Navodila za pripravo datotek korpusa SPOOK projekta Slovensko prevodoslovje – viri in raziskave

Tomaž Erjavec, 2009-08-16

Splošna navodila

1. Vsaka datoteka naj vsebuje eno enoto korpusa, t.j. bibliografske podatke o besedilu ter samo besedilo v originalu in prevodu v slovenski jezik.
2. Imena datotek naj vsebujejo samo znake a-z, A-Z, 0-9, _ in -. Predvsem naj ne vsebujejo šumnikov ali presledkov. Končnica datotek naj bo .xml
   Primer: `LMD001-Banlieues_Tissot.xml`
3. Zapis znakov v datotekah naj bo UTF-8.
4. Datoteke morajo biti zapisane v pravilnem formatu XML (več o tem spodaj). To preverite tako, da jih poskusite odpreti v Internet Explorer.
5. Sama besedila v datoteko XML vstavimo kar v izvornem formatu, razen prekodiranja v UTF-8.
6. Če je izvorno besedilo ni v XML, je potrebno predhodno še zamenjati vse »&« v besedilu z »&«, vse »<« pa z »<«.

Format XML

Spodaj podamo primer pravilno oblikovane datoteke. Obkrožene številke niso del datoteke, pač pa služijo za razlago spodaj.

```
1   <?xml version="1.0" encoding="utf-8"?>
2   <!DOCTYPE enota SYSTEM "http://nl.ijs.si/et/spook/spook.dtd">
3   <enota jezik="fra_slv">
4    <bibl>
5     <naslovIzvirnika jezik="fra">Comment la question sociale est dénaturée…</naslovIzvirnika>
6      <naslovPrevoda jezik="slv">Kako je družbeno vprašanje popačeno …</naslovPrevoda>
7      <avtor>Sylvie Tissot</avtor>
8      <prevajalec></prevajalec>
9      <datumIzvirnika>2007-10</datumIzvirnika>
10     <datumPrevoda>2007-10</datumPrevoda>
11     <zalozbaIzvirnika></zalozbaIzvirnika>
12      <zalozbaPrevoda></zalozbaPrevoda>
13      <cobissIzvirnika></cobissIzvirnika>
14      <cobissPrevoda></cobissPrevoda>
15     <pripravil>Adriana Mezeg</pripravil>
16     <datumObdelave>2009-07-18</datumObdelave>
```

```
17    </bibl>
18    <besedilo jezik="fra">
19    COMMENT LA QUESTION SOCIALE EST DENATUREE...
20    L’invention des “quartiers sensibles”
21    </besedilo>
22     <besedilo jezik="slv">
23    Kako je družbeno vprašanje popačeno ...
24    Iznajdba »občutljivih četrti«
25    </besedilo>
26    </enota>
```

- Vrstici 1,2 morata biti prisotni v vsaki datoteki točno tako kot sta napisani; služita za preverjanje pravilnosti oblikovanja datoteke.
- Vrstica 3 označi celo enoto (zaključi se v zadnji vrstici, torej 26), poda pa tudi jezika; slovenščina kot jezik prevoda je vedno druga.
- Oznake jezikov naj bodo po standardu ISO 639-3:
    - slv = slovenščina
    - ita = italijanščina
    - ger = nemščina
    - eng = angleščina
    - fra = francoščina
- Vrstice 4 – 17 zajemajo bibliografske podatke o enoti, ki morajo biti navedeni v podanem vrstnem redu. Če kateri od bibliografskih podatkov manjka (npr. COBISS) pustimo polje prazno, lahko ga pa tudi kar zbrišemo.
- Vrstici 15 in 16 zapišemo svoje ime in datum, kdaj smo besedilo pripravili.
- Vse datume zapišemo v obliki llll-mm-dd, npr. 1997, 2000-03, 2009-06-02; če besedilo ni izšlo (npr. v prevodu) potem naj bo datum 0000-00-00.
- Vrstice 18 – 21 vsebujejo izvirnik besedila, 22 – 25 pa prevod.
- Pomembno je, da je vsaka prevodna enota v svoji vrstici, in da se število vrstic originala ujema s številom vrstic prevoda; tako je vrstica 19 poravnana s 23, 20 pa s 24.
- Če je besedilo poravnano tako, da ima prevodna enota izvirnik in prevod v isti vrstici, ločena npr. s tabulatorjem, enoti vpišemo samo eno besedilo, npr. <besedilo jezik="fra_slv">