International Journal of Data Analytics
Volume 3 • Issue 1

A Phenetic Approach to Selected Variants of Arabic and Aramaic Scripts

Osama A. Salman, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Hungary

https://orcid.org/0000-0002-7077-5640

Gábor Hosszú, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Hungary*

https://orcid.org/0000-0002-2856-9552

ABSTRACT

This paper aims to introduce the phenetic method for processing paleographical datasets and evaluating their similarity relationships. The presented numerical taxonomic method was applied for selected varieties of the Arabic and Aramaic scripts. The phenetic model was evaluated by hierarchical clustering and—after applying multidimensional scaling—a centroid-based clustering method. The hierarchical clustering results were presented as dendrograms (phenograms), while the centroid-based results were given in 2- and 3-dimensional Cartesian coordinate systems. The obtained results demonstrate that the numerical taxonomy's phenetic approach is useful in describing the distances between different writing systems. The long-term goal of the research is to apply machine learning tools to clarify the relationships between the large number of Aramaic and Arabic script variants. This study belongs to the field of pattern evolution, in which machine learning methods of biological evolution are used to model evolving patterns (such as writing systems) over time.

KEYWORDS

Hierarchical Clustering, K-Means, Machine Learning, Multidimensional Scaling, Numerical Taxonomy, Pattern Evolution, Phenetics, Scriptinformatics, Writing Systems

INTRODUCTION

Scriptinformatics deals with the investigation concerning the evolution of graphemes in scripts and the exploration of relationships between scripts, where scripts could be any sequence of symbols of cultural origin, such as historical writing systems (Hosszú, 2017; Hosszú, 2021a, p. 9). Evolutionary modeling of scripts includes phylogenetic modelling, namely phenetic, evolutionary, and statistical analyses of the studied scripts' features (Hosszú, 2014; Hosszú, 2021b).

This paper's main goal is to demonstrate how the developed exploratory data analysis algorithm applies to processing paleographical datasets and evaluating their phenetic modelling. The presented phenetic method was used for the selected varieties of the Arabic and Aramaic scripts. This modelling demonstrates that phenetic modelling, known from biology, can be used to detect similarities between otherwise closely related writing variants. The obtained phenetic results correlate well with their evolutionary relationships.

DOI: 10.4018/IJDA.297519

*Corresponding Author

Background

To date, several script evolution types of research have been published. The evolution and the reasons for character complexity are examined in detail, and it was found that the character complexity in a script is influenced by the evolution of that script (Miton & Morin, 2021). The logical relationship among glyphs belonging to the same script was modelled mathematically as three logical layers: topology, visual identity, and phonetic layers (Pardede et al., 2012). It was later improved by adding the semantic layer (Pardede et al., 2016), and recently it was extended with the style layer (Hosszú, 2021a, pp. 14–15). When comparing graphemes that are the key components of scripts, the comparison of graphemes with more characteristic shapes (glyphs) is more reliable because there is less chance of random coincidence of evolutionary independent graphemes. The shape characteristic of graphemes is expressed by the glyph complexity parameter (Hosszú, 2015). An investigation provides deciphering algorithms for reading historical inscriptions (Tóth et al., 2015; Tóth & Hosszú, 2019). A research was conducted to determine the actual version of a script used in an inscription from an unknown origin; in this research, the cluster analysis was applied (Tóth et al., 2016). The modeling of the evolution of various writing systems (scripts) is based on various phylogenetics methods, including exploration of phenetic relations of scripts based on the topological properties of their graphemes (Hosszú & Kovács, 2016) and other features (Hosszú, 2021a). Another research direction is applying convolutional neural networks to determine the degree of visual similarity between pairs of glyphs in various scripts (Daggumati & Revesz, 2019).

Among the various scripts developed during the evolution of writing, the Aramaic and Arabic script variants are especially interesting since calligraphic influences fully survive, and glyphs are intended to reproduce cursive handwriting (Baird, 2014). The two most important varieties of Arabic script are the Kufic and the Nashk (Taylor, 1883, Vol. I, pp. 317–318); the Kufic is square and monumental, it has practically fallen into disuse since the 14th c., and the Nashk is the cursive Arabic script in which books and newspapers are printed. There are several altering classifications of varieties of the Arabic scripts with different names, e.g., African (Fry, 1799, p. 8), Qarmatians (Carmatic, Karmathian, used in the north of Arabia); Diwani (Divân, Dīwānī, used during the age of the early Ottoman Turks); Jeli Thuluth (used for stones and buildings); Kufic (Cufic, Kufic, Kufi, is the earliest calligraphic variety of the Arabic script), Maghrebi (Maġribī, was developed in North Africa and Al-Andalus [Iberia]); Mauritanian (used in Morocco, Fez and North Africa); Naskh (Nischi, Nasḫ, Nesh, Nesih, Neshkhi, used for text processing); Nastaliq (Nastaʿlīq, Nashk-i-taʿliq, used for Persian, Punjabi and Urdu languages); Reqaʿ (used for private correspondence); Ruqʿah (Ruqʿä, Riqʿa, Rikaʿa), Taliq (Taʿliq, Taʿalik, used for the Persian language) and Thuluth (Sulsi, Ṯuluṯ, Thulth, Sülüs, used for ornamental purposes) (Fry, 1799, p. 7; Milo, 2002, p. 114, 122; Milo, 2013; Shahzad, 2020, p. 1). These categories partly overlap with each other. The Arabic script also has national subvariants, differing chiefly in the number of graphemes, e.g., Afghan, Hindustani, Malay, Persian and Turkish. Each of them contains additional graphemes to express the peculiar sounds of these languages. Only some selected Arabic script varieties are involved in the current research, namely, the so-called early, African, Kufic, Mauritanian, modern and Naskh.

This paper presents investigations and find the relationships among 10 different script varieties using machine learning algorithms, where 73 features of the examined script variants were involved in the phenetic analysis.

MACHINE-LEARNING ALGORITHMS FOR PHENETICS

Phenetics

For modeling the evolution of various writing systems, it is useful to introduce a numerical taxonomic approach. Numerical taxonomic research recognizes similarities and classifies taxa employing numerically-oriented procedures. The majority of the numerical taxonomic studies are based on