Dictionary Structure Encoding via Style and Markup Parsers

Abstract

Textual Information and *Reading* are the central questions for enabling computers to process semantics, i.e. to replicate human's ability to read comprehensively. Knowing the exact mechanisms(models) and the nature(theory) of these concepts would allow for the delegation of the task to machines, via providing explicit instructions and digital constructions(abstractions) of the data... Authors start the paper reflecting on 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 representing its lexicographic structure. Results of the encoding should be placed in the global DH repository – TextGrid – to provide the basis for further ontology creation (Schneikel, 2009). Methods and approaches should be adopted for encoding of other lexicographic similar resources.

Processing textual information and the format choice are the questions of efficiency, convenience, and methodology soundness. Hierarchical representation of content objects (DeRose et al, 1991) was given the preference against other formats. Text is presented in (X)HTML&CSS format, what allows for easy-to-understand processing (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 (cf Fig. 1).

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. Entity recognition task is performed relying on text style or textual content or both.

Output of the processing is an XML/HTML tree directly reflecting lexicographic structure of the dictionary. Later conversion to TEI P5 terminology is implemented.