D

A Novel Computerized Paleographical Method for Determining the Evolution of Graphemes

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

INTRODUCTION

Examining the genealogy of the graphemes is useful for exploring the evolution of the writing systems, reading undeciphered inscriptions, and deciphering undeciphered scripts. Researching the evolution of various scripts needs sophisticated and automatized procedures for exploring the fine links among the different writing systems, which research belongs to the field of the computational paleography. The traditional method compares the whole grapheme repertoire of the scripts. However, it is inefficient if the graphemes in the writing system were borrowed from different scripts. The proposed, novel procedure is based on the topological comparison of each glyph of the graphemes. Its algorithm is based on the global minimization of the topological differences in each descent branch of the graphemes.

The studies related to glyphs of a particular script are challenging topics for linguists, archeologists and anthropologists, one of them are deciphering undeciphered glyph discovered through excavation, reading patterns in glyphs transformation, etc. The effective software may accelerate the research time and to provide more accurate result through the automatic process. Producing that software needs a support of a solid mathematical model. Therefore, our main challenge is to develop such descriptive mathematical model as a useful framework for the archeologist and anthropologist in supporting their research. Such model can be used for describing how a glyph could transform from one shape to another, or try to estimate how other external parameters (e.g. writing instrument) could affect the transformation of a glyph.

The chapter gives some results belonging to the field of the computational paleography. A detailed model for measuring the characteristicness of the glyphs called Glyph Complexity Parameter (GCP) is described. Moreover, a method of transforming the topological information of glyphs into numerical descriptions is also presented. Based on the elaborated numerical grapheme model, researching the genealogical ties of graphemes can be reached with solving a mathematical optimization problem.

The efficiency of the proposed approach is demonstrated examples in the field of the Rovash (pronounced "rove-ash", other spelling: Rovas) script family. The specialty of the rovash paleography that the Rovash script family shows a vital evolution during the last centuries; therefore, it is ideal subject to test the models of the evolution of the glyphs. The rovash graphemes are used by nations in the Eurasian Steppe and in the Carpathian Basin. The five most important rovash scripts are the following: the Proto-Rovash, the Early Steppean Rovash, the Steppean Rovash, the Carpathian Basin Rovash, and the Szekely-Hungarian Rovash. The examples of the rovash grapheme ties prove the usefulness of the introduced grapheme descend exploring procedure.

BACKGROUND

The writing system is a symbolic representation of a language described in terms of linguistic units (Malatesha Joshi & Aaron, 2006), where the *grapheme* (e.g. a letter) is the smallest semantically distinguishing fundamental unit in a particular writing system. Likewise, *glyph* refers to a specific shape, a representation of a grapheme. Different glyphs can be certain representations of the same, abstract grapheme.

In the article, the grapheme is a key term, its definition and further descriptions are listed in Table 1 (Hosszú, 2011). International Phonetic Alphabet (IPA)

DOI: 10.4018/978-1-4666-5888-2.ch194

Table 1. Abbreviations

Symbol, Term	Description
/ /	*Slashes* denote phonemic representation of graphemes (phonemes).
[ ]	Phonetic transcription (phonemes and allophones).
< >	Transliteration value, value in Latin alphabet of graphemes in other alphabets.
allograph	A variant of a grapheme, which refers to various shapes of the grapheme, e.g., cursive, printed, strokes, upper case, lower case.
character	Character is a more general object than a symbol, since characters can be graphemes, diacritic marks and accents. Moreover, the characters have additional properties to the graphemes; namely, the characters have code points, which are used in the computerized presentation.
glyph	The shape of the grapheme with topological information.
grapheme	It is a minimally distinctive unit in a writing system. Grapheme is the abstraction of a symbol. Graphemes can be letters, ligatures, numerals, or punctuation marks. In this book, the terms character and grapheme are used interchangeably. The grapheme has the following properties: *(i)* the script belonging into, *(ii)* glyphs of the grapheme, *(iii)* sound values of the grapheme, and *(iv)* periods of use.
grapheme types	A grapheme can be a letter, a ligature, a numerical digit, a punctuation mark, or an individual sign of any writing system.
letter	It is a grapheme that represents one or more different phonemes.
ligature	A grapheme, which is composed of two or more graphemes, and represents two or more sounds pronounced after each other, e.g. the symbol '&'.
morpheme	It is the smallest semantic unit in a language.
orthography	The visual representation of a language, which is determined by the phonological, syntactical, morphological, and semantic parameters of the language (Malatesha Joshi & Aaron, 2006).
phoneme	A contrastive unit of sound in a language; a class of allophones (Rogers, 2005: 289). This sound exists in the mind of the speaker.
script	A *writing system*, which includes different orthographies. E.g., the Latin script has several orthographies, including the French, German, English, Hungarian, etc. orthographies.
script family	A group of scripts, which are closely related to each other. The Rovash (pronounced "rove-ash") script family includes the Proto-Rovash, the Early Steppean Rovash, the Carpathian Basin Rovash, the Steppean Rovash, and the Szekely-Hungarian Rovash scripts
symbol	The minimal individual visual unit of the inscription. Typically, the symbol is a realized grapheme.

symbols were applied for representing phonemes and occasionally actual sounds (allophones). In Table 1, each cited grapheme is presented by its glyph then it is usually followed by its grapheme-name and its phonetic representation for increasing the accuracy of the discussion.

The writing system of a spoken language changes periodically after being established (Rogers, 1999). The changes can occur as changing of set of symbols, which encompasses the shape transformation of a glyph. A cause of writing system alteration is the establishment of more advanced writing media or instruments. The establishment of more advanced writing technology introduces new writing technique, which impacts the shape of the glyph of a grapheme.

A large field of the computational paleography deals with the analyses of scribal hands, identification of inks, recovery of hidden or deleted text, and the segmentation and dating the document. Shiel et al. perform quality text recovery, segmentation and dating of historical texts. (Shiel et. al., 2009, pp. 159-174). They used a kind of spectral imaging named Hyperspectral Imaging (HIS, in other name Optical Reflectance Imaging), which processes data from across the electromagnetic spectrum (Chang, 2003). HIS is a non-destructive optical method that measures reflectance characteristics of a document and utilizes its high spatial and spectral resolution. The computer-based techniques for paleographic research aims to help solve paleographic issues and provide quantitative evidence to paleographical arguments (Hassner et. al., 2013, p. 18).

There are several undeciphered writing systems, including the proto-writings from the early 3rd mil-