Projekt *InterCorp* – jak přidávat nové texty*

4. května 2005

Obsah

*Sepsal Alexandr Rosen s vydatnou pomocí Michala Křena a Pavla Vondřičky.

1 Varování

- Tento postup zatím nelze provést celý, protože některé nástroje nejsou ještě k dispozici.
- Platí pro texty, které ještě nebyly zarovnány.
- Texty, které již zarovnány byly, je třeba zaevidovat (viz níže bod 8-3j[1]). Tyto texty budou zpracovány odlišně po dohodě s ÚČNK.
- **Zarovnávání nových textů doporučujeme odložit do doby, kdy bude možné texty zpracovat podle tohoto postupu.** Snažíme se, aby to netrvalo dlouho.
- **V získávání textů (včetně skenování a OCR) lze samozřejmě pokračovat.**
- Připomínky všeho druhu uvítáme na adrese alexandr.rosen(zavinac)ff.cuni.cz, nebo intercorp(zavinac)ff.cuni.cz.

[1] tedy bod 3j v části 8

2 Získání textu

Text lze získat různými způsoby (v pořadí podle rostoucí obtížnosti zpracování):

1. ze zdrojů ÚČNK
2. elektronicky z externího zdroje
3. skenováním z papírové předlohy s pomocí programu na rozpoznávání znaků (např. FineReader)

3 Konverze do textového formátu

Texty z ÚČNK (bod 2-1) už budou připravené k zarovnávání. Texty získané jinak (body 2-2, 2-3) je třeba do takové podoby převést.

1. Je-li text (naskenovaný nebo jinak získaný) ve formátu editoru MS Word, bude možné využít makro, které provede export do textové podoby v kódování UTF-8, případně označkuje sekce, hranice odstavců a vět, kurzívu, tučné písmo apod.
2. Bude-li původní text v jiném formátu, je nutné ho upravit do textové podoby v kódování UTF-8, případně označkovat sekce, hranice odstavců a vět, kurzívu, tučné písmo apod. jiným způsobem.
3. Český text je třeba po konverzi v každém případě předat ÚČNK k registraci a označení hranic odstavců a vět (část 4). Až pak se může zarovnávat s cizojazyčným textem.
4. Cizojazyčný text lze po konverzi zarovnávat rovnou.

4 Příprava českých textů

Český text může vystupovat ve dvojici s více jazyky. Abychom si ušetřili práci s jeho přípravou a případně ho mohli později využít jako mezičlánek mezi dvěma nebo více cizími jazyky, je třeba dbát na to, aby text měl ve všech dvojicích jedinou neměnnou podobu.

1. V českých textech se před zarovnáním s cizojazyčnými texty vyznačí značkami hranice odstavců (`<p id=...>`) a vět (`<s id=...>`), případně sekcí (`<div id=...>`).

2. To provede ÚČNK pro všechny české texty, které má k dispozici, a také pro všechny další texty, které účastníci získají.

3. Jsou-li příslušné údaje k dispozici, v textech se vyznačí značkami také řezy písma (tučné, kurzíva apod.).

4. Takto zpracovaný český text předá ÚČNK katedře.[2]

5 Příprava cizojazyčných textů

1. Příprava cizojazyčných textů je úkolem kateder.

2. V cizojazyčných textech nemusí být před zarovnáním s českými vyznačeny hranice odstavců a vět.

3. Jsou-li v českém textu vyznačeny hranice sekcí, je pro zarovnávání výhodné, aby byly také v textu cizojazyčném (`<div id=...>`).

4. Pro cizojazyčné texty platí výše uvedené body 3-1 a 3-2.

6 Zarovnávání

1. Katedry provedou zarovnání cizího textu s českým textem, který dodá ÚČNK (viz část 4). V českém textu se při zarovnávání nesmí dělit věty ani odstavce.

2. Minimálním požadavkem je zarovnání po odstavcích. Je možné provést automatické zarovnání po větách (v programu ParaConc bez nutnosti dalších úkonů). Automatické zarovnání po větách lze zkontrolovat a opravit.

3. Při zarovnávání v programu ParaConc se do programu načte český text s označenými hranicemi odstavců, vět (případně sekcí) spolu s paralelním cizojazyčným textem (kde tyto hranice vyznačeny být nemusí).

[2] *Katedrou* je v tomto textu myšleno řešitelské pracoviště pro daný jazyk.

4. Při zarovnávání v programu ParaConc je nutné zarovnané texty exportovat do formátu se značkami `<seg id=...>`.

5. Při zarovnávání jiným způsobem je třeba sladit formát vstupu a výstupu s navazujícími kroky.

6. Výsledek zarovnávání katedry odevzdají ÚČNK.

7 Zpracování zarovnaných textů

1. Zarovnané texty lze využívat pomocí programu ParaConc.

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. V případě potřeby bude možné dodatečně označit v textech jednotlivá slova (`<w id=...>`) pro podrobnější značkování nebo zarovnávání.

8 Evidence

1. ÚČNK a katedry vedou na webových stránkách projektu evidenci textů a postupu jejich zpracování.

2. U každého textu se uvedou jeho bibliografické údaje, odkaz na osobu, která za text odpovídá, a příznaky aktuálního stavu. Z těchto údajů se generuje hlavička podle TEI-XML.

3. Příznaky:

   (a) text je v papírové podobě

   (b) text je v elektronické podobě

   (c) text je zajištěn po právní stránce

   (d) v textu jsou značky pro hranice odstavců a vět

   (e) text je zarovnán s českým po odstavcích (jen u cizojazyčných textů)

   (f) text je zarovnán s českým po větách (bez kontroly) (jen u cizojazyčných textů)

(g) text je zarovnán s českým po větách a zarovnání po větách je zkontrolováno (jen u cizojazyčných textů)

(h) text je označkován podle TEI-XML

(i) zarovnání je označeno podle TEI-XML (jen u cizojazyčných textů)

(j) text je zarovnán s českým jinak než doporučeným postupem (český text není zaregistrován a připraven v ÚČNK)

4. Metatextové údaje se budou generovat z údajů o textu zadaných ve webovém rozhraní. Katedry budou pracovat s jednoduše označkovanými texty bez hlavičky.