.5)

Software, který je třeba nainstalovat:

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

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 textu.
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, radši se ještě emailem přeptejte hlavního koordinátora.
   (b) Zjistěte si, zda by nebylo možné danou knihu získat v elektronické verzi volně na internetu nebo např. v podobě e-booku (zatím je ověřena poměrně snadná možnost exportu `.lit` soborů 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. Teprve pokud jste si ověřili, že text není možné získat jiným způsobem a že texty, které máte k dispozici, bude možné zarovnat, přistupte ke skenování z papírové předlohy.

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 používáme k tomuto účelu program *FineReader* (`http://www.abbyy.com/`).

⟹ Návod k programu FineReader najdete na adrese
`http://www.abbyy.com/download/?param=28619`.

3.1 Skenování a uložení textu

1. Údaje z tiráže zapište do databáze (snažte se o skenovaném textu zjistit co nejvíce informací).
2. Spusťte program ABBYY FineReader. Tím se automaticky otevře nová Nepojmenovaná dávka.

> Uvedený postup odpovídá verzi 8. U novější verze 9 jsou některé ovládací prvky uspořádány jinak.

3. V menu Soubor zvolte Uložit dávku jako... a dávku uložte (doporučujeme vytvořit snadno identifikovatelný název, ze kterého je hned zřejmé, o jakou knihu a verzi jde, např. `vassalli-labut-italsky`). Dávka je vlastně složkou, do které se bude ukládat úplně vše, tj. obrazy skenování, rozpoznaný text a výsledný dokument ve Wordu.
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 vašeho 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.
6. **Neskenujte**: obsah, různé tabulky, popisky obrázků, seznamy slov, předmluvu, doslov, tiráž.

7. Klepněte na Scan. Poté, co je obraz stránek 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 přečteno i něco „nežádoucího“ (např. čísla stránek – to když jste nastavili mřížku poněkud větší, než bylo nutné), můžete to následně vymazat tak, že kurzorem najedete do nežádoucích zelených rámečků a stisknete *delete* na klávesnici. Tyto akce ale můžete udělat až při korekturách ve Wordu.

9. Takto pokračujte s každou stránkou – jen s tím rozdílem, že už nebudete muset nastavovat mřížku, jestliže jste ji nastavili správně. Doporučujeme ale ujistit se, že je naskenováno všechno – to zjístíte i tak, že ve spodním okně FineReaderu máte zvětšený obraz pod lupou, takže je jasně vidět, zda např. nechybí spodní či horní řádky textu.

10. Text knihy musí být spojen **do jednoho velkého souboru**. Až budou naskenovány všechny stránky, přistupte k exportu (uložení) do formátu editoru MS Word. Klepněte na Uložit. Objeví se okno, ve kterém je potřeba nastavit následující parametry:

    - Název souboru: vytvořte podle vzoru `capek-valka_s_mloky.bg-00.doc`, tedy z příjmení autora, pomlčky a českého názvu textu (místo mezer použijte podtržítka), označení jazyka a verze překladu. Část názvu před první tečkou nesmí být delší než 22 znaků (podle potřeby zkraťte název díla či příjmení autora) a musí přesně odpovídat ID v databázi. Název smí obsahovat jen malá písmenka bez diakritiky, podtržítka a pomlčku (k oddělení příjmení autora a názvu textu). Mělo by z něj být napohled jasné, o jaký text se jedná.
    - 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:

- Obecně platí, že převedený text by měl co nejvíc odpovídat originálu, a to zejména v členění na odstavce, interpunkci a diakritických znaménkách.
- **Zachovávejte odstavce.** Dodržujte členění textu na odstavce podle předlohy. Odstavce je nutné oddělovat dvěma znaky konce odstavce (dvěma *entery*), tedy jedním prázdným řádkem, jinak je konverzní program při dalším zpracování textu nerozezná.
- **Pozor na konce stránek**: na konci stránky udělá program při rozpoznávání znaků automaticky odstavec, i když tam třeba odstavec nekončí. Znak konce odstavce (*enter*) na takovém místě je nutné odstranit.
- Všechny druhy **uvozovek** musí být reprezentovány 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. Může vybírat z následujících znaků, přičemž dvojice odlišných znaků mohou být pro další zpracování a využití korpusu výhodnější:[1]

anglické dvojité uvozovky	“abc”
rovné dvojité uvozovky	"abc"
české dvojité uvozovky	„abc“
francouzské dvojité uvozovky	«abc»
německé dvojité uvozovky	»abc«
anglické jednoduché uvozovky	‘abc’
rovné jednoduché uvozovky	'abc'
české jednoduché uvozovky	‚abc‘
francouzské jednoduché uvozovky	‹abc›
německé jednoduché uvozovky	›abc‹

Mezi nejčastější chyby tohoto druhu patří **dvě čárky** (, ,) místo českých dvojitých otevíracích uvozovek („) a **dva apostrofy** (’ ’) místo znaků pro české uzavírací uvozovky (“), rovné uvozovky ("), nebo anglické otevírací (“) nebo uzavírací (”) uvozovky.

- **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, závorky, uvozovky – pokud v originále nejsou oddělené od slova – musí být bez mezery spojené se slovem. Za interpunkční znaménka naopak mezera patří.

- Na **mezery** je vůbec třeba dávat pozor. Mezera je znak jako každý jiný, který slouží zejména k oddělování jednotlivých slov. Proto nesmí být mezera uprostřed slova. Opět platí zásada, že mezery mají být umístěny **stejně jako v předloze**, není však třeba napodobovat grafickou podobu předlohy vkládáním většího počtu mezer.

- Podle uvážení koordinátora lze rozlišovat **spojovníky** a **pomlčky**. **(je to tak?)** AR

- Naskenovaný text může a nemusí odpovídat předloze i co do řezů písma – textbftučné, *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.

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 zarovnáním

Texty, které už prošly zpracováním v ÚČNK, jsou ve správném formátu a připravené k zarovnávání (část2, bod 3a). Tyto texty lze stáhnout přímo z databáze textů `http://korpus.cz/intercorp/DocDatabase`. S těmito texty můžete dále postupovat podle části 5 a tuto část přeskočit.

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 skenováním, zkontrolujte dodržení zásad z části 3.2.
3. Exportujte text z editoru MS Word pomocí speciálního makra (viz dále část 4.1).

[1] Kromě uvozovek se na začátku přímé řeči používá také dlouhá nebo krátká pomlčka (– a —). Vzhledem k obtížím s určením konce přímé řeči v takovém případě ponecháme pomlčky.

Všechny úpravy textu je nutné provést pokud možno ještě v této fázi, před exportem pomocí makra a předáním výsledku do ÚČNK. Při zarovnávání v ParaConku už text není možné měnit, lze jen rozdělovat nebo spojovat odstavce. Po vytvoření konečné verze označkovaných souborů je každý zásah nepříjemný a může způsobovat špatně odhalitelné chyby. Makro tedy spusťte až po ujištění, že jste provedli všechny požadované kroky a že v textu nejsou chyby, které by bylo možné odstranit

4.1 Export textů z editoru MS Word

Úpravy textu před odesláním do ÚČNK provede speciální makro editoru Word a další programy, které si makro volá automaticky.

⟹ Návod k instalaci makra najdete v části B.

- Před použitím makra musí být text ve formátu, který umí Word zobrazit (`.doc` nebo `.rtf`).
- Pokud byste získali elektronické verze textu v jiných formátech a nebudete si vědět rady s jejich konverzí, pokusíme se Vám poradit.
- Popis operací, které makro provede, najdete v části A.

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.

   (Návod k použití makra.) AR
2. Po exportu předejte text do ÚČNK k registraci a označení hranic vět. Kromě vyexportovaného souboru ve formátu `.txt` je nutné odevzdat i původní soubor soubor `.doc` nebo `.rtf`.
3. ÚČNK vrátí takto zpracovaný text koordinátorovi pro daný jazyk k zarovnání (viz část 5).

4.2 Segmentace po větách

Texty exportované z Wordu pomocí makra zašlete do ÚČNK na adresu hlavního koordinátora (`martin.varin@ff.cuni.cz`). Zde projdou texty poslední fází úpravy před zarovnáním — pomocí programu pro segmentování budou označeny jednotlivé věty (značkami `<s id="n">` a `</s>`). Takto označkované soubory Vám pošle ÚČNK zpátky a můžete je začít zarovnávat.

♣ Zatím se takto věty označovaly jen v českých textech. Nově se segmentují i texty cizojazyčné.

Druhá možnost je, že text získáte přímo z archivu ÚČNK. Tyto texty projdou stejnou segmentací jako texty naskenované. Platí tedy bez výjimky, že každý text musí před zarovnáním projít kontrolou a segmentací v ÚČNK.

5 Zarovnávání

Při zarovnávání se k sobě srovnají jednotlivé odstavce, ParaConc poté automaticky, na základě hranic vět a jejich délky, přiřadí k českému textu cizí po úsecích — segmentech, které může tvořit jedna nebo více vět. V ideálním případě by ParaConc měl přiřadit jedné české větě jednu větu z druhého jazyka (přirazení 1:1). Pokud tuto dvojic nelze sestavit (na jedné straně je např. souvětí, zatímco na druhé dvě věty jednoduché), provede se zarovnání 1:2 a podobně.

V tabulce 1 je uveden příklad vstupních souborů. Nastavíte-li při exportu identifikaci segmentu pomocí značek `<seg id="n"> ... </seg>`, budou výstupní soubory vypadat jako v tabulce 2.

♣ Tabulka 2 ukazuje dvě omezení programu ParaConc:

1. Značky `<seg id="n"> ... </seg>` jsou umístěny bez ohledu na značky pro odstavce a věty. Tento nedostatek se odstraní při konverzi do formátu TEI-XML v ÚČNK.
2. Segmenty číslo 8 až 10 nejsou zarovnány správně. Zjistíte-li takovou chybu až u exportovaných souborů, je jednodušší zarovnání opravit v programu ParaConc a soubory exportovat znovu. Správně zarovnaný text je uveden v tabulce 3.

V části 5.4 uvidíte, jak se v ParaConku opravují chyby automatického zarovnávání po větách, včetně výše uvedeného příkladu.

Zatím jsme na věty segmentovali jen české texty. Cizojazyčné texty se do ParaConku načítaly segmentované pouze na odstavce, nikoli na věty. Segmentaci na věty prováděl ParaConc automaticky při zarovnání. Nově segmentujeme na věty před zpracováním ParaConkem i cizojazyčné texty. České i cizojazyčné texty tedy mají na vstupu do ParaConku stejnou formální strukturu.

5.1 Načtení textů do programu ParaConc

1. Zarovnávat budeme označkovaný český a cizojazyčný text ve formátu `.txt1`. Texty je třeba vůči sobě zarovnat v programu ParaConc.
2. Spusťte program ParaConc.
3. Klepněte na File→Load Corpus File(s). Objeví se okno Load Corpus Files.
4. Nastavení v okně by mělo být následující, překontrolujte/nastavte všechny parametry v následujícím pořadí.
   (a) Nastavení jazyků:
   Parallel texts: 2
   Jazyk vlevo: Czech
   Jazyk vpravo: cizí jazyk (*vybrat z rolovací nabídky*)
   ⟹ Pokud příslušný jazyk mezi nabízenými možnostmi nenajdete, je třeba příslušné národní prostředí do systému doinstalovat. (Obraťte se na svého správce systému, nebo vložte instalační CD systému Windows do mechaniky, klepněte na Start→Nastavení→Ovládací panely→Místní nastavení a dále postupujte podle pokynů.)
   (b) Klepněte na tlačítko Font... pro češtinu:
   Písmo: Arial
   Skript: Středoevropský
   (Položka Skript je dole uprostřed.)
   Klepněte na OK.
   (c) Klepněte na tlačítko Font... pro cizí jazyk:
   Písmo: Arial
   Skript: – podle odpovídající znakové sady:
   - Středoevropský pro češtinu, slovenštinu, polštinu, maďarštinu, chorvatštinu, slovinštinu
   - Západní pro angličtinu, němčinu, italštinu, španělštinu, francouzštinu, portugalštinu, nizozemštinu, norštinu, švédštinu, finštinu, dánštinu
   - Pobaltský pro litevštinu a lotyštinu
   - Cyrilika pro ruštinu, ukrajinštinu, bulharštinu, srbštinu, makedonštinu
   - Arabský pro arabštinu

   Klepněte na OK.

(d) Nyní je třeba nastavit informace o zarovnávání:

Align format: Not Aligned

(To je položka úplně dole, ale je třeba s ní začít.)

(e) Vraťte se doprostřed okna pod tlačítko Font... a pro český i cizojazyčný text provedeme nastavení formátu. Klepněte na příslušné tlačítko Format.... U obou jazyků je nastavení stejné:

Headings: Regular expression — hodnota prázdná, neboť nadpisy neoznačujeme

Paragraphs: HTML/SGML Markers

start tag: **p**

stop tag: **/p**

Sentences: HTML/SGML Markers

start tag: **s**

stop tag: **/s**

5. Klepněte na tlačítko Add a vyberte správný soubor s příponou `.txt1`. To proveďte u obou jazyků. Pak klepněte na tlačítko OK.

6. Po klepnutí na tlačítko OK v okně Load Corpus Files se pravděpodobně objeví chybové hlášení o tom, že počet sekcí nebo odstavců se v obou textech liší. Tuto chybu odstraníme při zarovnání.

   $\Longrightarrow$ Pokud počet odstavců odpovídá, na obrazovce se žádná tabulka neobjeví. I tak je nutné zkontrolovat, zda jsou odstavce zarovnány správně. To proto, že zarovnání nemusí být v pořádku, i když počet odstavců souhlasí. Tabulku zarovnávání lze otevřít takto:

   (a) Zvolte nabídku File→View Corpus Alignment nebo použijte klávesovou zkratku CTRL+V.

   (b) Objeví se okno s vybranými soubory, klepněte na název souboru a poté dole na tlačítko Alignment.

   (c) Měla by se objevit tabulka s přiřazením sekcí a další s přiřazením segmentů (vět). Objeví-li se jen jedna s přiřazením sekcí (obsahuje jedinou sekci zobrazenou jako jeden řádek textu), zkuste ji minimalizovat („srolovat“), měla by se ukázat i tabulka s přiřazením sekcí.

   (d) Mezi přiřazením vět a odstavců lze přepínat v nabídce Alignment→ Paragraphs a Alignment→ Aligned Sentences

5.2 Zarovnání po odstavcích

1. V okně Error klepněte na tlačítko Fix.

2. Objeví se dvě tabulky o dvou sloupcích. Jedna udává členění textů na sekce (Sections), druhá na odstavce (Alignment). Pracovat budeme s tabulkou Alignment. Odstavce jsou od sebe odděleny vodorovnou čárou. Skládá-li se odstavec z více vět, jsou tyto věty odlišeny barevně.

3. Zkontrolujte, zda dělení na odstavce zhruba odpovídá nastavení. Také si ověřte, zda se na obou stranách správně zobrazují znaky abecedy příslušného jazyka. Zjistíte-li nesrovnalosti, chyba může být v nastavení podle části 5.1, nebo přímo v textech.

4. Na konci tabulky Alignment je vidět, kolik odstavců přebývá. Výsledkem zarovnávání by měl být shodný počet správně přiřazených odstavců.

   $\Longrightarrow$ Je-li zarovnání odstavců hodně rozhozené, obraťte se na koordinátora.

5. Je třeba zjistit, které odstavce si navzájem neodpovídají, a zarovnání odstavců opravit. Po klepnutí pravým tlačítkem myši na odpovídající sloupec a odstavec lze věty přesouvat mezi odstavci a upravovat zarovnání odstavců pomocí těchto funkcí:

   - Split Paragraph
   - Merge with Next Paragraph
   - Merge with Previous Paragraph

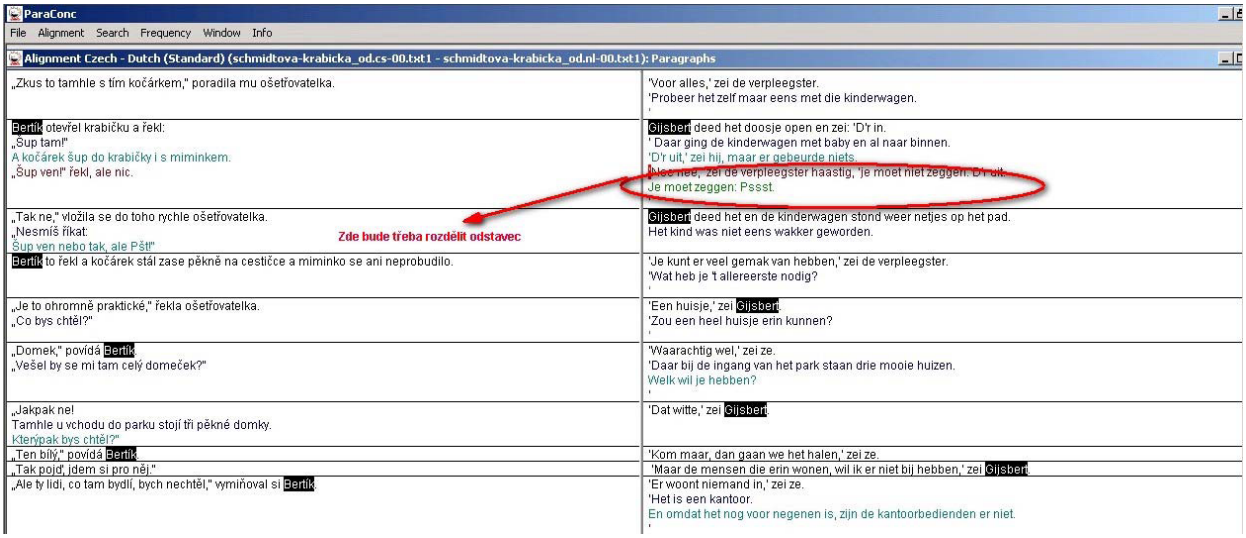
- Insert Empty Paragraph

Funkcí Split Paragraph dělte odstavce jen na hranici mezi větami, nikdy uprostřed věty.

6. Nepoužívejte funkce, které spojují a rozdělují věty:
   - Split Sentence
   - Merge with Next Sentence
   - Merge with Previous Sentence
7. Není třeba vše podrobně číst, stačí zrakem zachytit konce odstavců. Další pomůcky pro rychlejší práci:
   - Podívejte se na první a poslední odstavec na stránce. Shodují se? Pokud ano, stačí prohlédnout zbývající odstavce na stránce jen zběžně.
   - Pracujte s klíčovými slovy. V nabídce Alignment→Alignment Markers lze nastavit slova, která mají být na české i na cizojazyčné straně zvýrazněna. To se hodí hlavně pro vlastní jména. Výrazně to usnadní orientaci v textu.
   - Pokud Váš text obsahuje aspoň občas nějaká čísla, zapněte si v nabídce Alignment→Alignment Markers u obou jazyků volbu Hilite Numbers.
   - Sledujte, jestli v nějakém odstavci není výrazně víc vět. Rozdílný počet vět však není nezbytně chybou. Může se jednat pouze o odlišné dělení souvětí, což není potřeba opravovat.
8. Pokud zjistíte, že nějaký odstavec neodpovídá, klepněte do textu pravým tlačítkem myši. Členění textu můžete upravovat na české i cizojazyčné straně.
   - Klepnete-li pravým tlačítkem myši na první písmeno věty, která v odstavci „přebývá“, můžete odstavec rozdělit funkcí Split Paragraph. Je nepřípustné touto funkcí dělit věty, používejte ji vždy jen na hranici mezi větami.
   - Pokud je třeba spodní část odstavce spojit s následujícím odstavcem, vyberte Merge with Next Paragraph.
   - Pokud je v některém z jazyků nějaký odstavec navíc, vložte proti němu do textu v druhém jazyce prázdný odstavec funkcí Insert Empty Paragraph.
     - ♣ Věta, která není přeložena, může být součástí většího odstavce. Potom je nutné zarovnat okolní text správně a poznamenat si, kde se chybějící text vyskytuje. Po skončení zarovnávání po odstavcích se korpus automaticky zarovná po větách. V tomto okamžiku je nutné poznamenané místo vyhledat v textu, který je již rozdělen na segmenty, a proti nepřeložené větě vložit prázdný segment.
     - ♣ **(jeste do nejak podrobnejsi urovne zminit problem s prebyvajicim textem na konci zarovnani.. tam totiz nelze vlozit empty segment a Paraconk tedy nepovoli Align corpus.. Jestli to nikde neni, tak ten postup popisu.)** MV

   Totéž opakujte tak dlouho, až si odstavce navzájem odpovídají. Viz obr. 1.
9. Okna s tabulkami zavřeme.
10. Pokud nezvládnete zarovnat celý text najednou, lze si práci uložit: File→ Save Workspace As.... Příště už nemusíte soubory znovu načítat a od začátku zarovnávat, ale stačí uložený korpus otevřít (File→ Open Workspace...).

Při ukládání *workspace* dávejte pozor na to, abyste *workspace* ukládali na stejné místo, na kterém máte vstupní soubory `.txt1`. Pro opětovné otevření *workspace* musíte mít k dispozici jak vstupní soubory `.txt1`, tak oba soubory workspace `.bin` a `.pws`.



Obrázek 1: Úprava dělení odstavců

> ParaConc neumí uložit zarovnání odstavců jinak než jako *workspace*. Pokusíte-li se po zarovnání odstavců provést rovnou export, objeví se nesrozumitelná chybová hláška.

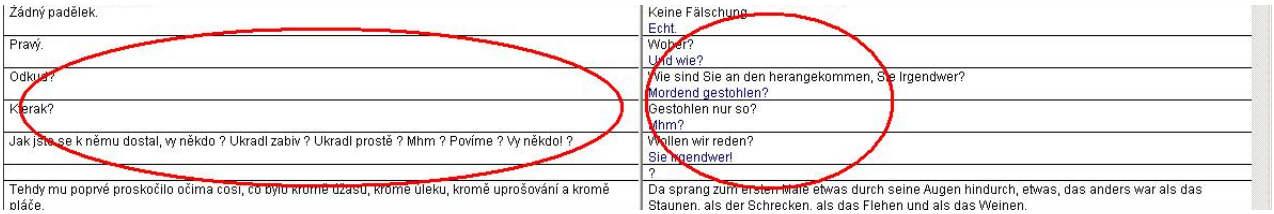
5.3 Automatické zarovnání po větách

1. Po zarovnání textu po odstavcích klepněte na File→Align Corpus. Program se pokusí texty automaticky zarovnat po větách.
2. Pokud máte okno Alignment už otevřené, můžete přejít k části 5.4.
3. Není-li okno Alignment už otevřené, klepněte na File→View Corpus Alignment.
4. Objeví se okno Select Files to View. Klepněte na soubory, které se mají zobrazit, a pak na tlačítko Alignment.
5. Opět se objeví dvě okna, ale tentokrát lze v okně Alignment prohlížet a opravovat zarovnání po větách.

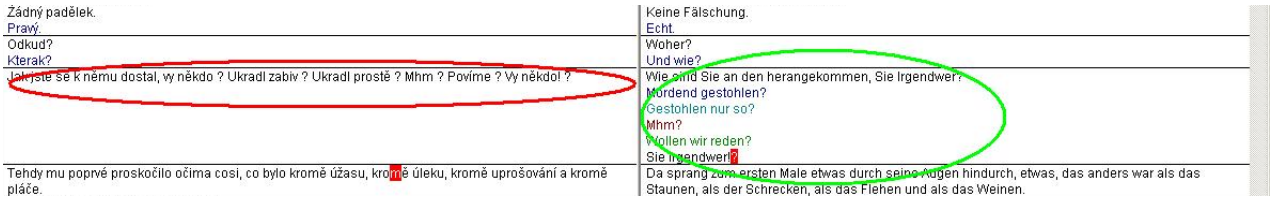
5.4 Manuální kontrola zarovnání po větách

Po automatickém zarovnání ParaConkem je nunté projít ručně i zarovnání vět na segmenty. ParaConc pracuje na základě statistických metod a délky textu, takže v některých případech může udělat chybu. Podíl nesprávně zarovnaných dvojic vět by neměl přesáhnout cca 5%.

1. Během této fáze budeme opět kontrolovat, zda obsah segmentů na jedné straně zarovnání odpovídá obsahu segmentů na druhé straně. Ve většině případů si se správným zarovnáním poradí sám ParaConc. Případy, se kterými si ParaConc sám neporadí jsou zejména tyto:
   (a) Na jedné straně zarovnání je dlouhé souvětí a na druhé 3 a více krátkých vět (obr. 2, po opravě obr. 3).
   (b) Na jedné straně je věta, která na druhé straně není vůbec přeložena (obr. 4, po opravě obr. 5).
   (c) Záludné jsou případy, kdy je na jedné straně věta odpovídající dvěma větám na straně druhé, a o kousek níže se situace opakuje, ale na opačných stranách. To byl náš příklad z tabulek 1–3. Po načtení souborů `.txt1` do ParaConku a automatickém zarovnání po větách uvidíme výsledek jako na obr. **??**.



Obrázek 2: Více krátkých vět proti souvětí



Obrázek 3: Více krátkých vět proti souvětí – opraveno

(d) Případy jako na obr. 9 není třeba opravovat.

Když přehlédnete drobné chybky, jako třeba zvolání nebo oslovení, které spadlo do špatného segmentu, tak se nic hrozného neděje, ale je třeba najít výše zmíněné chyby, které mají obvykle vliv na celý odstavec.

2. Zarovnání vět lze upravit takto:

   - Segment (text v jednom poli tabulky) lze rozdělit, ale jen na hranici mezi větami. Klepněte pravým tlačítkem myši na první písmeno „přebytečné“ věty a zvolte Split Segment. Segment se rozdělí těsně před kurzorem.
   - Merge with Next Segment, případně Merge with Previous Segment použijte tam, kde chcete spojit dva segmenty do jednoho. Vrátíme-li se k příkladu z obr. **??**, tak na obr. **??** je vidět výsledek spojení dvou anglických segmentů a na obr. **??** dvou českých.
   - Na obou stranách lze použít Insert Empty Segment tam, kde některá věta není přeložena.

♣ Stejně jako při zarovnávání odstavců si lze práci uložit: File→Save Workspace. Příště už nemusíte soubory znovu načítat a zarovnávat, ale stačí uložený korpus otevřít (File→Open Workspace...). Nezapomeňte, že soubor *workspace* musí být uložen ve stejné složce jako vstupní soubory `.txt1`.

♣ ParaConc si nepamatuje, kde jste se zarovnáváním minule skončili, ani neumí vyhledat místo v souboru podle zadaného řetězce. Proto se vyplatí si alespoň někam poznamenat přibližnou pozici posuvníku po straně okna, případně i konkrétní text, kam jsme při nedokončené opravě zarovnávání dospěli.

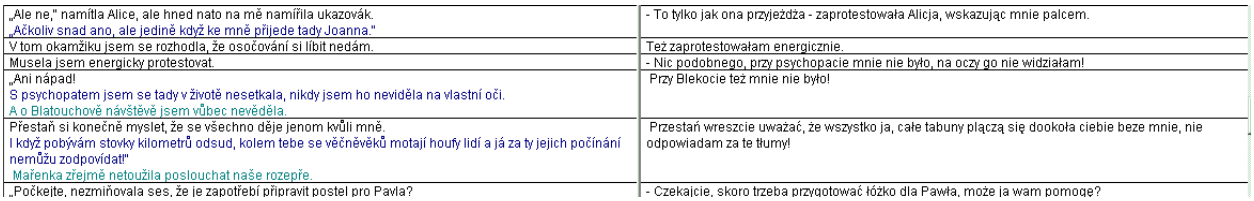
5.5 Export textů z programu ParaConc

1. K odevzdání textů do ÚČNK je třeba zarovnané texty exportovat: File→Export Corpus Files. Exportovaný soubor dostane automaticky předponu (přednastaveno `A_`).

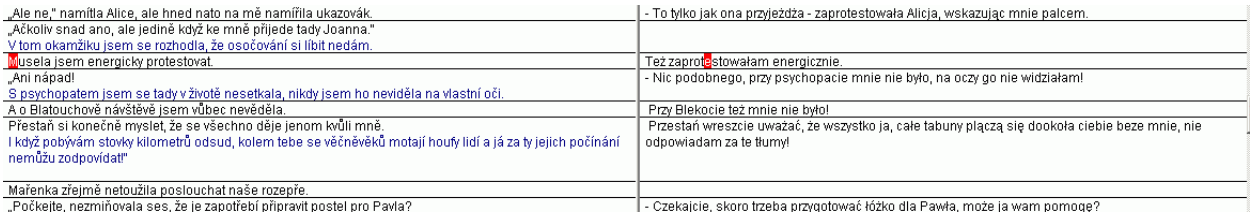
2. V Alignment Style je nutné zvolit možnost Tags. Jako způsob označení segmentů (Tag) zvolte značky `<seg>` s atributem (Attribute) `id`.

3. Po exportu přejmenujte soubory z podoby s předponou na původní jméno souboru, ale s koncovkou `.seg` místo `.txt1` (např. `A_Golding-PanMuch.cz-00.txt1` → `Golding-PanMuch.cz-00.seg`).

4. Exportované texty odevzdejte koordinátorovi a do ÚČNK hlavnímu koordinátorovi (`martin.vavrin@ff.cuni.cz`). Tyto texty lze do programu ParaConc znovu načíst jako zarovnané (Alignment style: Tags) a pracovat s nimi jako s korpusem.



Obrázek 4: Nepřeložená věta



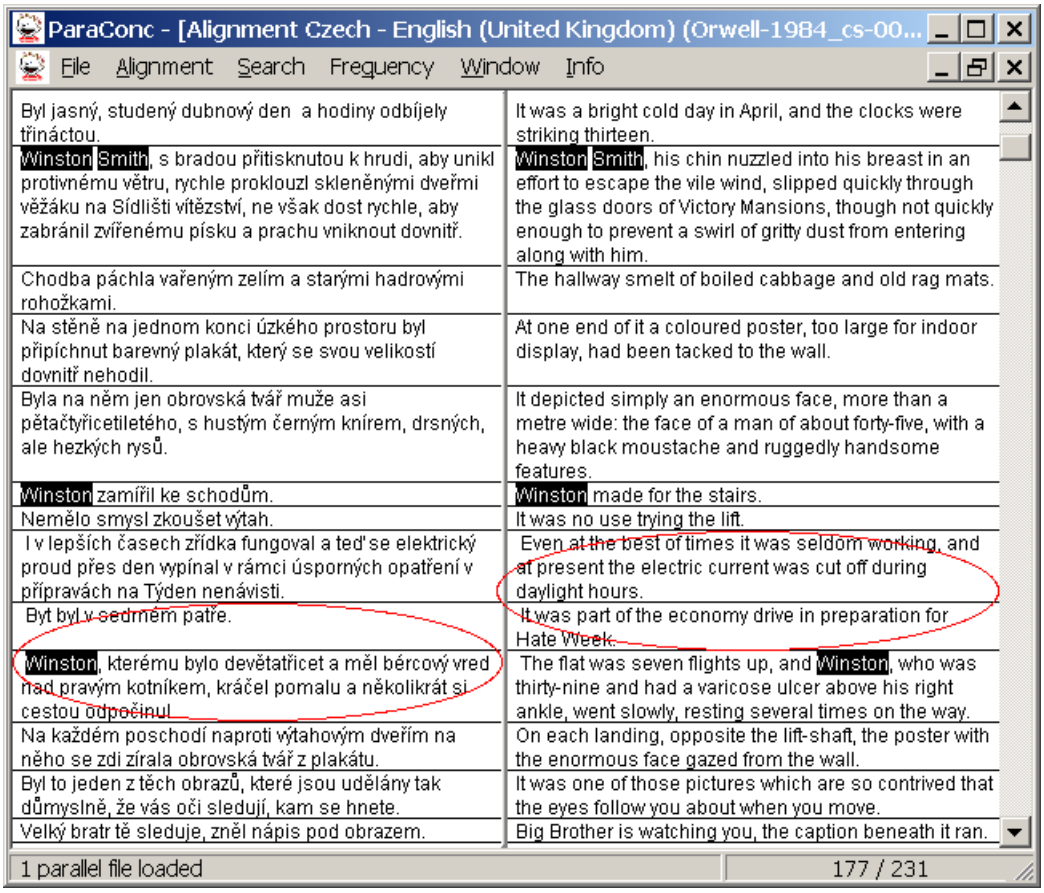
Obrázek 5: Nepřeložená věta – opraveno

6 Zpracování zarovnaných textů

1. Zarovnané texty lze využívat pomocí programu ParaConc. Návod k programu ParaConc (anglicky) najdete na adrese `http://www.athel.com/paraconc.pdf`.
2. ÚČNK zpracuje exportované texty do formátu TEI-XML tak, aby údaje o zarovnání byly ve zvláštním souboru a byly připraveny pro 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.

7 Evidence textů

1. Texty a jejich stav se sleduje v databázi textů projektu InterCorp na adrese `http://korpus.cz/intercorp/DocDatabase/`.
2. Informace o jednotlivých textech nejsou součástí textů samotných, ale jsou uloženy pouze v databázi. Vazba mezi záznamem v databázi a vlastním textem je zajištěna pomocí identifikátoru textu.
3. Koordinátoři pro jednotlivé jazyky a ÚČNK vedou na webových stránkách projektu evidenci textů a postupu jejich zpracování.
4. U každého textu se uvádějí jeho bibliografické údaje, odkaz na osobu, která za text odpovídá, typ textu, příznaky aktuálního stavu zpracování. Z těchto údajů se generuje hlavička podle TEI-XML.
5. Příznaky stavu textu:
   (a) text je v papírové podobě
   (b) text je v elektronické podobě
   (c) text je označkován
   (d) text je označkován ve formátu TEI-XML
   (e) stav zarovnání (u cizojazyčných textů):
   - text je zarovnán po odstavcích
   - text je zarovnán po větách (automaticky)
   - text je zarovnán po větách (zkontrolováno)



Obrázek 6: Po sobě jdoucí věty, které si odpovídají 1:2 a 2:1

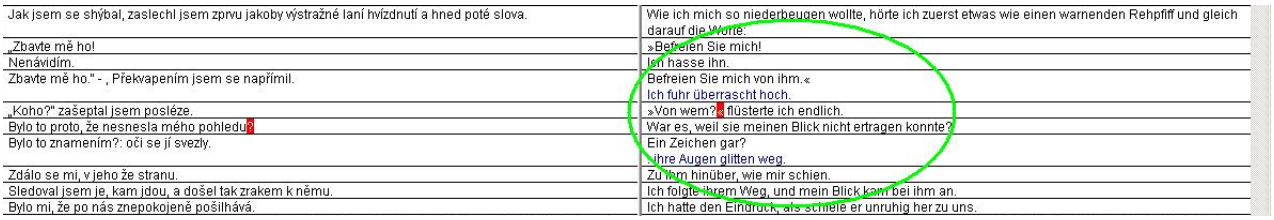
6. Příznaky právního zajištění:
   (a) žádná smlouva
   (b) omezené citování
   (c) otevřený text (nekomerčně)
   (d) otevřený text
   (e) ústně / s vědomím nakladatelství

7.1 Přístup do databáze

- K přístupu do databáze textů slouží stejné přístupové jméno a heslo jako k přístupu do diskusního fóra InterCorp. 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 ve „svém“ jazyce (a případně českou kanonickou verzi, kterou sám vytvořil).

- S databází se pracuje pomocí webového rozhraní přístupného na adrese: `http://korpus.cz/intercorp/DocDatabase`. Pro vstup je nejprve nutné zadat výše zmíněné uživatelské jméno a heslo. Při prvním přihlášení se zobrazí nabídka s položkami: moje texty, vybrat texty, přidat nový text a odhlásit se. Tlačítko odhlásit se slouží k ukončení práce s databází a návratu na domovskou stránku intercorpu.



Jak jsem se shýbal, zaslechl jsem zprvu jakoby výstražné laní hvízdnutí a hned poté slova.	Wie ich mich so niederbeugen wollte, hörte ich zuerst etwas wie einen warnenden Rehpfiff und gleich darauf die Worte:
„Zbavte mě ho!	»Befreien Sie mich!
Nenávidím.	Ich hasse ihn.
Zbavte mě ho." - , Překvapením jsem se napřímil.	Befreien Sie mich von ihm.« Ich fuhr überrascht hoch.
„Koho?" zašeptal jsem posléze.	»Von wem?« flüsterte ich endlich.
Bylo to proto, že nesnesla mého pohledu?	War es, weil sie meinen Blick nicht ertragen konnte?
Bylo to znamením?: oči se jí svezly.	Ein Zeichen gar? Ihre Augen glitten weg.
Zdálo se mi, v jeho že stranu.	Zu ihm hinüber, wie mir schien.
Sledoval jsem je, kam jdou, a došel tak zrakem k němu.	Ich folgte ihrem Weg, und mein Blick kam bei ihm an.
Bylo mi, že po nás znepokojeně pošilhává.	Ich hatte den Eindruck, als schiele er unruhig her zu uns.

Obrázek 7: Co se nemusí opravovat

7.2 Postup při zařazování nového textu do databáze

1. Nejprve zkontrolujeme, zda je k danému textu v databázi již zařazena česká verze. Nejjednodušší postup je ten, že zvolíte z menu odkaz vybrat texty. Na nové stránce je pak možné 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 jste tedy nenašli to, co jste čekali, zkuste si zkontrolovat, zda nemáte zapnutý některý filtr navíc.

2. Databáze je nastavena tak, že je nejprve nutné 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, proto výběru identifikátoru prosím věnujte zvýšenou pozornost. Zároveň je automaticky nastaven jazyk textu na češtinu. Vyplňte pokud možno co nejvíce údajů, které jste o dané knize schopni zjistit. Náplň jednotlivých položek, pokud není zřejmá z samotného názvu, je popsána přímo na stránce. Potom, co vyplněné údaje uložíte, by se měla kniha zobrazit v seznamu moje texty.

3. Když máte uložen záznam o české verzi, můžete klepnout na odkaz tvořený identifikátorem dané knihy. Zobrazí se Vám informace o knize nebo seznam verzí, pokud je jich už v systému víc.

4. Pokud je zobrazen seznam verzí, jsou v menu k dispozici nové odkazy vybrané texty a přidat novou verzi.

   - Odkaz vybrané texty zobrazí seznam textů podle aktuálně nastaveného filtru (změna pomocí odkazu vybrat texty).
   - 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.

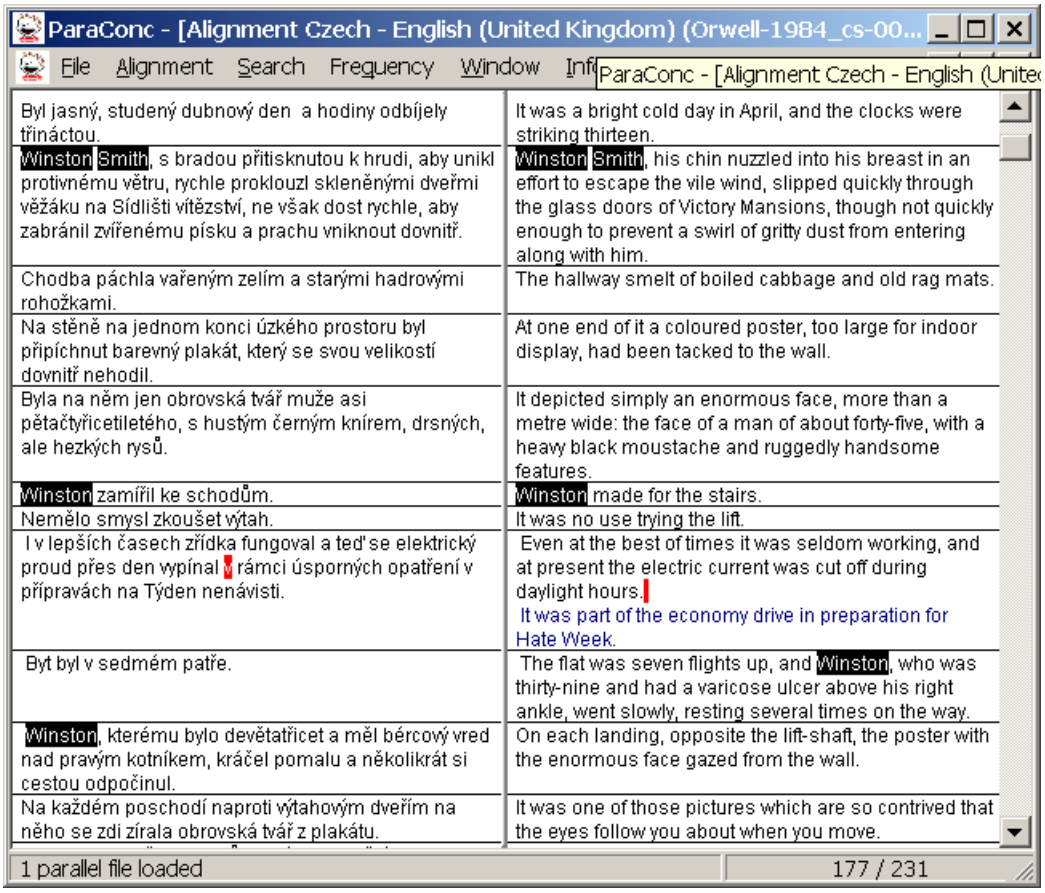
5. Pokud klepnete v seznamu verzí na některý identifikátor knihy, zobrazí se detaily dané verze spolu s detaily se zpřístupní v menu opět několik nových odkazů. všechny verze textu - zobrazí opět seznam verzí pro danou knihu. A dále, pokud jste osoba odpovědná za daný text, se zobrazí možnosti editovat záznam a smazat záznam.

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 celého životního cyklu textu v rámci projektu InterCorp spolu s tím, jak by měly být jednotlivé fáze práce zaznamenány v databázi.

1. Nejprve si v databázi ověřte, zda už není česká verze textu k dispozici.

2. Pokud student dostane za úkol naskenovat koordinátorem vybrané texty, nebo se koordinátor rozhodne nějaký text skenovat sám, měl by být text zapsán do databáze společně s poznámkou, že text není k dispozici. Stejný stav je možné použít jak u českého, tak u cizojazyčného textu. Tento stav je důležitý zvlášť u české verze, protože je možné, že ve stejné době budou dva týmy zbytečně shánět stejnou knihu.



Obrázek 8: Po spojení segmentů (vět) na anglické straně

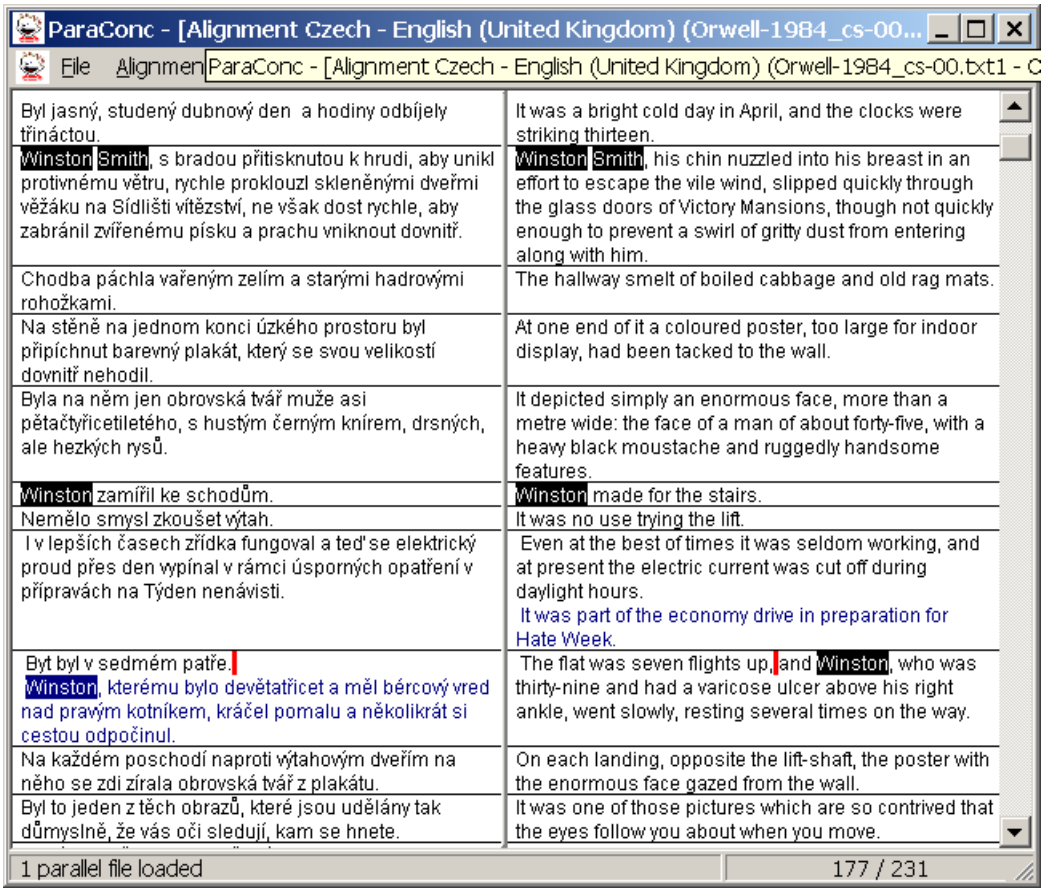
3. Poté, co daný text získáte v knižní nebo jiné papírově podobě, je potřeba doplnit v databázi bibliografické údaje a změnit stav knihy na: k dispozici v papírové (knižní) podobě (v papírové formě). Tím sdělíte ostatním účastníkům projektu, kteří by mohli chtít začít pracovat na stejném textu, že na skenování se už pracuje.

4. Až student odevzdá naskenovaný text, koordinátor zkontroluje cizojazyčnou verzi. (Předpokládá se zhruba: pečlivé přečtení několika náhodně vybraných odstavců a letmá kontrola celého textu, aby se určila míra chybovosti. Při velkém počtu chyb vrátit text studentovi k opakované korektuře.) Po této kontrole změní koordinátor v databázi stav cizojazyčného textu na stav: v elektronické podobě. Tím je dáno na srozuměnou, že text je už k dispozici a že koordinátor zkontroloval kvalitu naskenování cizojazyčné verze – na základě toho bude vyplácena odměna za naskenování cizího textu studentům!

5. Koordinátor by měl zkontrolovat i českou verzi, pokud ji student skenoval sám a nebyla vydána ze zdrojů ÚČNK. Provedení této kontroly zaznamená koordinátor stejným způsobem v databázi u záznamu o české verzi (stav: v elektronické podobě).

6. V této době je potřeba zaslat text do ÚČNK, kde bude provedena segmentace a označení vět. Dokončení této operace je označeno změnou stavu na: text je označkován. Stejně bude označen text, pokud je vydán z archivu ÚČNK.

7. Koordinátor musí zkontrolovat i zarovnání obou textů po větách v programu ParaConc. Po kontrole zarovnání změní koordinátor stav u cizojazyčné verze textu na: zarovnán po větách (zkontrolováno).



Obrázek 9: Po spojení segmentů (vět) na české straně

Přílohy

A Co dělá makro ICorpExport

- **Úpravy nutné pro zpracování v Paraconku** – změny, které je nutné do textu zanést, aby ho bylo možné korektně zobrazovat v programu Paraconc:

  1. Převedení do podoby holého textu v jednobytovém kódování, které je schopné ParaConc zobrazit v závislosti na volbě znakové sady (při načtení do ParaConku se znaková sada volí pro jednotlivé jazyky v položce Skript).
  2. Nahrazení některých znaků, které je třeba vyhradit pro značkování (<,>, &).

     ♣ Znaky, které v dané sadě nemají svůj kód, se převádějí do podoby nezávislé na použitém kódu – do takzvaných *znakových entit*, řetězců začínajících znakem `&` a končících `;`. Např. `&agrave;` reprezentuje písmeno à. Seznam znakových entit najdete třeba zde: `http://www.evolt.org/article/ala/17/21234/`.
  3. Další úpravy usnadňující další práci s textem (např. nahrazení tří teček znakem „výpustek“, 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, tučné písmo atd.)

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

B Instalace makra ICorpExport

Instalace potřebných komponent má dvě části:

1. Instalace Java Runtime Enviroment (JRE) – balíku utilit nutných k běhu programů napsaných v jazyce *java*, ve kterém je napsána větší část makra
2. Instalace makra pro Word (Word 2000 a pozdější; je možné, že starší verze balíku MS Office bude nutné upgradovat — není ověřeno).

B.1 Je nutné instalovat javu?

Prostředí JRE je mimo jiné nutné i ke správnému zobrazování některých webových stránek. Může se tedy stát, že správná verze JRE je na Vašem počítači už nainstalována. Zjistíte to takto:

1. Spusťte příkazový řádek:
   Start→Programy→Příslušenství→Příkazový řádek.

   ⟹ Nenajdete-li příkazový řádek v nabídce Start, klepněte na Start→ Spustit... a do políčka Otevřít: zadejte tento řetězec: `C:\WINDOWS\system32\cmd.exe`. Totéž lze provést tak, že poklepáte na soubor `cmd.exe` ve složce `C:\WINDOWS\system32`.
2. V příkazovém řádku zadejte `java -version`.
3. Počítač může odpovědět dvěma způsoby:
   (a) Pokud je java na Vašem počítači nainstalovaná, zobrazí se výpis s verzí javy a JRE. Pro běh makra je nutná verze JRE 1.5 (používá se však i označení 5.0), na verzi update nebo build-u nezáleží (označení build 1.5.0_02-b09 znamená: verze 1.5.0, update 02, build 09).
   (b) V opačném případě vypíše počítač zprávu o tom, že žádný program jménem „java“ nezná.
4. Máte-li na počítači JRE verze 1.5, můžete následující část přeskočit a rovnou si nainstalovat makro – viz část B.3. Jinak je třeba nejprve provést instalaci JRE – viz následující část B.2.

B.2 Instalace javy

1. Instalaci lze provést jen tehdy, máte-li k systému na vašem počítači administrátorská práva. Jinak se musíte obrátit na správce systému, na FF je to Laboratoř výpočetní techniky `http://www.ff.cuni.cz/lvt/`.
2. JRE lze stáhnout volně ze stránek `java.sun.com`:
   (a) Vpravo si pod Popular Downloads vyberte Java SE.
   (b) Najděte si Java Runtime Environment (JRE) a klepněte na Download.
   (c) Jako Platform zvolte Windows, zaškrtněte I agree a klepněte na Continue.
   (d) Můžete zvolit třeba Windows Online Installation. Klepněte na stažený soubor `jre-...` a postupujte podle instrukcí.
3. Po instalaci zkusíme funkčnost opět pomocí povelu `java -version` v příkazovém řádku. Pokud se zobrazí verze javy, instalace proběhla úspěšně a dále můžete pokračovat instalací makra – viz následující část B.3.

4. Pokud se i po instalaci na příkazovém řádku objevuje zpráva, že počítač program java nezná, je pravděpodobně nutné nastavit proměnné systému PATH a CLASSPATH. Postup se může lišit v jednotlivých operačních systémech. Následující postup platí pro nastavení těchto proměnných v prostředí Windows 2000:

   (a) Vyvoláme tabulku s vlastnostmi systému: pravým tlačítkem myši klepneme na Tento počítač, pak levým na položku Vlastnosti.

   (b) Zvolíme záložku Upřesnit, klepneme na tlačítko Proměnné prostředí....

   (c) Najdeme-li v dolním okně Systémové proměnné proměnnou Path, klepneme na tlačítko Upravit.... Proměnná Path již bude nějaké cesty obsahovat. Cesty se v poli hodnota oddělují středníkem, najdeme tedy poslední cestu a na její konec přidáme středník a cestu k adresáři se spustitelnými příkazy javy – viz bod 4e.

   (d) Pokud v dolním okně Systémové proměnné proměnná Path chybí, klepneme místo na Upravit... na tlačítko Nová... a zadáme název proměnné Path).

   (e) Do hodnoty proměnné Path uvedeme cestu k adresáři se spustitelnými příkazy javy. Při běžné instalaci by měla vypadat asi takto: `C:\Program Files\Java\jdk1.5.0_<cislo updatu>\bin`.

   (f) Klepneme na OK. Tím je cesta nastavena a my můžeme spouštět javu bez nutnosti zadávat vždy celou cestu ke spustitelnému souboru.

B.3 Instalace makra

1. Stáhněte si adresář zabalený pomocí programu WinZip z adresy `http://korpus.cz/intercorp/files/CharConvert.zip` a rozbalte si ho do adresáře Program Files. Měli byste vidět adresář C:\Program Files\CharConvert\ .

2. Nyní spusťte Příkazový řádek (Start→Programy→Příslušenství→Příkazový řádek). V příkazovém řádku zadejte postupně:

   ```
   C:
   cd \program files\charconvert
   java SystemCheck
   ```

   Program by měl vypsat zhruba následující hlášení (verze programu nebo systému se mohou lišit):

   ```
   C:\Program Files\CharConvert>java SystemCheck line.separator:

   file.separator: \
   java.version: 1.5.0_02
   java.vendor: Sun Microsystems Inc.
   java.home: C:\Program Files\Java\jre1.5.0_02
   java.vm.version: 1.5.0_02-b09
   java.vm.vendor: Sun Microsystems Inc.
   java.vm.name: Java HotSpot(TM) Client VM
   os.name: Windows 2000
   os.arch: x86
   os.version: 5.0
   ```

   Pokud se tak skutečně stalo, je všechno v pořádku a můžete přejít k bodu 5. Pokud se to nepořilo, nejprve si zkontrolujte podle bodu 2 předešlý postup (při psaní příkazu `java SystemCheck` je důležité zachovat velká a malá písmena).

3. Pokud se stále vypisuje chybové hlášení, je pravděpodobně nutné nastavit proměnnou systému CLASSPATH. Postup je stejný jako při nastavení proměnné Path (viz část B.2, bod 4).

$\Longrightarrow$ Proměnná CLASSPATH pravděpodobně mezi systémovými proměnnými Vašeho počítače zatím nebude. Je tedy nutné zvolit tlačítko Přidat.... Název proměnné zadejte CLASSPATH a hodnotu pouze tečku (.). Tím počítači řeknete, že když se pokoušíte spustit program pomocí javy, má hledat jeho zdrojové kódy v aktuálním adresáři.

4. Zkuste opět spustit program SystemCheck (viz bod 2), pokud jste ani teď nedošli k žádoucímu výsledku, kontaktujte svého administrátora nebo hlavního koordinátora projektu (`martin.vavrin@ff.cuni.cz`). Pokud program SystemCheck pracuje správně, máte ověřeno, že java je na Vašem počítači správně nainstalována.
5. Nyní zkopírujte soubor ICorpExport.dot z adresáře c:\Program Files\CharConvert do adresáře C:\Program Files\Microsoft Office\Office\Startup.
6. Spusťte Word. V menu Nástroje→Šablony a doplňky by se Vám měla zobrazit nová šablona ICorpExport.dot, pokud není zaškrtnuté políčko před jejím názvem, zaškrtněte ho. Tím jsou potřebná makra připravena k použití.
7. Makra lze spouštět buď pomocí menu Nástroje→Makro→Makra, nebo si vytvořit tlačítka v nástrojové liště pro pravidelné a snadné použití:

Aby se nastavení tlačítek správně uložilo, je třeba mít během vytváření tlačítek otevřené pouze jedno okno Word-u, v tom změnu provést a po dokončení úpravy Word ukončit. Tím se uloží nastavení nástrojové lišty a Word si bude nově vytvořená tlačítka pamatovat při příštím spuštění.

(a) Pravým tlačítkem myši klepněte kamkoli na panel nástrojů a v kontextovém menu zvolte Vlastní....
(b) Nejprve zvolte záložku Panely nástrojů, klepněte na tlačítko Nový. Název zvolte dle vlastního uvážení, např.: Makra a potvrďte tlačítkem OK. Vytvoří se Vám nový panel nástrojů, který můžete buď umístit mezi ostatní do lišty, nebo nechat volně jako samostatné okno.
(c) Přepněte na záložku Příkazy a ve sloupci Kategorie zvolte položku Makra. Ve sloupci Příkazy byste teď měli vidět mimo jiné i položky TemplateProject.NewMacros.CheckParagraphs a TemplateProject.NewMacros.ICorpExport.
(d) Přetáhněte myší obě tlačítka na panel, který jste si před chvílí vytvořili.
(e) Nechte okno Vlastní otevřené a klepněte pravým tlačítkem myši na nová tlačítka, která jste vytvořili.
(f) V menu, které se zobrazí, můžete zkrátit název tlačítek na ICorpExport a CheckParagraphs, nebo si zvolit vlastní názvy. Dále si v tomto menu můžete přidat k tlačítku ikonu, nebo před tlačítko přidat oddělovací čáru zaškrtnutím volby Začátek skupiny.
(g) Aby se změny uložily, program Word ukončete.

C 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 (včetně XP)
  - 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).

D Tabulky

Tabulka 1: Český a anglický text před zarovnáním

Český text	**Anglický text**
<s id="1.1">Byl jasný, studený dubnový den a hodiny odbíjely třináctou. </s>	<s id="1.1">It was a bright cold day in April, and the clocks were striking thirteen. </s>
<s id="1.2">Winston Smith, s bradou přitisknutou k hrudi, aby unikl protivnému větru, rychle proklouzl skleněnými dveřmi věžáku na Sídlišti vítězství, ne však dost rychle, aby zabránil zvířenému písku a prachu vniknout dovnitř. </s>	<s id="1.2">Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though not quickly enough to prevent a swirl of gritty dust from entering along with him. </s>
<s id="2.1">Chodba páchla vařeným zelím a starými hadrovými rohožkami. </s>	<s id="2.1">The hallway smelt of boiled cabbage and old rag mats. </s>
<s id="2.2">Na stěně na jednom konci úzkého prostoru byl připíchnut barevný plakát, který se svou velikostí dovnitř nehodil. </s>	<s id="2.2">At one end of it a coloured poster, too large for indoor display, had been tacked to the wall. </s>
<s id="2.3">Byla na něm jen obrovská tvář muže asi pětačtyřicetiletého, s hustým černým knírem, drsných, ale hezkých rysů. </s>	<s id="2.3">It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features. </s>
<s id="2.4">Winston zamířil ke schodům. </s>	<s id="2.4">Winston made for the stairs. </s>
<s id="2.5">Nemělo smysl zkoušet výtah. </s>	<s id="2.5">It was no use trying the lift. </s>
<s id="2.6"> I v lepších časech zřídka fungoval a teď se elektrický proud přes den vypínal v rámci úsporných opatření v přípravách na Týden nenávisti. </s>	<s id="2.6"> Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours. </s>
<s id="2.7"> Byt byl v sedmém patře. </s>	<s id="2.7"> It was part of the economy drive in preparation for Hate Week. </s>
<s id="2.8"> Winston, kterému bylo devětatřicet a měl bércový vřed nad pravým kotníkem, kráčel pomalu a několikrát si cestou odpočinul. </s>	<s id="2.8"> The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way. </s>
<s id="2.9">Na každém poschodí naproti výtahovým dveřím na něho se zdi zírala obrovská tvář z plakátů. </s>	<s id="2.9">On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall. </s>
<s id="2.10">Byl to jeden z těch obrazů, které jsou udělány tak důmyslně, že vás oči sledují, kam se hnete. </s>	<s id="2.10">It was one of those pictures which are so contrived that the eyes follow you about when you move. </s>
<s id="2.11">Velký bratr tě sleduje, zněl nápis pod obrazem. </s>	<s id="2.11">Big Brother is watching you, the caption beneath it ran. </s>

Tabulka 2: Chybně zarovnaný český a anglický text

Český text	**Anglický text**
<seg id="1"> <s id="1.1">Byl jasný, studený dubnový den a hodiny odbíjely třináctou.</seg> <seg id="2"></s> <s id="1.2">Winston Smith, s bradou přitisknutou k hrudi, aby unikl protivnému větru, rychle proklouzl skleněnými dveřmi věžáku na Sídlišti vítězství, ne však dost rychle, aby zabránil zvířenému písku a prachu vniknout dovnitř.</seg> </s> <seg id="3"> <s id="2.1">Chodba páchla vařeným zelím a starými hadrovými rohožkami.</seg> <seg id="4"></s> <s id="2.2">Na stěně na jednom konci úzkého prostoru byl připíchnut barevný plakát, který se svou velikostí dovnitř nehodil.</seg> <seg id="5"></s> <s id="2.3">Byla na něm jen obrovská tvář muže asi pětačtyřicetiletého, s hustým černým knírem, drsných, ale hezkých rysů.</seg> <seg id="6"></s> <s id="2.4">Winston zamířil ke schodům.</seg> <seg id="7"></s> <s id="2.5">Nemělo smysl zkoušet výtah.</seg> <seg id="8"> </s> <s id="2.6">I v lepších časech zřídka fungoval a teď se elektrický proud přes den vypínal v rámci úsporných opatření v přípravách na Týden nenávisti. </seg> <seg id="9"></s> <s id="2.7">Byt byl v sedmém patře. </seg> <seg id="10"></s> <s id="2.8">Winston, kterému bylo devětatřicet a měl bércový vřed nad pravým kotníkem, kráčel pomalu a několikrát si cestou odpočinul. </seg> <seg id="11"></s> <s id="2.9">Na každém poschodí naproti výtahovým dveřím na něho se zdi zírala obrovská tvář z plakátů.</seg> <seg id="12"></s> <s id="2.10">Byl to jeden z těch obrazů, které jsou udělány tak důmyslně, že vás oči sledují, kam se hnete.</seg> <seg id="13"></s> <s id="2.11">Velký bratr tě sleduje, zněl nápis pod obrazem.</seg> </s> 	<seg id="1"> <s id="1.1">It was a bright cold day in April, and the clocks were striking thirteen.</seg> <seg id="2"></s> <s id="1.2">Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though not quickly enough to prevent a swirl of gritty dust from entering along with him.</seg> </s> <seg id="3"> <s id="2.1">The hallway smelt of boiled cabbage and old rag mats.</seg> <seg id="4"></s> <s id="2.2">At one end of it a coloured poster, too large for indoor display, had been tacked to the wall.</seg> <seg id="5"></s> <s id="2.3">It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features.</seg> <seg id="6"></s> <s id="2.4">Winston made for the stairs.</seg> <seg id="7"></s> <s id="2.5">It was no use trying the lift.</seg> <seg id="8"> </s> <s id="2.6">Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours. </seg> <seg id="9"></s> <s id="2.7">It was part of the economy drive in preparation for Hate Week. </seg> <seg id="10"></s> <s id="2.8">The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way. </seg> <seg id="11"></s> <s id="2.9">On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall.</seg> <seg id="12"></s> <s id="2.10">It was one of those pictures which are so contrived that the eyes follow you about when you move.</seg> <seg id="13"></s> <s id="2.11">Big Brother is watching you, the caption beneath it ran.</seg> </s>

Tabulka 3: Správně zarovnaný český a anglický text

Český text	**Anglický text**
<seg id="1"> <s id="1.1">Byl jasný, studený dubnový den a hodiny odbíjely třináctou.</seg> <seg id="2"></s> <s id="1.2">Winston Smith, s bradou přitisknutou k hrudi, aby unikl protivnému větru, rychle proklouzl skleněnými dveřmi věžáku na Sídlišti vítězství, ne však dost rychle, aby zabránil zvířenému písku a prachu vniknout dovnitř.</seg> </s> <seg id="3"> <s id="2.1">Chodba páchla vařeným zelím a starými hadrovými rohožkami.</seg> <seg id="4"></s> <s id="2.2">Na stěně na jednom konci úzkého prostoru byl připíchnut barevný plakát, který se svou velikostí dovnitř nehodil.</seg> <seg id="5"></s> <s id="2.3">Byla na něm jen obrovská tvář muže asi pětačtyřicetiletého, s hustým černým knírem, drsných, ale hezkých rysů.</seg> <seg id="6"></s> <s id="2.4">Winston zamířil ke schodům.</seg> <seg id="7"></s> <s id="2.5">Nemělo smysl zkoušet výtah.</seg> <seg id="8"> </s> <s id="2.6">I v lepších časech zřídka fungoval a teď se elektrický proud přes den vypínal v rámci úsporných opatření v přípravách na Týden nenávisti. </seg> <seg id="9"> </s> <s id="2.7">Byt byl v sedmém patře. </s> <s id="2.8">Winston, kterému bylo devětatřicet a měl bércový vřed nad pravým kotníkem, kráčel pomalu a několikrát si cestou odpočinul. </seg> <seg id="10"></s> <s id="2.9">Na každém poschodí naproti výtahovým dveřím na něho se zdi zírala obrovská tvář z plakátů.</seg> <seg id="11"></s> <s id="2.10">Byl to jeden z těch obrazů, které jsou udělány tak důmyslně, že vás oči sledují, kam se hnete.</seg> <seg id="12"></s> <s id="2.11">Velký bratr tě sleduje, zněl nápis pod obrazem.</seg> </s> 	<seg id="1"> <s id="1.1">It was a bright cold day in April, and the clocks were striking thirteen.</seg> <seg id="2"></s> <s id="1.2">Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though not quickly enough to prevent a swirl of gritty dust from entering along with him.</seg> </s> <seg id="3"> <s id="2.1">The hallway smelt of boiled cabbage and old rag mats.</seg> <seg id="4"></s> <s id="2.2">At one end of it a coloured poster, too large for indoor display, had been tacked to the wall.</seg> <seg id="5"></s> <s id="2.3">It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features.</seg> <seg id="6"></s> <s id="2.4">Winston made for the stairs.</seg> <seg id="7"></s> <s id="2.5">It was no use trying the lift.</seg> <seg id="8"> </s> <s id="2.6">Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours. </s> <s id="2.7">It was part of the economy drive in preparation for Hate Week. </seg> <seg id="9"> </s> <s id="2.8">The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way. </seg> <seg id="10"></s> <s id="2.9">On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall.</seg> <seg id="11"></s> <s id="2.10">It was one of those pictures which are so contrived that the eyes follow you about when you move.</seg> <seg id="12"></s> <s id="2.11">Big Brother is watching you, the caption beneath it ran.</seg> </s>