Jak na paralelní texty s programem ParaConc

verze 0.3

Alexandr Rosen*
alexandr.rosen@ff.cuni.cz

14. dubna 2005

1 ParaConc – základní údaje

- 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`

2 Instalace

Předpoklady: 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.

Soubor o velikosti asi 1,4 MB zkopírujeme kamkoli na disk (nejlépe do složky *Program Files*, se zástupcem na ploše).

*S poděkováním Martinu Sváškovi za cenné připomínky.

3 Můj první paralelní korpus

Následuje návod, jak programem ParaConc vytvořit paralelní korpus co nejjednodušším způsobem. Postup předpokládá jednu z typických situací: máme k dispozici dva soubory ve formátu MS Word (text a jeho překlad) a pracujeme na počítači se systémem MS Windows (ověřeno pro verzi XP) a s editorem MS Word (ověřeno pro verze 2000, 2002, 2003).

3.1 Kontrola formátování

Některé texty nejsou v takovém formátu, aby je bylo možné v programu ParaConc bez úprav použít. To se týká zejména nevhodně umístěných znaků konce odstavce, tabelátorů, mezer apod.

Znak konce odstavce by měl v souborech oddělovat odstavce, nadpisy, položky seznamů apod. Může oddělovat i jednotlivé věty.[1] Neměl by ale oddělovat jednotlivé řádky. Chybně formátovaný text, kde znak konce odstavce leží uprostřed věty, je většinou výsledkem převodu textu z jiného formátu, který není určen pro další zpracování (pdf, HTML).

Stejně tak by se uprostřed věty neměly nacházet tabelátory a mezery.

Postup:

1. Otevřeme soubor v programu MS Word.

2. V nabídce na horní liště klepneme levým tlačítkem myši na Nástroje, pak na Možnosti. Vybereme kartu Zobrazení a v části Značky formátování zaškrtneme políčka Znaky tabulátoru a Konce odstavců. Výběr potvrdíme tlačítkem OK.

3. Zkontrolujeme, zda znak konce odstavce, zobrazený jako symbol ¶, není uprostřed věty. Může stát pouze na konci věty, odstavce, nadpisu, položky výčtu nebo na prázdném řádku.

4. Podobně zkontrolujeme, zda někde uprostřed věty nestojí znak tabulátoru, zobrazený jako →. Slova by od sebe neměla být oddělena víc než jednou mezerou.

5. Nevhodně umístěné znaky lze odstranit pomocí volby Úpravy/Nahradit/Více/Formát/Speciální/....

[1] Pokud nechceme dělení na odstavce ignorovat, je třeba pak odstavce vyznačit jiným způsobem, např. značkami.

3.2 Konverze do textového formátu

Z formátu MS Word (.doc) musíme oba soubory převést do textového formátu (.txt) v kódování Unicode UTF-8.

Postup:

1. Otevřeme soubor v programu MS Word.
2. Klepneme na Soubor/Uložit jako.
3. V okénku Typ souboru vybereme možnost, která závisí na verzi editoru Word:
   - Word 2000: Kódovaný text (*.txt)
   - Word 2002/2003: Prostý text (*.txt) a Jiné kódování
4. Klepneme na Uložit.
5. Na dotaz „Styly, obrázky a jiné formátování nelze uložit jako Kódovaný text. Chcete soubor ... v tomto formátu uložit?“ odpovíme klepnutím na Ano.
6. Objeví se okno Převod souboru. Na výzvu Zvolte kódování pro uložení tohoto souboru zareagujeme zaškrtnutím možnosti Jiné kódování a v rámečku vpravo vybereme možnost Kódování Unicode (UTF-8). V rámečku Náhled: by se měl objevit text se správnými znaky.
7. Klepneme na tlačítko OK.

3.3 Označení struktury textu

Před načtením textů do programu ParaConc je často vhodné v nich označit hranice sekcí (kapitol, oddílů), odstavců a vět. Toto označení se pak zachová i v souborech, které z programu ParaConc exportujeme.

K označení hranic mezi úseky textu (zejména odstavci a větami) lze použít různé programové nástroje. **!!! sem doplnit doporučení**

Jsou-li hranice mezi úseky textu vyznačené jen znaky konce odstavce (¶), při exportu (File/Export Corpus Files) je třeba zvolit možnost Alignment Style: Tags. Jinak se informace o struktuře textu ztratí.

3.4 Načtení textů do programu ParaConc

Předpokládáme-li, že paralelní texty jsou už zarovnány po odstavcích nebo po větách pomocí znaků „konec odstavce“, je vhodné před jejich načtením ověřit, že obsahují stejný počet odstavců. Lze tak učinit například ve Wordu pomocí volby Nástroje/Počet slov (zobrazí se i údaj o počtu odstavců).

Zvolíme-li při načítání možnost, že soubory už jsou zarovnané, ParaConc nás na nestejný počet odstavců (případně vět) neupozorní. Zvolíme-li při načítání možnost, že soubory zarovnané nejsou, ParaConc při nestejném počtu odstavců (vět) oznámí chybu. V obou případech máme možnost dělení na odstavce (věty) v ParaConku opravit.

1. Spustíme program ParaConc.
2. Klepneme na File/Load Corpus File(s).
3. Objeví se okno Load Corpus Files.
4. Počet paralelních textů v okně Parallel texts ponecháme na hodnotě 2.
5. Nejprve vybereme parametry prvního souboru:
   (a) Nastavíme **jazyk**. Pokud příslušný jazyk mezi nabízenými možnostmi nenajdeme, je třeba příslušné národní prostředí do systému doinstalovat. (Vložíme instalační CD systému Windows do mechaniky, klepneme na Start/Nastavení/Ovládací panely/Místní nastavení a dále postupujeme podle pokynů.)
   (b) Po klepnutí na tlačítko Font vybereme písmo. Vhodné písmo může být např. Arial. Je velmi důležité zvolit správný Script. Např. pro západoevropské jazyky zvolíme Western, pro češtinu Central European. Není-li v nabídce vhodný skript, zvolíme jiné písmo a tento krok zkusíme znovu.
   (c) Volba Format závisí na formátu zarovnávání (Align format).
      i. Jsou-li texty už zarovnané (Align format: New line delimiter, Delimiter nebo Start/stop tags), stačí po klepnutí na tlačítko Format zadat pouze způsob rozpoznání hranic mezi větami: Při volbě Automatic recognition se konec věty určuje na základě interpunkce; při volbě HTML/SGML Markers se předpokládá, že každá věta je v textu vyznačena značkami, např. takto:

```
<s>Toto je první věta.</s>
<s>Toto je druhá věta.</s>.
```

Do políčka Start tag pak zapíšeme s, do políčka Stop tag zapíšeme /s.

ii. Pokud texty zarovnané nejsou (Align format: Not aligned), je třeba po klepnutí na tlačítko Format zadat způsob, jak rozpoznat hranice mezi většími úseku textu – sekce (kapitoly, oddíly), odstavci i větami. U kratších textů lze ponechat nastavení Headings: HTML/SGML Markers s nevyplněnými políčky Start tag a Stop tag (text se pak na sekce nedělí, celý se považuje za jedinou sekci). Jsou-li odstavce oddělené znakem konce odstavce, ponecháme Paragraphs: New Line Delimited. Určení způsobu oddělování vět (Sentences) je popsáno výše v bodě 5(c)i.

(d) Klepneme na tlačítko Add a vybereme správný soubor.

(e) Klepneme na jméno souboru v okně Load Corpus Files a klepnutím zaškrtneme UTF-8 (soubor je ve formátu Unicode UTF-8).

6. Body 5a až 5e zopakujeme pro druhý soubor.

7. Klepneme na tlačítko OK.

3.5 Úpravy segmentace a zarovnání

Po klepnutí na tlačítko OK v okně Load Corpus Files se může objevit chybové hlášení o tom, že počet sekcí nebo odstavců se v obou textech liší. V takovém případě soubory nelze zarovnat a chybu je třeba opravit.[2]

Postup

1. V okně Error klepneme na tlačítko Fix.

2. Objeví se dvě tabulky o dvou sloupcích. Jedna udává členění textů na sekce, druhá na odstavce. Je-li chyba v různém počtu odstavců, je navrchu tabulka s odstavci v příslušné sekci. Skládá-li se odstavec z více vět, jsou tyto věty odlišeny barevně. (Dělení na věty nemusí být vždy správné, je to výsledek dělení odstavce na věty podle zadaných kritérií.)

3. Najdeme v tabulce místo, kde na jedné straně text končí (zbývá prázdné místo), zatímco druhý sloupec pokračuje dalšími větami nebo odstavci.

[2] Poznámka: Pokud jsme při načítání souborů uvedli, že jsou již zarovnané (Align format: New line delimiter, Delimiter nebo Start/stop tags), počet sekcí a odstavců se nekontroluje a rovnou se zobrazí tabulka odpovídající zarovnaným souborům. Buňky tabulky („segmenty“) i předpokládané věty v rámci buněk lze rozdělovat a spojovat, ale věty už nelze zarovnávat automaticky (volba File/Align corpus není dostupná.)

Klepneme pravým tlačítkem myši na první písmeno „přebývajícího textu“ a z místní nabídky vybereme Split paragraph (nebo Split current section, opravujeme-li dělení na sekce). Totéž opakujeme tak dlouho, až si odstavce (sekce) navzájem odpovídají.

4. Okna s tabulkami zavřeme.

5. Klepneme na File/Align Corpus.

6. Klepneme na File/View Corpus Alignment.

7. Objeví se okno Select Files to View. Klepneme na soubory, které se mají zobrazit, a pak na tlačítko Alignment.

8. Opět se objeví dvě okna, tentokrát lze v okně Alignment prohlížet a opravovat zarovnání nejen na odstavce (Alignment/Paragraphs), ale i na věty (Alignment/Aligned Sentences). (Nabídka Alignment se objeví na horní liště po klepnutí na okno Alignment.

9. Zarovnané věty lze rozdělovat nebo spojovat po klepnutí pravým tlačítkem myši do příslušného pole tabulky a volbě možnosti Split segment (věta se rozdělí v místě kurzoru) nebo Merge with Next Segment, případně Merge with Previous Segment. Jiným způsobem text upravovat nelze.

10. Zarovnanou tabulku zavřeme a práci si uložíme: File/Save Workspace. Příště už nemusíme soubory znovu načítat a zarovnávat, ale stačí uložený korpus otevřít (File/Open Workspace...).

3.6 Export textů

Paraconc obsahuje funkci pro export textů korpusu: File/Export Corpus Files. V současné verzi (269) se texty ukládají v kódování ANSI, kódování UTF-8 nelze použít.

Při exportu textů, v nichž jsou odstavce, případně i věty, oddělovány znakem konce odstavce, je vhodné zvolit možnost Alignment Style: Tags, jinak se informace o struktuře textu ztratí.

Exportované texty lze do programu ParaConc znovu načíst jako zarovnané.