# An Example of Lexicographic Structure Recognition Based on Text Formatting and Delimiters

**Abstract**

This article discusses technical and theoretical questions aroused during the process of lexicographic structure recognition of human readable text provided by Jumakunova’s Turkish-Kyrgyz dictionary. The structuring task was accomplished according to logical lexicographic schema of the dictionary and benefited from markup and parsing techniques. Parsing techniques rely on text formatting patterns and text delimiters. Output of the parsing is a machine-readable XML tree which conveys dictionary logic.

**Key words:** text reading, lexicographic structure, parsing, formatting, delimiter, dictionary

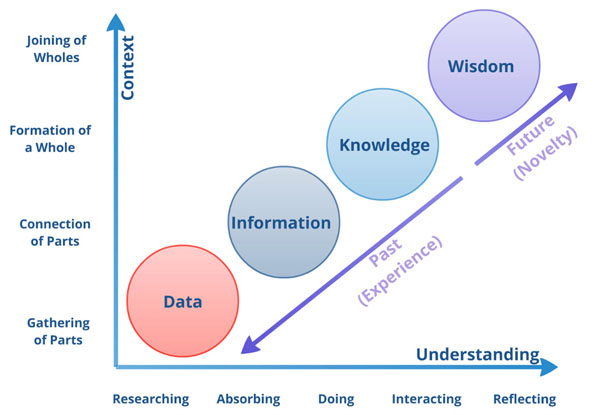
**Organization of the Article**

First section provides theoretical background for text reading process both by human and computer.

**Introduction**

Human readable dictionary texts may be viewed as logical (lexicographical), organizational (editorial), and esthetical (typographical) entities [TEI Consortium, 2015], or simply saying they have a form (expressed via layout and font) and logic[[1]](#footnote-1). Logically dictionary consists of series of entries, and entries in turn have their own tree like structure which is determined by a lexicographical standard (add Figure where entry is annotated by hand). Every structural unit performs its logical function, i.e. it conveys some metadata about its content. But the central question of the work is how human beings discern that units and how this process may be delegated to computers.

In the lexicographic context logical structure recognition process doesn’t invade inside the word boundaries, such an analysis in Natural Language sometimes referred as syntactic analysis(cite is needed). As an output of this analysis an intelligence agent has a syntactic structure which reflects semantic relationship between text level units (cite is needed). The process of syntactic (in our context lexicographic) structuring is the one of the main steps towards semantics extraction process which is the final point of meaningful reading. The chain of processes leading to understanding of the text being read depicted in Figure1. This processes start from seeing and end with semantics extraction or simply saying getting the meaning of the text.



The process of logical structure recognition by machine is copied from the process of logical structure recognition by human.

Logic is expressed by means of form. Form may be different but logic stays permanent.

Figure 1. Text Processing in Terms of Human Being

Figure 2. Text Processing in Terms of Computer Agent

TEXT

Visual Information

Color Distinction Knowledge

Shapes Distinction Knowledge

Shaped Information

Character Knowledge (Alphabet)

Shaped Textual Information

Word and Phrases Delimitation Knowledge (Syntax)

Syntactically Structured Information

Su(a/pre)ffix Delimitation Knowledge (Morphology)

Syntactically & Morphologically Structured Information

Concepts Knowledge (Lexicon & Grammar)

Semantics

TEXT

Image

Colors DB

Area Patterns DB

Layouted Data

Charasets DB

Layouted Textual Data

Text Units Delimiters or Patterns DB

Syntactical Parse Tree

Word Units DB

Syntactic& Morphologically Annotated Data

Concepts and Rules DB

Semantics

Computer

1. Programmers usually call it as ‘front-end’ and ‘back-end’ [↑](#footnote-ref-1)