Dictionary Structure Recognition Using Style and Markup Parsers

Abstract

This article discusses technical and theoretical questions aroused during the process of lexicographic structure recognition of human readable text, namely the text of Jumakunova's Turkish-Kyrgyz dictionary[1]. The goal of structure recognition is to create a machine-readable version/edition. Data structuring task was accomplished according to the lexicographic schema of the given dictionary using markup technology and programmatic text entity recognition techniques. Recognition task rely on the text style (formatting and layout) patterns and text delimiters (syntax of the dictionary) identification by means of (X)HMTL and CSS parsers. Output of the parsing is a machine-readable (X)HTML tree which conveys dictionary logic. Later this (X)HTML tree must be transformed to XML/TEI P5 terminology in order to provide global data interchange.

Key words: text entity recognition, lexicographic structure, parsing, text, delimiter, markup

Organization of the Article

First section provides an introduction in reading process both by human and computer.

Introduction

With the invention and intensive development of computerized information processing (ICT) new opportunities allowing for almost unlimited efficiency and scalability started to replicate human being tasks/abilities. Today cutting edge computer science branches such as artificial intelligence or machine learning are being widely used in tasks recently considered as only human doable such as visual object recognition, driving etc [Google AI, 2017]. The task of text reading is not an exception in this context.

Human being ability to read is limited physically, but machines can analyze and process texts of thousand and thousand times bigger than human being is able to do[S.Dehaine, 2010] & [TextGrid, 2015]. The central question here is to 'teach' computer to read and analyze texts as we do or at least as we want them to do. During thousands of years human communities have been developing complex reading/writing systems helping to encode, I.e to record information presented in a natural language they speak. In order to read and get the meaning of the written text every community member has to be able to derive/decode explicit information making use of implicit or meta information which is in fact commonly agreed rules instructing a reader the way he/she should interpret a text. Writing systems have syntactic, grammatical, orthographic etc rules. Above all this reader operates on the level of common semantic context which may

1

the most complex problem to be solved. This is caused by infinite variation of speech, abundance of implicit information it may contain and complex contextual relationships between lexical conceptions.

Although Natural Language Processing on the level of semantics still remains to be complex task for computer linguists (regardless of significant achievements in this field), computers are widely used in structural analyses of NL. Computers rely on formal instructions to process data, and lack common sense to make intelligent decision making [Fensel, 2000]. In this context they are still regarded as extension of human mind but not a replacement.

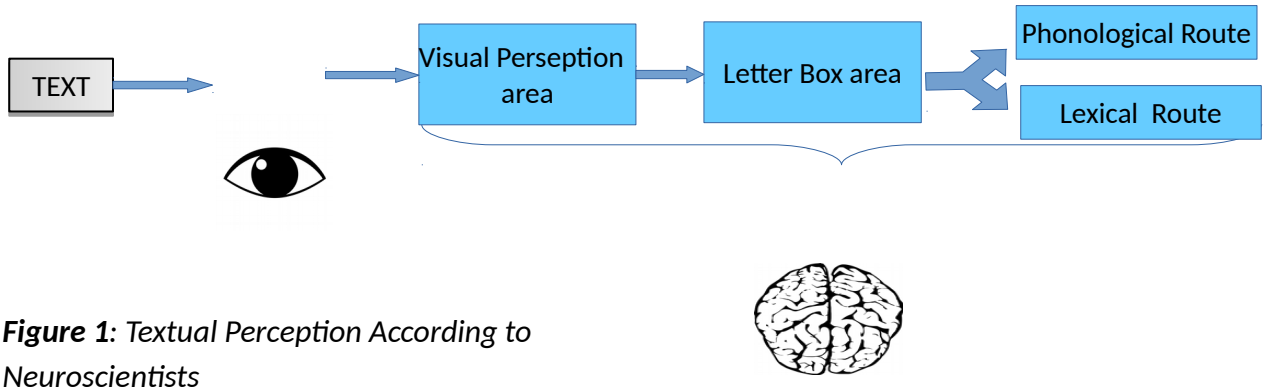


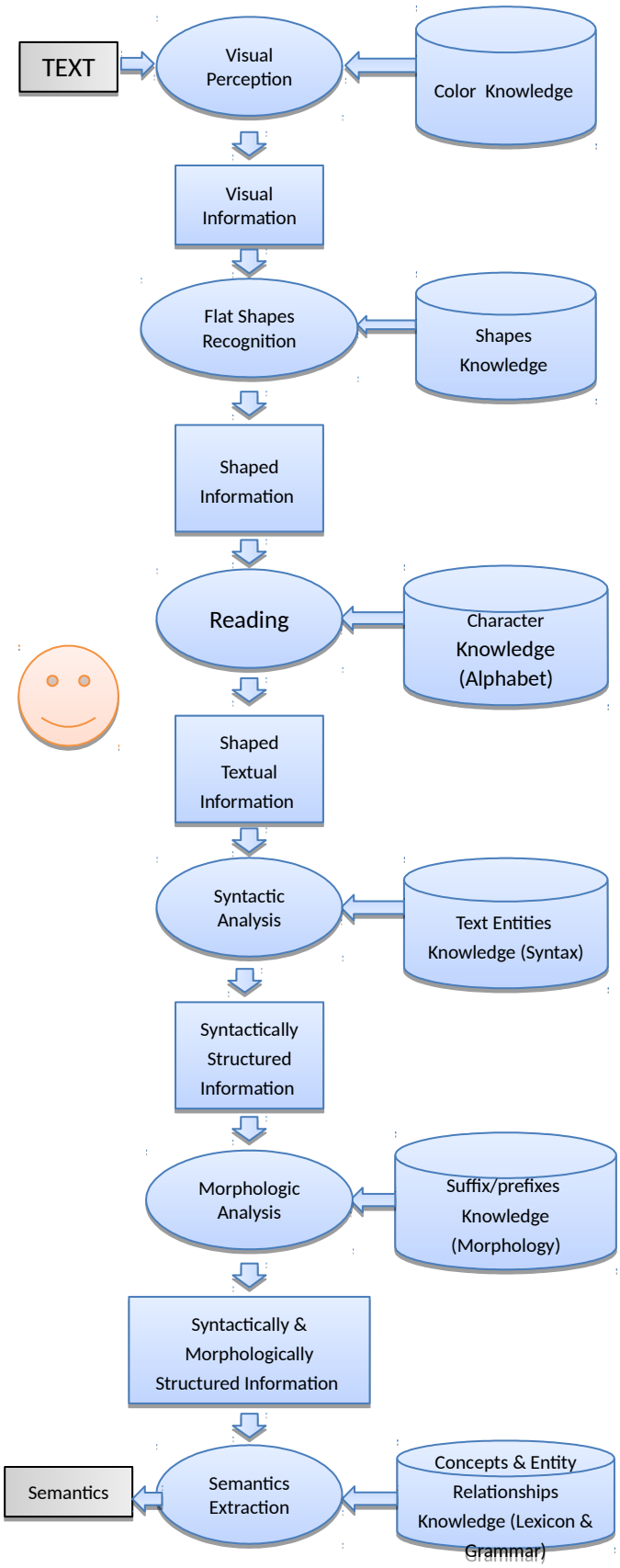
***Figure 1**: Textual Perception According to Neuroscientists*

Human-Computer Analogy

In order to convert human readable texts to machine-readable ones it is reasonable to understand information processing mechanisms or logic at first, and then the infrastructure and technology involved in this processes. Saying infrastructure for human being we imply its brain working principles, and for computer or machine we imply working principles of different kinds of information processors or software agents. The architecture I.e, physical structure is not concerned for computers, but in case of human being differentiating brain areas helpes to identify some reading phases.

Model of Textual Information Perception By Human Being

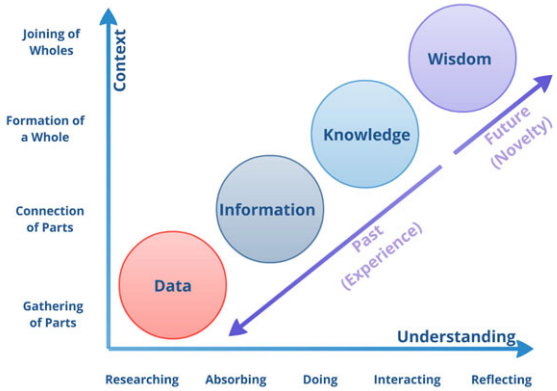
Figure 2. Prototypical Model of Textual Information Processing by a Human Being



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

According to the recent works in field of neuroscience [S.Dehaine, 2010], cognition and cognitive science our brain percepts and processes information by certain schema or model similar for all people. This model illustrates main stages of perseption and provides information about which brain area is responsible for that or this reading subtask. Even though these studies haven't provided comprehensive information about deeper mechanisms of semantic extraction or different kinds of thinking processes(see Figure2), they still are able to cast some light on textual perception or what we call reading mechanism.

This studies showed that information passes certain phases such as pictorial(visual), phonological (graphemes to phonemes), orthographic, and lexical, that our brain captures bigger patterns at first rather than processes smaller ones [S.Dehaine, 2010]



Perception of a text starts at our eyes. This process relies on color distinguishing abilities

Figure 3. Text Processing in Terms of a Computer Agent

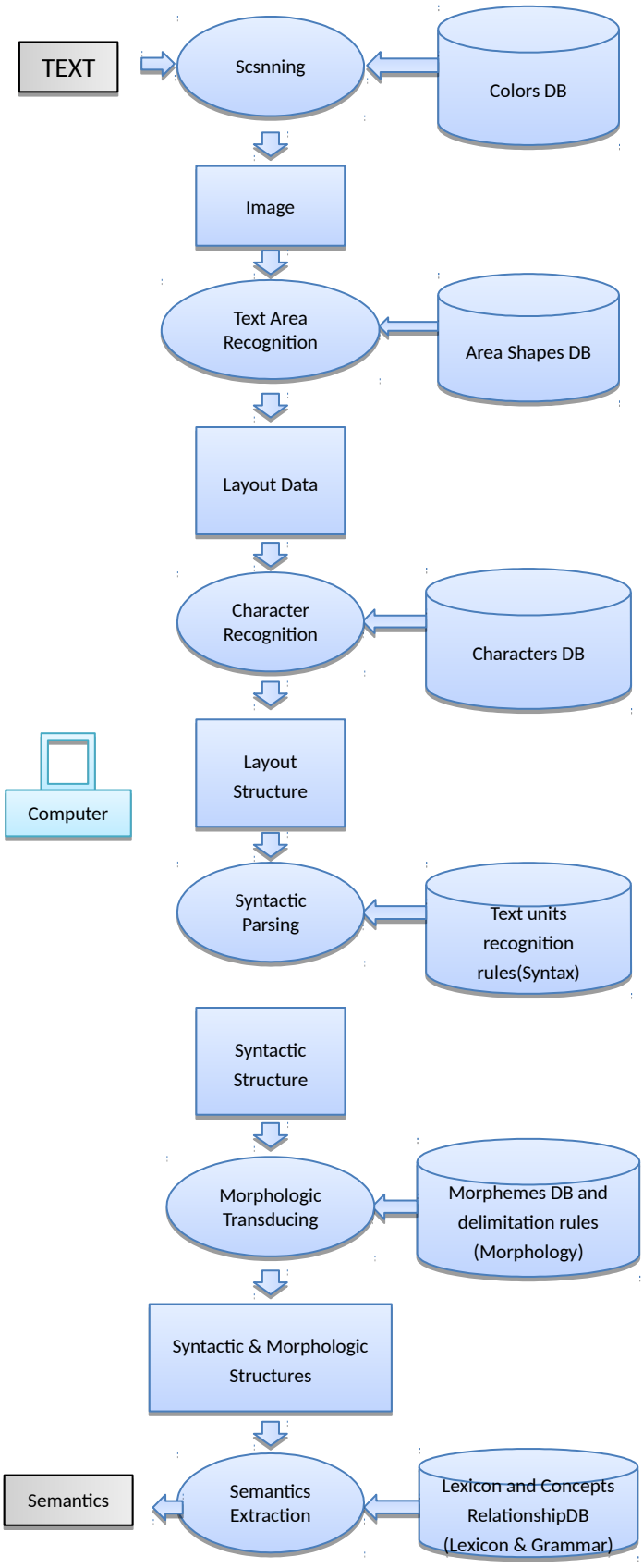


Figure 4: Evolution of Data

Human readable dictionary texts may be viewed as logical (lexicographical), organizational (editorial), and aesthetical (typographical) entities [TEI Consortium, 2015], or simply saying they have a form (expressed via layout and font) and logic[2]. 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 (human or computer) 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 Figure2 and Figure3. This processes start from seeing and end with semantics extraction or simply saying getting the meaning of the text.

2Programmers usually call it as 'front-end' and 'back-end'

Different Options to Process Textual Information

Bibliography

- ***Reading in the brain*. New York: Penguin, 2009. ISBN 0-670-02110-5 .[32]**
- *Consciousness and the Brain*: *Deciphering How the Brain Codes Our Thoughts*. Viking Adult, 2014. ISBN 978-0-670-02543-5.