Student projects for course "Standards for digital encoding"

Tomaž Erjavec
2006-11-14

The project counts for 70% of the final grade, and is composed of the practical work + written report. The project work is to be presented and discussed at the last lecture (1.12.2006) and the report handed in by the end of the term, (1.2.2007) at the latest.

Projects can be done individually or in groups. If students prepare the work together, the end result should be proportionately more substantial.

The written report should be written as a standard ACL paper; the stylesheet with instructions is available at http://www.aclweb.org/acl2005/index.php?stylefiles The report should contain an introduction and explain the task undertaken. Esp. interesting is the discussion of problems encountered, and their solutions.

Students are free to come up with their project proposals, which should, however, be first discussed with me. Follow some suggestions, which mostly involve doing the following steps:

1. get appropriate text(s)
2. decide how to encode them in TEI, consulting existing projects and publications
3. parameterise the TEI to get your schema
4. convert the texts to TEI
5. write a detailed header
6. adapt the TEI XSLT stylesheets and use them to produce an HTML version

Manuscripts and transcriptions

- The Newton Project http://www.newtonproject.ic.ac.uk/ is making Newton's writings freely available online, where each text includes the critical and diplomatic transcriptions, and, possibly, the facsimile. They use a tagset based on TEI (although not TEI conformant), but the TEI version is not available – they offer the writings only in (fancy) HTML. The student project consist of getting well acquainted with the Newton project (e.g. also their Tagging & Transcription Guidelines), then taking one of the *manuscripts* and encoding the two transcriptions in TEI, followed by a conversion to HTML. How much detail will be included in the tagging depends on the complexity of the chosen text.
- A similar task to the above is encoding a sample of one of the texts available at the University of Virginia Library, http://etext.lib.virginia.edu/eaf/
  The main task is to appropriately mark up the metadata and the linkage to the facsimiles.
- For the ambitious: The project *Scholarly Digital Editions of Slovenian Literature* http://nl.ijs.si/e-zrc/ currently contains two editions, e-Slomšek and e-Zois. For both, the complete edition can be downloaded, i.e. the TEI P4 source, a dedicated tei2html XSLT script and the HTML version. The project would attempt to update these resources, changing them from P4 to P5, and using (as much as possible) the standard tei2html scripts for the conversion. The headers and texts are currently written only in Slovene, but I'd be happy to offer help.

Corpora

- Construct a monolingual tagged corpus of a certain language / domain using BootCat http://corpora.fi.muni.cz/bootcat/ For tagging you can click on “Tag corpus” in BootCat, or, for some other languages, use the tagging service at http://nl2.ijs.si/analyze/ (choose CLOG lemmatiser and TEI output, but note that this is in fact not quite proper TEI). Convert the corpus into TEI and give it a detailed header, also explicating the tagset used. The HTML output should nicely display also tags and lemmas.
- Take one language pair for the Open Office component of OPUS parallel corpus, http://logos.uio.no/opus/ and encode it, using in-line encoding, in TEI. For 30 sentences / segments also encode word alignment. XSLT should be able to output aligned sentences.

Onomastica

- Take a newspaper or journal article from the Web, which contains at least 30 names (of people, places, companies, etc). Encode the article in TEI, annotating all names and dates, and link 30 of the names to a TEI encoded database (appendix) of names (using the module for Names and Dates). Make use of WikiPedia to find out more about the names mentioned.

Dictionaries

- From the on-line dictionary http://dictionary.reference.com/ choose 30 entries and encode them in TEI, using the dictionary module. The XSLT stylesheet should nicely display the dictionary sample