Dictionary Encoding Based on CSS and XML/HTML Parsers

Abstract

As long as computers require explicit instructions and digital data organized in certain ways appropriate objects/abstractions/models and their processing mechanisms should be clearly stated for comprehensive theoretical background. In this concern '*What is Text?*' and '*What is Reading?*' are the central research questions to enable computers process textual information on all levels from basic character rendering to semantics extraction.… Authors start the paper discussing these topics and seeking the answers in human brain (neuro-, cognition&cognitive) sciences, linguistics, and computer science.

Whereas the main goal of the practical part of the work is obtaining an XML/TEI P5 encoding of Jumakunova's Turkish-Kyrgyz dictionary, i.e. its lexicographic view. Results of the encoding should be placed to the global DH repository – TextGrid – to provide the basis for further deeper annotations. Methods and approaches should be used in similar lexicographic data encoding tasks.

Processing textual information and the format choice are the questions of efficiency, convenience, and methodology soundness. Ordered hierarchical representation of content objects was given the preference against other formats. Text is presented in (X)HTML&CSS format, what allows for easy-to-understand workflow (markup transformations) and has the abundance of powerful open source tools (markup and style parsers). These technologies provide all necessary control and manipulation opportunities for textual data processing.

The workflow of the dictionary structuring runs according to the lexicographic schema and obeys up-down approach of tokenization: when larger elements (of the higher level) are tokenized first and provide the output for the next level tokenization of smaller elements and so on. This approach helps to keep logic clear and thus reduce complexity. Structure recognition is implemented using Java parsing libraries based on CSS stylesheet and Document Object Models.

Output of the processing is an XML/HTML tree directly reflecting lexicographic structure of the dictionary. Conversion to TEI terminology to be implemented in XSLT.

Key words: structure, dictionary, recognition, text encoding, parsing

Autobiography

Personal Information:

Name: Kadyr
Surname: Momunaliev
Date of birth: 06/08/1985
Sex: male
Citizenship: Kyrgyz Republic
Nationality: Kyrgyz

1985: was born in the former Soviet Socialist Kyrgyz Republic. In 1991 it was renamed to Kyrgyz Republic or simply Kyrgyzstan.

1988-1997: obtained primary education (reading, counting) partially in kindergarten and partially in the first 5 years of Novopavlovka's Secondary School #2. The language of education was Russian, whilst Kyrgyz as a mother tongue was mainly used in the family.

2003: passed the Common Republical Examination with a score letting to be enrolled into Kyrgyzstan-Turkey Manas University with preference to study computer engineering.

2003-2008: Turkish, Intermediate English, literate Kyrgyz, programming languages, and other computer engineering subjects were learned. Bachelor thesis was concerned with development of a Turkish-Kyrgyz machine translation system.

2010: started career in academic community being enrolled in the master program at the same university's Natural Sciences Graduate School, and employed as a research associate. Research on creation of dictionary and encyclopedia databases for StarDict open source dictionary shell was conducted. Databases of Kyrgyz-Turkish (and vice versa) dictionaries (total 16 000 entires) as well as Kyrgyz Encyclopedia of Computer&Internet (700 entries) were created. XML markup for both dictionary and encyclopedia data annotations was used.

2010: graduation from Graduate School of Natural Sciences, and continuation of work at the university creating digital versions/editions of Kyrgyz disciplinary encyclopedias.

2014: with the help of a group of undergraduate students, faculty dean and Kyrgyz Encyclopedia chief editor 14 disciplinary encyclopedias, all in all 15 000 terms were automatically annotated. The data of the project was shared with other projects in field of education, such as Kyrgyz Wikipedia and others. Participated in the international conference on ICT promotion in Asia, ICTPA2014, where published materials of the work. Together with the faculty dean and the Kyrgyz Encyclopedia chief editor registered the database in the national patent agency.

2015: applied for a candidate program (stage in soviet style education system lasting for 3-4 years) of the Kyrgyz Technical State University, proposing a work on automatic annotation methods for Kyrgyz lexicographic texts.

2015-2016: researched annotation/encoding techniques of Camp's dictionary(TextGrid: Sneikel, Seipel). Sketched new approach for Turkish-Kyrgyz dictionary by G. Jumakunova (50 000 entries) based on markup transformations. Participated in the regional conference of Turkic languages processing, TurkLang2016, where published the main idea of the work.

Family Information: My farther is a teacher of physics, and mother is a doctor, therapist. I have two siblings, elder brother and younger sister.