Hosszú *Heritage Science* (2024) 12:110
https://doi.org/10.1186/s40494-024-01211-7

Heritage Science

RESEARCH

Open Access

Validation of graph sequence clusters through multivariate analysis: application to Rovash scripts

Gábor Hosszú[1*]

Abstract

This paper introduces the concept of pattern systems that evolve, with a focus on scripts, a specific type of pattern system. The study analyses the development of different script systems, known as scriptinformatics, with a focus on the historical Rovash scripts used in the Eurasian steppe. The aim is to assess the traditional classification of historical inscriptions, referred to as script relics, into distinct Rovash scripts. Clustering and ordination techniques were used to perform multivariate analyses on Rovash scripts and inscriptions. The study presents two new measures, the script-specific holophyletic index and the joint holophyletic index, for evaluating trees produced by hierarchical clustering. The results indicate that holophyletic indices can validate the traditional assignment of inscriptions to scripts through phylogenetic tree evaluation. This method can be extended to include pattern systems with evolutionary properties and graph sequences derived from them, as well as additional scripts and inscriptions.

Keywords Hierarchical clustering, Ordination, Pattern system, Phylogenetic inference, Principal component analysis, Scriptinformatics

Introduction

The different scripts and variants used by humanity have evolved. Comprehending the evolution of these scripts could be crucial in interpreting the numerous unread inscriptions of the past, commonly referred to as script remains or script relics [1, 2]. In addition, deciphering the origins of ancient manuscripts may require an accurate description of the evolution and interaction of many script variants associated with a single script. The study of the evolution of historical scripts and script variants is known as scriptinformatics [3, 4]. The focus of the present research is on the evolution of historical scripts.

The script is generalised through the pattern system, which is a specific form of symbolic communication. It includes symbols, syntax, and layout rules that determine their use. The research focuses on studying pattern evolution, which involves pattern systems with evolutionary properties.

When examining the background of scriptinformatics research, it is worth noting that although the evolution of scripts is typically a centuries-long process, only a few script relics have often survived, which is insufficient to trace the entire evolution of a historical script. Another problem is that not all inscriptions in some surviving script relics can be identified and deciphered. Archaeologists find thousands of short inscriptions, and even the particular script used for each inscription is difficult to identify, so it is often impossible to decipher the inscription. Uncovering the evolution of scripts can help interpret and decipher these ancient script remains. The efforts of the author's research group cover a wide range of topics, such as the use of data mining methods to discover similarities between scripts [1], the reconstruction of the descent lines of symbols in different scripts [2], the

*Correspondence:
Gábor Hosszú
hosszu.gabor@vik.bme.hu
[1] Department of Electron Devices, Budapest University of Technology and Economics, Budapest 1111, Hungary

deciphering of inscriptions written with unknown script variants [5], and methods for testing the correctness of reconstructed lines [6].

The best way to reconstruct evolution is to use phylogenetic statistical methods. Evolutionary processes are observable in various human, natural, and engineering sciences fields. Hence, methods of phylogenetic analysis can be employed in these areas [7]. Nakhleh et al. applied several phylogenetic reconstruction algorithms to explore the evolution of the Indo-European languages [8]. It is helpful to use various data mining algorithms to achieve this goal, particularly some exploratory approaches—identification of frequent subtrees as common patterns were invented by Deepak et al. [9]. Phylogenetic methods have been used to study manuscript versions of the Mahabharata written in various Brahmic scripts [10]. The phylogenetic relationships of the scripts used for each manuscript have been found to differ from the evolutionary relationships of the text's parts, also called stemmatic relationships (Table 1). It has been suggested that the evolutionary relationships of the scripts used for the manuscripts should be viewed as external (codicological) data, while the phylogenetic relationships of the textual versions of the manuscripts should be viewed as internal (stemmatological) data [11].

Biolcati et al. discovered the original order of poems by paleography, codicology, X-ray fluorescence spectroscopy and statistical analysis [12]. They used a type of artificial neural network, a self-organising map, as an unsupervised machine learning method applied to create a low-dimensional representation of a higher-dimensional data set while preserving the topological structure of the original data.

The article is structured as follows: Firstly, the problem to be solved is defined. Secondly, a summary of the statistical methods used is provided. Thirdly, the new algorithm developed to validate the traditional classification of graph sequences into individual pattern systems is described. Fourthly, the results obtained are presented. Finally, the paper concludes with an analysis of the results and conclusions.

Identifying the problem

Notions of pattern evolution research and scriptinformatics

The concept of a pattern system is broader than a set of scripts. Pattern systems encompass various historical scripts, the Morse code system, and the design rules of microelectronic layout design. Pattern systems typically undergo evolution. For instance, the Morse code system

Table 1 Scientific fields related to scripts and their evolution [3]

Term	Description
Computational paleography	It provides algorithmic support for deciphering ancient inscriptions and is a part of scriptinformatics [6]
Digital paleography	Digital paleography [13–16], also known as computerized paleography [17] or computer-aided paleography [18], is a subfield of digital humanities. It combines traditional paleography with computer methods such as digitization of old codex data, author identification through image recognition, and categorization of writing patterns [19, 20]
Epigraphy	Epigraphy is a branch of the humanities that deals with the study and decipherment of ancient inscriptions created through carving, scratching, or engraving
Evolutionary analysis	It is used to reconstruct phylogenetic trees or networks
Grammatology	Gelb introduced the term 'grammatology' to study writing systems and their relationship with speech, religion, and art [21]. This concept falls under the humanities
Graphemics	Graphemics, also known as graphematics, is a field of linguistics that studies writing systems and their essential elements, graphemes. It focuses on the articulatory properties of written language and their relationship to spoken language. In contrast, scriptinformatics explores the evolution and interaction of individual scripts and identifies the graphs of various inscriptions
Paleography	Paleography is a field within the humanities that involves the study of ancient writing, including the interpretation and dating of historical manuscripts. In a broader sense, paleography is the study of all types of historical inscriptions, documents and scripts, including epigraphy
Pattern	In the context of pattern evolution research, a pattern is a symbol or graph sequence; see Table 2 for definitions
Pattern evolution research	It is the study of the temporal evolution of pattern systems using methods from data mining, multivariate analysis, evolutionary analysis and bioinformatics. On the one hand, it is an evolutionary discipline because it models the evolution of pattern systems; on the other hand, it is a kind of applied computer science
Pattern system	It is a form of symbolic communication, which various features can characterise
Scriptinformatics	Scriptinformatics models the evolution of scripts as unique pattern systems. It is a subfield of pattern evolution research and belongs to evolutionary disciplines. Additionally, it is a type of applied computer science
Scriptology	Blatner proposed the term 'scriptology' to describe the scientific field of writing [22]. Scriptinformatics differs from scriptology in that it primarily examines the evolutionary properties of scripts
Stemmatology	The study of the evolution of traditions recorded in manuscripts is also known as stemmology [23–25]

Table 2 Basic terms of pattern evolution research with special emphasis on script informatics [3]

Term	Description
Glyph	A property of the symbol, i.e. the drawing of the symbol
Graph	A self-contained, visually or otherwise perceptible formal unit. It is the implemented glyph of the symbol
Graph sequence	It implements a symbol sequence with a technology, considering the pattern system's layout rules. A graph sequence is composed of graphs. An example of a graph sequence is a sequence of measured quantities (a measurable record) that can be interpreted as a symbol sequence (i.e. text) created with Morse code. Another example of a graph sequence is an inscription (in its physical reality) that represents a symbol sequence (i.e. text) according to the symbol set, syntax and layout rules of a script. The concept of a graph sequence in pattern evolution research is analogous to a fossil in biological evolution
Inscription	A particular case of the graph sequence where the pattern system is a script used to create the graph sequence. In this case, the graph sequence is an engraving or a written text
Layout rules	A feature of pattern systems, they regulate the appearance of a graph sequence representing the realisation of a symbol sequence with a given technology. Suppose the used pattern system is a script. In that case, the layout rules describe some properties of the script, such as rendering rules (alignment, placement), emphasis (highlighting), text separation (e.g. by parallel lines) and overall appearance of the inscription
Symbol	A symbol can be a grapheme (having a sound value or meaning), a tamga (having a specific meaning, e.g. property mark) or a decorative sign (having a specific decorative function)
Symbol sequence	A data sequence of symbols belonging to a pattern system, edited according to the syntax and layout rules of the pattern system. An example of the symbol sequence is a message created with Morse code or a text created using a specific script. The concept of symbol sequence in pattern evolution research is analogous to the genetic stock of an organism in biological evolution
Syntax	It refers to the syntactic rules that govern the formation of symbol sequences in pattern systems, including punctuation, hyphenation, writing direction, and line order

in use today (in different versions) was developed by many inventors before and after Samuel Finley Breese Morse [26] .

In scriptinformatics, which is concerned with the evolution of scripts, and in pattern evolution research, which is concerned with the modelling the evolution of pattern systems (as generalisations of scripts), the term taxon (taxonomic unit) known from phylogenetics and numerical taxonomy can refer to any pattern system (e.g. a script). However, in multivariate methods used in phylogenetics, numerical taxonomy, and pattern evolution research, the basic unit is an object (or a data point). When multivariate methods are used in pattern evolution research, an object must be assigned to either a taxon or something else to be analysed, e.g. a graph sequence. Phylogenetic terms in pattern evolution research, including scriptinformatics are presented on Table 3.

The purpose of phylogenetic inference is to reconstruct the temporal evolution of the individuals or their groups (taxa) under study. The graphically visible result is the phylogenetic tree or network. In terms of phylogenetic inference, it is an optimisation whose goal is to decide

Table 3 Phylogenetic terms in pattern evolution research, including scriptinformatics

Term	Description
Distance (in general, dissimilarity)	It is the dissimilarity of two objects. In a broader sense, distance has the same meaning as dissimilarity (e.g. it is used this way in the expression 'distance-based phylogenetic inference method', see below), but in a narrower sense only the dissimilarity can be called distance, in a mathematical sense it is metric, and it satisfies—among other conditions—the so-called triangle inequality; see Eq. (3) below
Feature	It is used to describe a pattern system. Features can be symbols, syntax and layout rules
Feature state	The specific value of a particular feature
Object	In multivariate analysis, the examined entities are objects. In pattern evolution research, an object can be, e.g. a taxon or a graph sequence. In the present analysis, objects are scripts or inscriptions. Features can describe these objects
Taxon	In phylogenetics, the examined entities are taxa, a specific object type. In the present research, a script can be a taxon. Typically, the leaves or internal nodes of a phylogenetic tree
Phylogeny	The history of the evolution of an object (species or any evolutionary object). A phylogenetic tree or network can represent it
Phylogenetic tree	The result of phylogenetic analysis when this result is not a reticulation (i.e., not a phylogenetic network). Its versions include additive tree and ultrametric tree
Branch length	It represents the evolution between each node on the tree and the number of changes in feature states
Additive tree	Its alternative name is phylogram. Here, the branches on the tree and the branch lengths are informative
Ultrametric tree	An additive tree can be rooted so that all paths from the root to a leaf have the same length; it describes times of divergence

between possible phylogenetic trees or networks based on a given criterion. There are two main types of these methods, feature-based (also known as phylogenetic character-based) and distance-based. In this research, only the distance-based methods are used. A phylogenetic tree (also known as a phylogeny) is created when individual lines of descent are not connected during descent. On the other hand, in the case of the phylogenetic network, the individual lineages can be connected, which is called reticular evolution. Current research is limited to phylogenetic trees. In the future, with a larger database, it is conceivable that the possibility of reticular evolution will also be taken into account to obtain a finer evolutionary model.

The type of phylogeny we study is called a phenogram. A phenogram is the result of a special case of phylogenetic inference called phenetic analysis. It is a statistically constructed tree that gives only the degree of similarity. In phenograms, the length of branches represents the similarity between taxa. Phenetic analysis is the simplest form of phylogenetic inference; however, the primary purpose of phenetic modelling is not to explore phylogenetic relationships. Another type of phylogenetic analysis, cladistics, compares characters in related taxa to determine relationships between ancestors and descendants. It is a specific method that assumes a relationship between taxa [27]. Cladistics focuses on inferring evolution from changes in individual features (characters in the biological sense) or changes in feature states (character states in the biological sense). While phenetic analysis primarily expresses similarity between taxa, cladistic analysis determines phyletic relatedness.

If there is no clear ancestor–descendant relationship between the analysed taxa, phenetic methods should be used instead of cladistic methods. This is because cladistic methods rely on a clear ancestor–descendant relationship [28]. In cases where the direction of development of individual features (known as feature polarity) is unknown, a phenetic approach is more appropriate. Due to the limited number of surviving inscriptions for the four scripts analysed in this study, the feature polarity is not always known. It is recommended to start with phenetic analysis, which requires less prior knowledge, rather than opting for a more advanced cladistic analysis based on assumptions about feature changes.

Clustering is a vital technique in phenetics as it models the evolutionary relationships of taxa for phylogenetic analysis. Various clustering methods are used in phenetics, including distance-based phylogenetic inference methods such as UPGMA [29], WPGMA [30], neighbour-joining (NJ) [31], and the Ward method [32]. For this analysis, we used WPGMA and NJ, while the results of other linkage methods are reported in the Additional file 1.

The problem of classifying inscriptions in different scripts

The present study performed various multivariate analyses, including clustering and ordination. The aim was to clarify whether the inscriptions belonging to the scripts under study were correctly assigned to each script. This question may arise because all but one of the scripts under study are long extinct, so their evolution can only be known from reconstruction. Notably, these scripts were initially used by the Turkic people of the Eurasian steppe and later applied to various languages. Of these scripts, Turkic Rovash (TR, also known as Turkic runic or runiform), Carpathian Basin Rovash (CBR) and Steppe Rovash (SR) have been extinct for a millennium. In contrast, the use of Székely-Hungarian Rovash (SHR) as an additional script in the Carpathian Basin has survived to the present day [6, 33]. During the first millennium BC, more and more populations migrated from east to west across the Eurasian steppe. Therefore, the similar Rovash inscriptions found from Inner Asia to the Carpathian Basin must be related; their similarities in the data can typically be considered a phylogenetic signal. In principle, the representatives of all scientific fields dealing with Rovash scripts agree that they are somehow related. Of course, the possibility of the independent evolution of some features must also be considered, but this has been largely considered in the previously completed feature engineering phase [3]. Table 4 clarifies related terms due to the terminological confusions and misunderstandings that exist in the literature.

The TR was used in Inner Asia, and the SHR in the Carpathian Basin. They were located far apart from each other. The substantial quantity of TR inscriptions, some of which are quite extensive, and the continued understanding of SHR allow for the categorisation of newly discovered inscriptions as either TR or SHR, despite their differences. In the case of CBR and SR, the situation is quite different. From Inner Asia to the Carpathian Basin, archaeologists have continued to find unreadable inscriptions for a long time. Today, many of them have been deciphered by linguists, but this is still controversial. Most of these inscriptions found in the Carpathian Basin are similar to each other but different from TR and SHR. Their collective name is CBR. The other inscriptions from Inner Asia to the Carpathian Basin are similar and different from the well-defined TR or SHR. Their collective name is Steppe Rovash (SR). Note that there are many other names for different groups of these inscriptions. Only deciphered inscriptions have been used in this study. References and brief

Table 4 Concepts to be clarified in relation to the present research

Term	Description
Runic script family	Various Germanic peoples used the Runic scripts between the second and fifteenth centuries AD. The oldest Runic scripts are Elder Fuþark (AD ca. 150–650), Anglo-Frisian Fuþorc (AD ca. 450–1000), and Younger Fuþark (7th–11th c. AD) [34–36, 38]. The Runic script family evolved independently from the Rovash scripts
Old Hungarian orthography of the Latin script; shortly, Old Hungarian script	Its stages: (i) medieval systems (11th c. AD—1530 s) and (ii) modern system (1530s—1832, up to the publication of the first Hungarian spelling rules)
Hungarian orthography of the Latin script	1832—present
Proto-Rovash	It is a hypothetical script that was developed in Inner Asia during the fifth and sixth centuries AD. It was based on Aramaic-Middle Iranian and Brahmic scripts and may have been influenced by Eurasian tamgas [3]
Turkic Rovash (also known as Turkic runic or runiform) script (TR)	TR was used by Turks in Inner Asia. Its inscriptions are believed to date back to the 7th to 10th centuries AD [38]
Carpathian Basin Rovash script (CBR)	CBR was used in the Carpathian Basin between the seventh and eleventh centuries AD to record various languages, primarily Hungarian, and to a lesser extent Turkic, as well as sporadically Alan and Slavic [39]
Steppe Rovash script (SR)	SR was used in the Eurasian Steppe and sporadically in the Carpathian Basin to record Turkic and possibly Alan languages during the 8th to tenth centuries AD [40]
Székely-Hungarian Rovash script (SHR)	SHR was first used by Hungarians in the Carpathian Basin and the earliest deciphered inscription dates back to around 900 AD [41] or the first half of the tenth century AD [42] in Bodrog-Alsóbű, Hungary. The majority of SHR inscriptions are in Hungarian, although there are also some in Cuman and Latin. The script known as Székely-Hungarian Rovash is sometimes erroneously referred to as 'Old Hungarian'

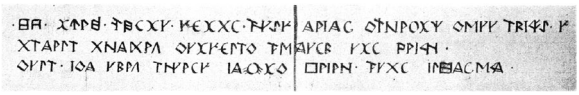

Fig. 1 Copy of the original mural inscription in Constantinople [6]

descriptions of these inscriptions can be found in the following sources: [3, 43, 44].

Due to the limited knowledge available about the examined scripts, the traditional classification of the Rovash inscriptions into scripts is inherently uncertain. In this situation, modelling evolutionary relationships based on clear and objective criteria can be helpful.

An example of the Rovash scripts is shown in Fig. 1. The background to the inscription is that in 1515 in Constantinople (Turkish: Istanbul in Turkey), Barnabas Bélay, the ambassador of the Hungarian King Vladislaus II (1490–1516), found that he had to wait 2 years for the Sultan Selim I (1512–1520) to let them go home. During this time, a Hungarian named Thomas Kidei Székely wrote this SHR inscription on the wall of the Ambassadors' House. Between 1553 and 1555, it was discovered and copied by the numismatist and epigraphist Hans Dernschwam. An accidental fire later destroyed the building. Translation of the inscription: "Written in the year one thousand five hundred and fifteen; delegate of King Vladislaus was sent here. Barnabas Bélay waited here for 2 years; the emperor did not do [anything for them]. Thomas Kidei Székely wrote here; Emperor Selim housed here with one hundred horses."

Fig. 2 Two pages of the Patakfalvi inscription with SHR text [45]

Another example is the Patakfalvi inscription dating from 1776 to 1785; see Fig. 2. The Székelys were responsible for protecting the eastern borders of the country and had an autonomous legal system within the Kingdom of Hungary's legal system. This manuscript describes the vital inheritance law of the Székelys. According to medieval-origin law in the Kingdom of Hungary, only sons inherited family property, while daughters had the right to a dowry. However, in the absence of a son, the privilege of the Székelys was that daughters inherited family property in the same way as sons. This was called 'boy-daughterhood' in Székely law [45].

Very few surviving inscriptions are known for some of the Rovash scripts examined in this study, and most of

them are either of unknown age or their age can only be roughly estimated. The so-called feature engineering step must precede the data mining algorithm to analyse their relationships properly. This feature engineering typically means feature selection to filter out the uninformative features and increase data analysis efficiency. Feature selection is usually a dimensionality reduction method [46].

In the applied feature selection process, the possible ancestors of the Rovash scripts have been identified [3]. These preliminary studies found that each Rovash script contains similar proportions of Aramaic-Middle Iranian and Brahmic origin features. Therefore, the Rovash scripts under study likely had a common ancestor (Proto-Rovash) that showed Aramaic-Middle Iranian and Brahmic influences. In the present study, the set of $f = 119$ features defined in the analysis carried out earlier [3] was used. Table 5 presents the name and the traditional classification of inscriptions into different Rovash scripts. Their usual classification is based on the literature on Rovash scripts. Their detailed description is presented in the literature [3, 6, 33, 43].

As shown in Table 8, the number of inscriptions included in the analysis is $n = 57$. During the tests carried out, the examined objects included in the clustering are, in some cases, scripts, and in other cases, inscriptions belonging to these scripts.

An inscription as a knowable part of a script version

When scriptinformatics models the evolution of some inscriptions, the aim is to model the evolution of the knowledge of the individual inscription makers (scribes) from generation to generation. This knowledge also includes those features of the given script (symbols, syntax, or layout rules) that were not needed to create each inscription. The complete knowledge of the scribe (writer) can only be determined if the inscriptions in question are longer texts or even abecedarien explicitly intended to present a complete script (writing system). In many cases, however, the inscriptions under consideration are too short, so only some of the script's features were needed to create them. In other words, each inscription only approximates the full knowledge of the writer who created it. This error may limit the accuracy of evolutionary models of script versions based on extant inscriptions.

It's worth noting that the inscriptions being studied typically aren't replicated from each other, which means that the phylogeny of the inscriptions is fundamentally distinct from the stemmatic tree or network of the different versions of a particular text.

Phylogenetic trees consisting of a non-trivial number of input sequences are generated using computational phylogenetic methods. Dissimilarity matrix methods, such as WPGMA or NJ, were used to calculate an object-object dissimilarity matrix from the object-feature data matrix using a dissimilarity measure. Another method is PCA, which is also used to detect similarities between inscriptions. All these phylogenetic tree reconstruction procedures are used in the developed method.

The Rovash inscriptions in the study exhibit variations. Correctly attributing them to different Rovash scripts allows for multivariate analyses to group together inscriptions typically assigned to the same script and differentiate those belonging to different scripts.

The input data structure

A data structure consisting of $f = 119$ features was created during the previous feature engineering step. These were used to characterise both individual scripts and individual inscriptions. The book Scriptinformatics [3] provides a detailed description of the scripts used in

Table 5 Traditional classification of inscriptions under study, along with their estimated date of creation, where available

Class (script)	Name and date of the inscription	No. of inscriptions
Turkic Rovash (TR)	Almaly II, Bichiktu Boom III, XV, Bilge Khagan (AD 735), Kalbak Tash II (8th c. AD), Koytübek, Kül Tegin (AD 732), Kuljabasy I (second half of 8th c. or 9th–10th c.), Kuljabasy II, Kurgak I, Mendur Sokkon I, Tamgaly (9th–10th c.), Tuva III, Urkosh (8th–9th c.), Yabogan, Zhon Aryk (first half of 8th c. AD)	16
Székely-Hungarian Rovash (SHR)	Bodrog-Alsóbű (around AD 900 or first half of 10th c.), Vargyas (12th–13th c.), Homoródkarácsonyfalva (around 13th c. AD), Stick Calendar (ca. 15th c. AD surviving in a 17th c. copy), Bágy (15th c. AD), Nikolsburg (1490–1526), Székelyderzs (1490 s), Bögöz (end of 15th – beginning of 16th c.), Csíkszentmihály (1501), Constantinople (Fig. 1, 1515), Szamosköz (partly before 1593, partly in 1604), Wolfenbüttel (1592–1666), Rudimenta (1598), Farkaslaki (1624), Bonyhai (1627), István Csulyak (1610–1645), Kájoni-Ancient (1673), Bél (1718), Patakfalvi (Fig. 2, 1776–1785)	19
Carpathian Basin Rovash (CBR)	Ozora-Tótipuszta (last third of 7th c. AD), Jánoshida (last third of 7th c. AD), Kiskőrös-Vágóhíd (last third of 7th c. AD), Környe (end of 7th c. AD), Szarvas (first half of 8th c. AD), Kiskundorozsma (end or the last third of 8th c. AD), Nagyszentmiklós (8th–11th c. AD)	7
Steppe Rovash (SR)	Jitkov (8th c. AD), Achik Tash (8th c. AD), Mayaki (8th–9th c. AD), Mayatskoe 1, 2, 5, 10 (9th c. AD), Khumara 6, 7, 8 (mid-9th – beginning of 10th c. AD), Kermen Tolga (8th–10th c. AD), Novocherkassk (8th–10th c. AD), Homokmégy-Halom (10th c. AD), Algyő (first half of 10th c. AD), Kievan Letter (934–938)	15

this article. The present study utilizes the similarity feature groups (SFGs) outlined in that book. Table 6 shows some features and the corresponding feature values (variable values describing the presence of each feature in each object) for some scripts and inscriptions. The symbols and notations in Table 6 can be found in the book Scriptinformatics.

The process of deciphering an inscription (in general, graph sequence) is always conducted in a specific language. However, for the purposes of this study, the language used for decipherment is irrelevant. Instead, the focus is on the properties of the assigned symbols within the inscriptions, which are typically sequences of graphs. These properties are displayed in the form of feature states in the object-feature data matrix. The feature state (1 or 0) is determined by whether the current symbol can be assigned to at least one graph with a specific inscription (graph sequence) (1) or not (0).

The research goals and used approaches

One type of the methods used in data mining is represented by the various cluster analysis algorithms, which are a vital component of exploratory data analysis. There are various clustering and classification methods. A particular problem is evolutionary clustering, where the object-feature data matrix evolves dynamically over time; therefore, the cluster analysis result is looked for at each time step [47, 48]. Evolutionary clustering models the temporal evolution of observed data by describing their typical evolution phases [49]. A general requirement of evolutionary clustering is that it should be robust to short-term variations.

Another type of data mining is principal component analysis (PCA), which is a multivariate method used to reduce the complexity of the dataset while preserving data covariance [50]. Further example of data mining is the application of convolutional neural networks (CNN) to measure the degree of visual similarity between pairs of glyphs in various scripts [51]. In the case of extinct scripts, their characteristics can be identified from surviving inscriptions, for example, in the case of the Elymian script once used in Sicily [52] or a type of Brahmic script [53]. A similar line of research is identifying scribes and clustering scripts based on surviving inscriptions

Table 6 A part of the input data (object-feature data matrix)

Feature no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Feature description	<a>	<á>	<A, e>	<A>	<ö>	<A>; <e>	<Ẅ>	<A>	<A>	<e>	<e>	<é>	<e>	<b^1>	<b>	<β>	<b^1>	<g^2>; <g>
TR	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	1	1
SHR	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	1	0	1
CBR	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0
SR	0	0	0	0	0	1	0	1	1	1	0	0	0	1	1	1	1	0
Yabogan	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Almaly II	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
Kalbak Tash II	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Kül Tegin	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1
Kurgak I	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Tuva III	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1
Vargyas	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Stick Calendar	1	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1
Nikolsburg	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1
Csíkszentmihály	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1
Constantinople	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Rudimenta	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1
Kájoni-Ancient	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
Bél	1	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	0	1
Patakfalvi	1	1	0	0	0	0	1	0	0	0	1	1	0	0	1	0	0	1
Szarvas	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0
Nagyszentmiklós	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0
Jitkov	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0
Achik-Tash	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
Mayatskoe-10	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	1	0
Novocherkassk	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0

[54, 55]. An essential goal in data mining is to ensure the homogeneity of each class. Metzner et al. derived a theoretical limit of classification accuracy for overlapping data categories. They selected the most efficient classifiers, such as perceptrons and Bayesian models, that perform the separation at these accuracy limits [56].

The objective of this research is to evaluate the accuracy and consistency of assigning graph sequences (in this case, inscriptions) into distinct pattern systems (in this case, scripts). The question is whether the traditional classification of inscriptions placed in one of the examined pattern systems (the Rovash scripts) is correct, that is, whether all inscriptions should be classified according to the traditional classification for each script.

In the field of machine learning, there are specific methods for classification. The goal of classification is to assign individuals to predetermined categories. Therefore, classification methods require a learning database with feature vectors describing the individuals and their corresponding correct class labels. The classification algorithm aims to create a model that can accurately classify individuals not seen during teaching, even after processing the learning database. However, in our case, since there is no learning database, only unsupervised machine-learning methods can be applied.

The developed method is that, on the one hand, a phylogenetic tree of the investigated scripts is created. On the other hand, a phylogenetic tree is also created from the investigated inscriptions. Given the diversity of the Rovash inscriptions in this study, if their assignment to different scripts is correct, the various multivariate analyses place inscriptions traditionally assigned to the same script closely together while placing those belonging to different scripts far apart.

Statistical background

This section briefly describes the standard statistical methods used in the new algorithm developed.

Similarity and dissimilarity measures

Matching the dissimilarity measure to a given data structure can affect the performance of data mining algorithms for data analysis. The determination of a suitable dissimilarity measure between objects is a critical step in data mining [57].

The features of the objects studied in this research are of a binary type: either present in the object or absent. The object can be a script or an inscription, and the feature states are the presence or absence of a symbol, a syntactic rule or a layout rule. The measurement scale of a feature is an ordered list of possible values of the feature (presence/absence). This nominal scale is a measurement scale with no meaningful order, and equality is the only relevant relation. The measurable quantity of the nominal scale is treated after a binary-valued quantification, i.e., given a binary variable, a binary feature (a.k.a., presence/absence, alternative) is created [58]. If the number of objects is n, and the number of features is f, then the matching of the objects $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ $(i,j \in \{1,\ldots,n\})$ in binary features can be described by the four values (a,b,c,d) in Table 7, where $a+b+c+d=f$.

The similarity or dissimilarity of objects $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ can be described by various functions (measures). One of them is the Sørensen–Dice coefficient [59], which emphasises the effect of the co-existence of the feature states; see Eq. (1).

$$s_{SD}(\boldsymbol{x_i},\boldsymbol{x_j}) = \frac{2a}{2a+b+c},\ i,j \in \{1,\ldots,n\}, \tag{1}$$

where n is the number of objects to compare. The advantage of the Sørensen–Dice coefficient, which highlights similarities in the presence of features, is appropriate for the dataset in the present study since the absence of a feature in an object (script or inscription) is not characteristic. This is because each object only contains a minority of the features found in the examined scripts, especially when dealing with inscriptions. Hence, the Sørensen–Dice coefficient and its form Eq. (2) expressing the dissimilarity will be used in the following.

$$d_{SD}(\boldsymbol{x_i},\boldsymbol{x_j}) = 1 - s_{SD}(\boldsymbol{x_i},\boldsymbol{x_j}),\ i,j \in \{1,\ldots,n\}, \tag{2}$$

where $d_{SD}(\boldsymbol{x_i},\boldsymbol{x_j})$ is the Sørensen–Dice dissimilarity between objects $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. It is noteworthy that the Sørensen–Dice dissimilarity is not a distance, since it does not satisfy the so-called triangle inequality (3); where d is a distance. Therefore, the Sørensen–Dice dissimilarity is not a metric and that is why it is not called 'distance', only 'dissimilarity'.

Table 7 Contingency table describing the matching probabilities of two objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ $(i,j \in \{1,\ldots,n\})$ in binary features

	Number of feature states with 1 in the object x_j	Number of feature states with 0 in the object x_j
Number of feature states with 1 in the object $\boldsymbol{x_i}$	a (number of feature states with 1 in both objects)	b (number of feature states with 1 in the object $\boldsymbol{x}_i$, and with 0 in the object $\boldsymbol{x_j}$)
Number of feature states with 0 in the object $\boldsymbol{x_i}$	c (number of feature states with 0 in the object $\boldsymbol{x}_i$, and with 1 in the object $\boldsymbol{x_j}$)	d (number of feature states with 0 in both objects)

$$\forall \boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k} : d(\boldsymbol{x_i}, \boldsymbol{x_k}) \leq d\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) + d\left(\boldsymbol{x_j}, \boldsymbol{x_k}\right),\ i, j, k \in \{1, \ldots, n\} \tag{3}$$

It is noteworthy that $\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k}$ $(i, j, k \in \{1, \ldots, n\})$ are vectors of their features, e.g., $\boldsymbol{x_i} = \left[x_i^1, x_i^2, \ldots, x_i^l, \ldots, x_i^f\right]$, $l \in \left\{1, \ldots, f\right\}$, where $f = 119$ is the number of features.

Distance-based phylogenetic inference methods

Given existing phylogenetic inference methods, it is essential to note that an appropriate evolutionary model would be required to apply the advanced and widely used standard methods of maximum likelihood estimation or Bayesian inference in phylogenetic analyses. Such an evolutionary model provides how and with what probability the set of features of one object can be transformed into the set of features of another object, describes the evolutionary processes behind the transformation, and affects the accuracy of the tree construction. However, there needs to be more knowledge to develop such an evolutionary model. Therefore, it is more appropriate to use distance-based phylogenetic inference methods (phenetic approach) that do not require such an evolutionary model.

Phenetic tools, such as cluster analysis, rely on overall similarity rather than evolutionary relationships. To provide a more advanced, cladistic description, knowledge of feature polarity (character polarity) is necessary. Feature polarity refers to the direction of evolution of each feature state, assuming an evolutionary model. However, due to limited knowledge about the development of the examined scripts, it is not possible to make sufficient evolutionary assumptions. Therefore, it is recommended to use a procedure that does not rely on assumptions. The phenetic method has an advantage over cladistics in that it is objective; it always produces the same result based on the same features. In contrast, cladistic analysis distinguishes between features based on their relationship to descent and unique development, resulting in varying outcomes in cladistic studies conducted with different assumptions.

There are several clustering methods to calculate the distance between clusters during the hierarchical cluster analysis. In the present study, phylogenetic inference methods based on dissimilarity matrices, such as the weighted pair group method with arithmetic mean (WPGMA) and neighbour-joining (NJ), are used to find the best phylogenetic trees (Table 3) based on the object-feature dataset under study. Among the various linkage schemes available, the popularity of the WPGMA and NJ in computational phylogenetics justifies their selection; both are agglomerative hierarchical clustering algorithms.

However, additional clustering (linkage) methods are also applied, their results are presented in the Additional file 1, including unweighted pair group method of agglomeration (UPGMA). Both UPGMA and WPGMA classify each taxon into a separate cluster and then gradually merge them, always merging the two nearest clusters. The both algorithms search for pairwise similarity in the dissimilarity matrix and thus build the hierarchical cluster structure agglomerative. They continuously compute a new (one element more minor) similarity matrix by taking the average of the two most similar clusters, thus computing the average distance of one cluster from the other. The process gradually brings the clusters closer together.

The difference between WPGMA and UPGMA is that for UPGMA, in deciding which pair of clusters to merge, a cluster with a larger number of elements is considered with more impact during an intermediate step of the algorithm. In contrast, with WPGMA, clusters with a smaller number of elements are given more weight than the clusters with many elements. In this case, a cluster with a larger number of elements will not have a greater impact than a smaller one on deciding which clusters should be merged in the subsequent step. In scriptinformatics, WPGMA is preferable because it could be that fewer number of scripts evolved in one cluster of scripts, while in another there are more. However, when comparing different clusters of scripts based on their similarity, it is irrelevant to consider the number of scripts in the cluster. This statement also applies to the clustering of inscriptions.

Both UPGMA and WPGMA are effective when the evolutionary process adheres to the evolutionary clock assumption. The biological equivalent of the evolutionary clock is the molecular clock, which states that the number of changes at each site within a molecular sequence should be proportional to time [60]. However, it has not been proven that the evolutionary clock exists in the evolution of scripts. In fact, it is highly likely that the different historical circumstances of the users of each script resulted in varied evolutionary speeds of the scripts. Therefore, WPGMA is favoured over UPGMA. A special type of additive trees obtained as a result of phenetic analysis is the ultrametric tree (refer to Table 3), for which the condition of ultrametricity [61] is fulfilled, see (4), where d is a dissimilarity.

$$\forall \boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k} : d(\boldsymbol{x_i}, \boldsymbol{x_k}) \leq \max \left\{d\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right), d\left(\boldsymbol{x_j}, \boldsymbol{x_k}\right)\right\},\ i, j, k \in \{1, \ldots, n\} \tag{4}$$

Among others, the single linkage (furthest neighbour), the complete linkage (nearest neighbour), the UPGMA, the WPGMA, and the Ward method (minimum variance

linkage) clustering algorithms generate ultrametric trees [62]. In the calculated tree, the measured pairwise dissimilarities, i.e. ultrametrics, differ more or less from the original dissimilarity values. Let $\boldsymbol{D}$ be the matrix of pairwise dissimilarities between objects, and let $\boldsymbol{Z}$ be the matrix of pairwise so-called cophenetic dissimilarities between objects. Matrices are denoted by bold uppercase, italicised letters, while vectors are denoted by bold lowercase, italicised letters. The cophenetic dissimilarity measures the similarity required for two objects to be grouped into the same cluster. The smaller the difference between matrices $\boldsymbol{D}$ and $\boldsymbol{Z}$, the better the inference method. This matrix difference is measured by the cophenetic correlation coefficient (CPCC) [63, 64], which is the Pearson correlation coefficient [50] between the elements of the original $\boldsymbol{D}$ dissimilarity matrix and the $\boldsymbol{Z}$ cophenetic dissimilarity matrix, see Eq. (5).

$$CPCC(\boldsymbol{D},\boldsymbol{Z}) = Cor(\boldsymbol{D},\boldsymbol{Z}) = \sum_{i<j}\left(d_{ij}-\overline{d}\right)\left(d_{ij}^{coph}-\overline{z}\right) \Big/ \sqrt{\sum_{i<j}\left(d_{ij}-\overline{d}\right)^2 \sum_{i<j}\left(d_{ij}^{coph}-\overline{z}\right)^2}, \tag{5}$$

where d_{ij} is an element of the $\boldsymbol{D}$ dissimilarity matrix, and d_{ij}^{coph} is an element of the $\boldsymbol{Z}$ cophenetic dissimilarity matrix; moreover, $\overline{d}$ and $\overline{z}$ are the average of the d_{ij}, and d_{ij}^{coph} values, respectively. *CPCC* is the measure (suitability index) of how well a specific hierarchical clustering (a tree) preserved pairwise dissimilarities between objects. In other words, it measures the amount of distortion the clustering method leading to the tree has imposed upon the system [64]. If *CPCC* is close to 1 (or 100% as a percentage), it is a fit.

The principle of the NJ method for phylogenetic inference [60] is to find pairs of objects—as closest mutual neighbours—that minimise the sum of least square branch lengths (this is called the minimum evolution criterion) at each stage of clustering objects starting from a star-shaped tree. The closest pair of objects is found and merged into a new hypothetical object, and the original pair of objects is deleted until the dissimilarity matrix is reduced to a single object. NJ is based on parsimony (preference for simplicity) [65]; however, NJ does not attempt to obtain the shortest possible tree for an object-feature data matrix. Therefore, NJ does not necessarily lead to a phylogenetic tree with minimal evolution. NJ is a greedy algorithm for optimising a tree since, at each step, it joins the pair of objects that causes the greatest reduction in the estimated tree length using a locally optimal choice.

In the NJ method, the branches of the phylogenetic tree calculated with NJ show the path of transmission of heritable feature information from one object to the next. The NJ tree is an additive tree, in other word, a phylogram (Table 3) since in this tree, the branch lengths are directly related to the extent of genetic change. The longer the branches of a tree, the larger the phylogenetic change (change in heritable feature states) that has occurred. Unlike UPGMA and WPGMA, NJ does not assume that all lineages evolve at the same rate over time. It is therefore suitable for heterotachous evolution, where heterotachy refers to differences in lineage-specific evolutionary rates.

The principal component analysis (PCA) ordination

Ordination is a data mining technique that represents similarity relationships in a few dimensions; its goal is to reduce the dimensionality of large data structures with the most minor loss of information. Ordination extracts artificial variables to reduce the dimensionality of the original feature set of objects. Principal components analysis (PCA) is a statistical method for ordinal data analysis [50], which is widely used in data mining [57]. The input is an object-feature matrix of multivariate data. The purpose of PCA is to show how the different variables (in this case, features) change about each other and how they are related. This is done by transforming the correlated original variables into a new set of uncorrelated underlying variables using the variance–covariance matrix. PCA finds hypothetical variables (components) that account for as much variance in multivariate data as possible. The prerequisite for its use is that all variables are quantitative.

The principal components are eigenvectors of the variance–covariance matrix of the data. PCA involves the calculation of eigenvalues and their corresponding eigenvectors. The principal components are linear combinations of the original variables and are ranked in descending order of how much variance they account for in the original set of variables. Taken together, all principal components account for 100% of the variation. PCA can be thought of as exploring the internal structure of the data in a way that best explains the scatter in the dataset. The proportion of variances represented by every eigenvector can be determined by dividing the eigenvalue of the eigenvector by the total sum of all eigenvalues.

New algorithm

The holophyletic index (HI) and the joint holophyletic index (JHI)

The inscriptions were subjected to hierarchical cluster analysis, resulting in a phylogenetic tree with objects as leaves. Each subtree of the tree represents a holophyletic group if all its leaves belong to the same script. A holophyletic group includes all descendants of the group's most recent common ancestor [59]. This method allows for determining the proportion of objects belonging to a group in a phylogenetic tree that are included in a subtree. The subtree is considered homogeneous if all its leaves are objects belonging to the same group.

If the phylogenetic tree is created with WPGMA, it is an ultrametric tree; if created with NJ, it is an additive tree (Table 3). The l_i leaves of the phylogenetic tree are the objects included in the study, denote their set $\boldsymbol{l}$; their number is $|\boldsymbol{l}| = n$ (the number of objects), and $i \in \{1, \ldots, n\}$. The intermediate vertex of the phylogenetic tree v_j $(j = 1, 2, \ldots)$ represent objects, denote their set by the vector $\boldsymbol{v}$.

Let the n objects be a priori grouped into different classes. E.g. if objects are inscriptions, then the classes of inscriptions are a priori classified into different scripts. Let the set of classes be $\boldsymbol{c}$, denote the number of classes by $|\boldsymbol{c}|$. Represent each class by $\boldsymbol{c_p}$ and $\boldsymbol{c_q}$, where $p, q \in \{1, \ldots, |\boldsymbol{c}|\}$. Suppose that $\boldsymbol{c_p} \cap \boldsymbol{c_q} = \varnothing$ for any $\boldsymbol{c_p}, \boldsymbol{c_q} \in \boldsymbol{c}$, and $\sum_{p=1}^{|\boldsymbol{c}|} |\boldsymbol{c_p}| = |\boldsymbol{l}|$, i.e. the classification is exclusive and complete. As above, (6) is true.

$$|\boldsymbol{l}| = \sum_{p=1}^{|\boldsymbol{c}|} |\boldsymbol{c_p}| = n \tag{6}$$

Let $\boldsymbol{l}^r(\boldsymbol{c_p}) \subseteq \boldsymbol{l}(\boldsymbol{c_p})$ be a holophyletic subset of the set $\boldsymbol{l}(\boldsymbol{c_p}) \subseteq \boldsymbol{l}$ of leaf objects that belong to the class $\boldsymbol{c_p}$, $p \in \{1, \ldots, |\boldsymbol{c}|\}$, and $r = 1, 2, \ldots$, the number of disjoint $\boldsymbol{l}^r(\boldsymbol{c_p})$ holophyletic groups of a specific $\boldsymbol{c_p}$ class. The $\boldsymbol{l}^r(\boldsymbol{c_p})$ forms a *holophyletic group* for a class $\boldsymbol{c_p}$ if for $\forall l_i \in \boldsymbol{l}^r(\boldsymbol{c_p})$ leaf (object) it is true that $\exists v_j \in \boldsymbol{v}$ intermediate node in the tree, from which $\forall l_i \in \boldsymbol{l}^r(\boldsymbol{c_p})$ leaves are less dissimilar than any other $l_j \notin \boldsymbol{l}^r(\boldsymbol{c_p})$ leaf of the tree, $i, j \in \{1, \ldots, n\}$. Let $|\boldsymbol{l}^r(\boldsymbol{c_p})|$ denote the multiplicity of the set $\boldsymbol{l}^r(\boldsymbol{c_p})$. The dissimilarity between two nodes (objects) is measured by an appropriate measure, which in the present research is the Sørensen–Dice dissimilarity (2).

Let's introduce the HI_{c_p} *holophyletic index* measure. HI_{c_p} is based on a fraction of objects of class $\boldsymbol{c_p}$ in a phylogenetic tree that belongs to one $(r = 1)$ or more $(r > 1)$ disjoint $\boldsymbol{l}^r(\boldsymbol{c_p})$ holophyletic groups of a specific $\boldsymbol{c_p}$ class, see (7).

$$HI_{c_p} = \frac{\sum_r |\boldsymbol{l}^r(\boldsymbol{c_p})|}{|\boldsymbol{c_p}|};\ \boldsymbol{l}^r(\boldsymbol{c_p}) \cap \boldsymbol{l}^s(\boldsymbol{c_p}) = \varnothing,\quad \text{if } r \neq s;\ r, s = 1, 2, \ldots;\ p \in \{1, \ldots, |\boldsymbol{c}|\} \tag{7}$$

Given the classification of objects into classes $\boldsymbol{c_p} \in \boldsymbol{c}$ $(p \in \{1, \ldots, |\boldsymbol{c}|\})$, the *JHI joint holophyletic index* represents a fraction of the set of leaves l_i of a phylogenetic tree that belongs to one of the holophyletic groups $\boldsymbol{l}^r(\boldsymbol{c_p})$ of any $\boldsymbol{c_p}$ class, see (8).

$$JHI = \frac{\sum_{p=1}^{|\boldsymbol{c}|} \sum_r |\boldsymbol{l}^r(\boldsymbol{c_p})|}{n} = \frac{\sum_{p=1}^{|\boldsymbol{c}|} (HI_{c_p} \cdot |\boldsymbol{c_p}|)}{n} \tag{8}$$

The HI_{c_p} holophyletic index and the *JHI* joint holophyletic index describe the separation of objects forming a phylogenetic tree according to classification $\boldsymbol{c}$. This is true for the set of values of these indices: $HI_{c_p}, JHI \in [0, 1]$, $(p \in \{1, \ldots, |\boldsymbol{c}|\})$. If the value HI_{c_p} or *JHI* is close to 1, then this fact supports the a priori (traditional) classification of l_i objects $(i = 1, \ldots, n)$ into the classes $\boldsymbol{c_p} \in \boldsymbol{c}$ $(p \in \{1, \ldots, |\boldsymbol{c}|\})$.

General flow of the algorithm

The flow chart of the developed composite phylogenetic analysis is presented in Fig. 3. Its main purpose is to validate the traditional classification of the inscriptions and explore their phylogenetic relations.

From the comparison of the different multivariate analysis methods, conclusions can be drawn about the

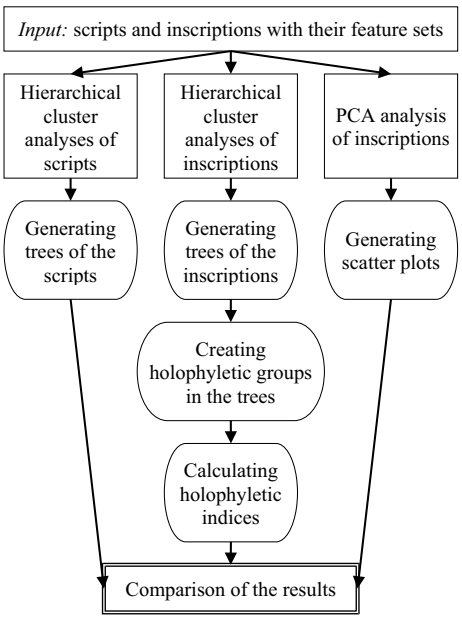

Fig. 3 Main steps of the composite phylogenetic analysis of scripts and inscriptions

Table 8 Actual parameter values for the present study

Description of the parameter	Value
Number of objects (inscriptions)	$n = 57$
Number of features	$f = 119$
Number of classes (inscriptions classified into scripts)	$\|\boldsymbol{c}\| = 4$
Names of the classes	$\boldsymbol{c}_1 = \boldsymbol{c}_{TR}$, $\boldsymbol{c}_2 = \boldsymbol{c}_{SHR}$, $\boldsymbol{c}_3 = \boldsymbol{c}_{CBR}$, $\boldsymbol{c}_4 = \boldsymbol{c}_{SR}$
Multiplicity of sets $\boldsymbol{c}_p, p \in \{1, \ldots, \|\boldsymbol{c}\|\}$	$\|\boldsymbol{c}_1\| = 16$, $\|\boldsymbol{c}_2\| = 19$, $\|\boldsymbol{c}_3\| = 7$, $\|\boldsymbol{c}_4\| = 15$

correctness of classifying the given groups of inscriptions into scripts.

Results

Common data of the different tests

Hierarchical cluster analysis and PCA were performed using Matlab R2023b software [66]. The Sørensen-Dice dissimilarity measure (2) was applied, along with additional measures. The results obtained with these measures can be found in the Additional file 1. Table 8 presents the values of the parameters used in the applied methods.

Trees of the scripts

The results (phylogenetic trees) of the clustering being interpreted as phylogenetic trees describe the similarities between the four taxa (TR, SHR, CBR and SR). The trees calculated by WPGMA and NJ are presented in Fig. 4.

A comparison of trees in Fig. 4 reveals that both cases' CBR and SR scripts are close.

Trees of the inscriptions

The calculations of phylogenetic trees of $n = 57$ inscriptions as objects were performed by WPGMA and NJ algorithms. The results are presented in Figs. 5 and 6, where the holophyletic groups are highlighted. Each holophyletic group is represented by the following colours: TR: blue, SHR: green, CBR: orange and SR: violet.

The evaluation of the hierarchical clustering is presented in Table 9, where the cophenetic correlation coefficient (CPCC) is based on Eq. (5), the holophyletic indices for each classification of the objects based on (7) and the joint holophyletic index based on Eq. (8) are calculated using data in Table 8.

Ordination of the inscriptions

The PCA function in Matlab centres the data and uses the singular value decomposition (SVD) algorithm. The 2-dimensional PCA ordering yielded Fig. 7. In the PCA analysis, the input data is a matrix of objects (inscriptions)—features (variables) with objects in rows and features in columns. The principal components scores are the representations of the object–feature data matrix in the principal component space. The analysis employed the variance–covariance matrix since all variables were measured in the same units. The variables were centred but not normalised. The PCA ordering calculated the eigenvalues and eigenvectors of the variance–covariance matrix.

The three principal components with the largest variances are presented in Table 10.

Table 10 displays the principal component variances, which are the eigenvalues of the variance–covariance matrix of the input object-feature data matrix, in the Eigenvalue column. The Variance column presents the percentage of the

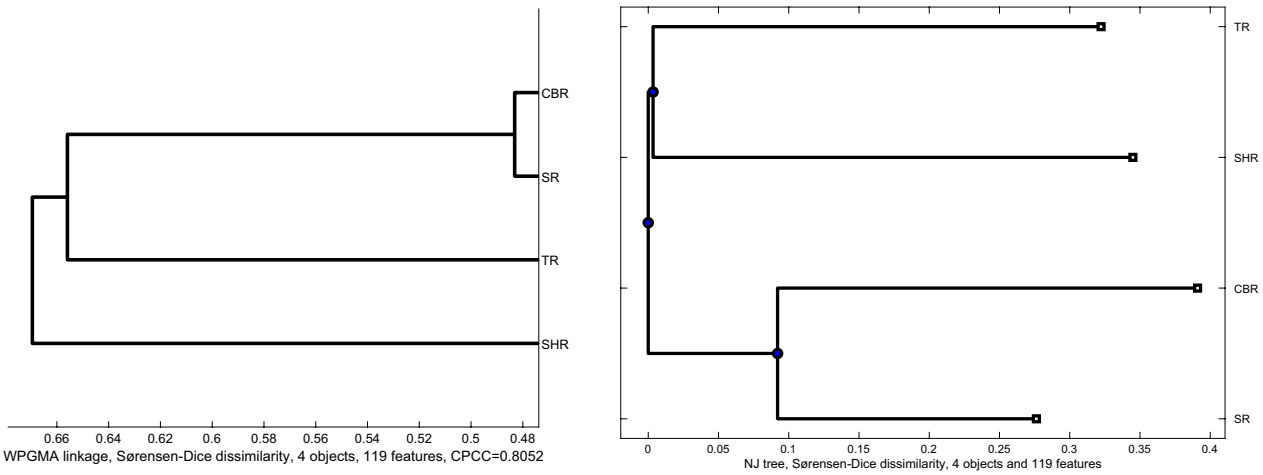

Fig. 4 The trees of the Rovash scripts as taxa (objects) calculated by WPGMA (an ultrametric tree, left) and by NJ (an additive tree, right)

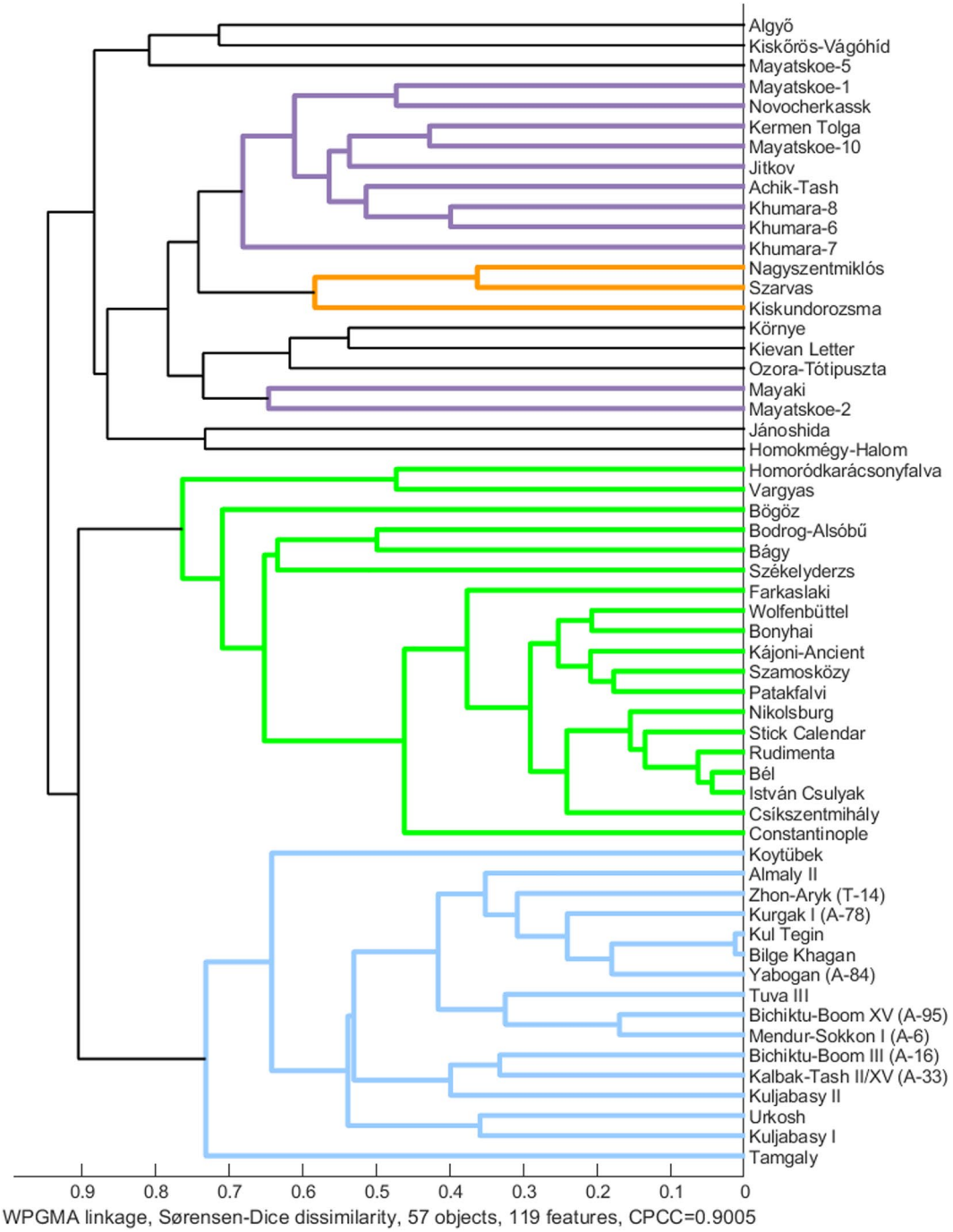

Fig. 5 The WPGMA (ultrametric) tree of the examined Rovash inscriptions as objects with highlighting the holophyletic groups with different colours

total variance explained by each principal component. These results justify the use of a 3D scatter plot, as shown in Fig. 8 (with identical colour codes to Fig. 7).

Figure 8 shows that the third principal component is not useful in distinguishing between the CBR and SR inscriptions. Apart from the three principal components, the remaining components represent a very small proportion of the total variance and can therefore be disregarded.

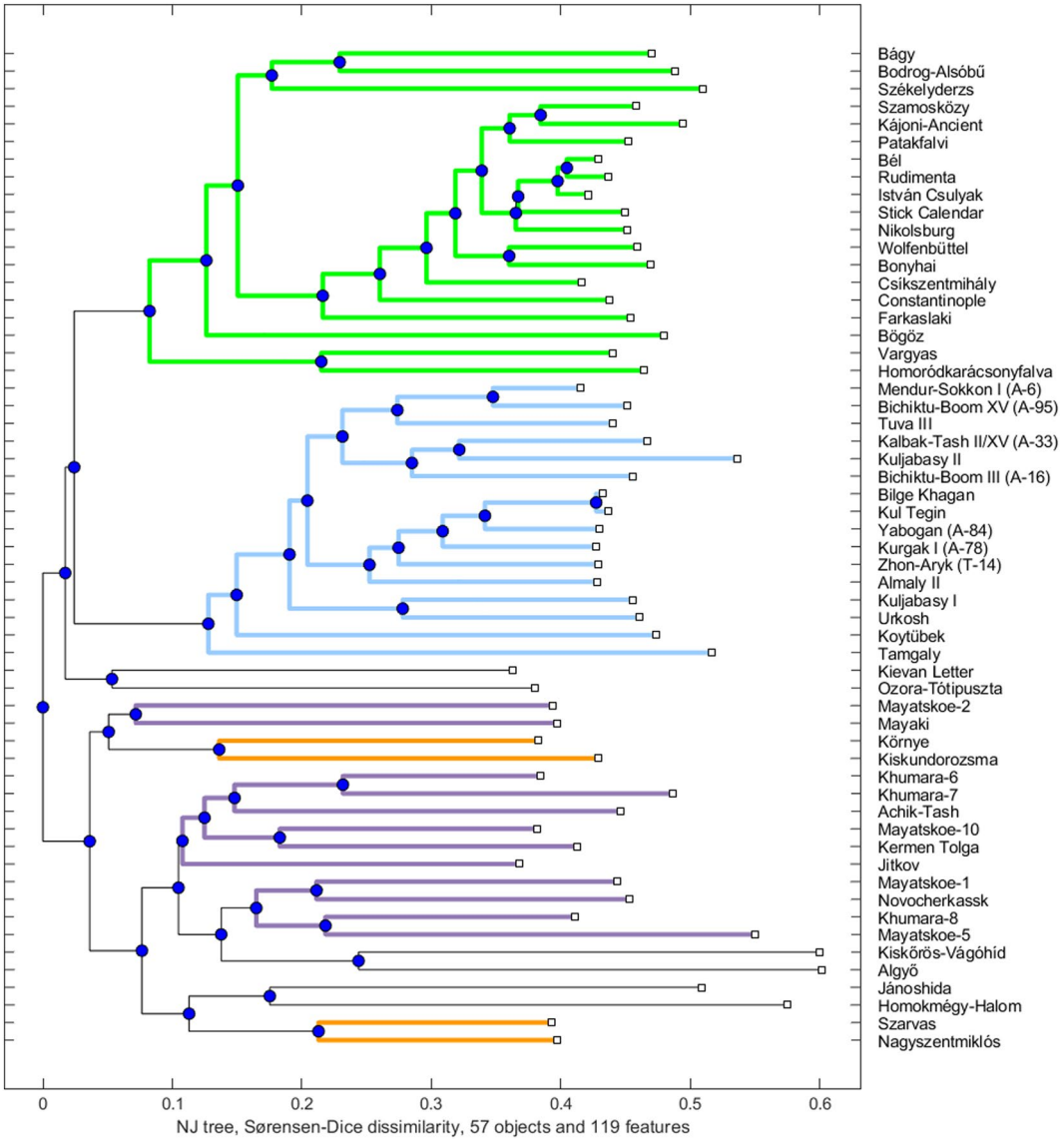

Fig. 6 The NJ (additive) tree of the examined Rovash inscriptions as objects with highlighting the holophyletic groups with different colours

Discussion and conclusions

Pattern evolution research examines the relationships between pattern systems and studies their evolution. Scriptinformatics, a specific area of pattern evolution research, focuses on the evolution of special pattern systems, namely scripts or writing systems. Scriptinformatics, as well as pattern evolution in general, employs methods of modelling evolutionary processes in

Table 9 Qualifying the calculated trees resulting from hierarchical clustering using Sørensen–Dice dissimilarity

Tree (linkage with Sørensen–Dice dissimilarity)	*CPCC*	HI_{TR}	HI_{SHR}	HI_{CBR}	HI_{SR}	*JHI*
Tree of scripts by WPGMA	0.8052	–	–	–	–	–
Tree of scripts by NJ	–	–	–	–	–	–
Tree of inscriptions by WPGMA	0.9005	1	1	0.43	0.73	0.86
Tree of inscriptions by NJ	–	1	1	0.57	0.80	0.89

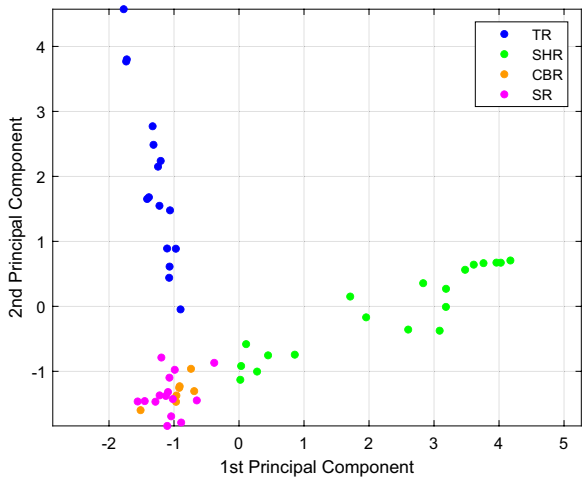

Fig. 7 The 2D scatter plot of the ordination of the Rovash inscriptions as objects

Table 10 Principal components with the largest variances calculated in PCA

Principal component	Eigenvalue	Variance [%]
1st	3.4590	25.4279
2nd	2.3994	17.6381
3rd	0.8458	6.2177

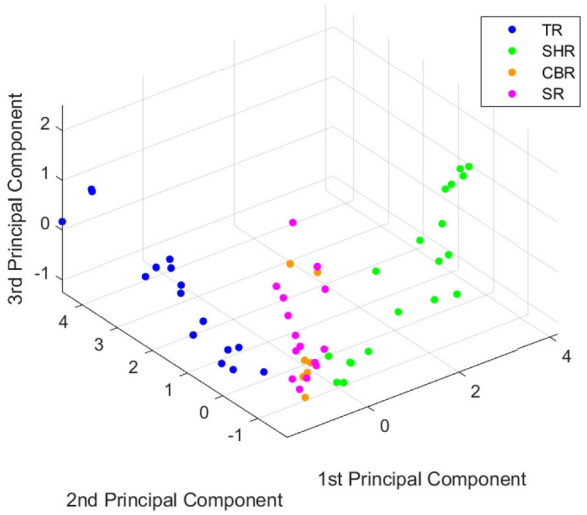

Fig. 8 The 3D scatter plot of the PCA result

phylogenetics to analyse the development of scripts and other pattern systems. The use of scripts typically results in graph sequences, referred to as inscriptions in scriptinformatics. In the case of historical inscriptions, they may be the only evidence for a script that has long been disused and forgotten, in other words, extinct. In such cases, the features that define a particular script can be determined based on the deciphered inscriptions that have survived. In cases like these, knowledge of the scripts may be incomplete, depending on the number and length of the surviving inscriptions. Conversely, the reverse situation can also be problematic. For instance, when new inscriptions are found, it can be challenging to determine which of the associated scripts they belong to. This is an important issue because each newly discovered inscription helps to clarify the features of the script used to create it.

The aim of this study was to assess the accuracy of classifying $n = 57$ inscriptions from a selected family of scripts, namely the Rovash scripts. To achieve this, multivariate methods were used to analyse the inscriptions and determine if the resulting groups corresponded to one of the four Rovash scripts (TR, SHR, CBR and SR).

This article presents the analysis of the relationships between the four Rovash scripts using various dissimilarity types and linkage methods. The results of the Sørensen-Dice dissimilarity-based neighbour-joining (NJ) and weighted pair group method with arithmetic mean (WPGMA) linkage methods are discussed. The hierarchical cluster analysis revealed that CBR and SR are the closest taxa, while the relationship between TR and SHR remains uncertain (Fig. 4). It is important to note that the hierarchical cluster analysis is a type of phylogenetic examination, but it does not explore the evolutionary relationships of the taxa (objects). However, the resulting structure of the objects, in this case, scripts, can be treated as a kind of phylogenetic tree of these objects.

The article presents a new algorithm that validates the traditional classification of inscriptions made with ancient scripts or script variants. The algorithm uses an evolutionary approach to assign individual scripts as classes. The study examines the descent of the script variants used to create individual inscriptions and reconstructs a phylogenetic tree of the inscriptions. During this process, the inscriptions were clustered hierarchically using WPGMA and NJ. Additional calculations are provided in the Additional file 1. A newly introduced measure, the holophyletic index, was used to evaluate the resulting phylogenetic trees. The holophyletic index represents the ratio of the number of objects belonging to a holophyletic group in the phylogenetic tree to the total number of objects. The holophyletic index can be calculated for inscriptions that belong to a particular script, resulting in a *script-specific holophyletic index* (*HI*). Additionally, it can be computed for inscriptions belonging to all examined scripts, resulting in a *joint holophyletic index* (*JHI*). Table 9 displays the results, indicating that for TR and SHR, the *HI*s (script-specific holophyletic

indices) are equal to 1. This confirms the correctness of the traditional classification of the inscriptions associated with these scripts. However, for CBR and SR, WPGMA and NJ linkages showed noticeably lower *HI* values. Both WPGMA and NJ produced similar results in their main characteristics. The WPGMA clusterings were also evaluated using the cophenetic correlation coefficient (*CPCC*). The obtained *CPCC* values were 0.8052 for scripts and 0.9005 for inscriptions (see Table 9), indicating excellent quality of the WPGMA clusterings. The Additional file 1 presents the trees constructed using additional linkage methods and dissimilarity measures. The *HI*, *JHI*, and *CPCC* values were determined from these trees, which were found to be similar to those reported in the article.

A principal component analysis (PCA) was applied to the 57×119 object-feature data matrix of the inscriptions to extract uncorrelated underlying variables. The results show a clear division between TR and SHR inscriptions as objects, while CBR and SR objects remain largely undifferentiated. This suggests that the classification of inscriptions as TR or SHR is precise and well-established, while the separation of inscriptions belonging to CBR and SR scripts is uncertain. This suggests that CBR and SR scripts are more closely related than TR and SHR.

The study's findings are consistent with previous assumptions. Figures 4, 5, and 6 demonstrate a relationship between the TR and SHR scripts. However, it should be pointed out that the inscriptions of SHR and TR evolved separately, based on the knowledge of the scribes who created each one. The earliest known SHR inscription dates back to the tenth century in the Carpathian Basin, while the latest TR inscriptions were found in Inner Asia and also belong to the tenth century. Therefore, it is not reasonable to expect that these inscriptions will reveal the common history of SHR and TR. Any earlier SHR inscriptions that are deciphered may shed light on this matter.

Rovash paleographers generally accept that the CBR and SR scripts are closely related. However, the test results do not provide sufficient evidence to conclude that the inscriptions previously classified as CBR and SR belong to the same script. It is possible to divide both CBR and SR scripts into script varieties. Furthermore, a script variety in CBR may be closely related or even identical to another in SR. In summary, this method facilitates the identification of subgroups within deciphered inscriptions as more inscriptions become known. Figures 7 and 8 illustrate the separate distribution of TR and SHR inscriptions, which have developed independently.

The Carpathian Basin was often the final destination for steppe peoples migrating from the East. The earliest inscriptions in this region do not exhibit a uniform writing culture. Instead, they resemble the script used in ancient Greece. This suggests that, like the ancient Greeks, different scripts were used in each polis, but they were connected to each other through the horizontal transmission of features. The research did not consider the possibility of horizontal transmission due to the limited number of Rovash inscriptions. Therefore, the accuracy of the results is naturally limited. If more inscriptions become available, further refinement of this phylogenetic modelling will yield more accurate results, possibly leading to the creation of a phylogenetic network.

The composite phylogenetic analysis results indicate deficiencies in conventional inscription classifications. To make progress, two actions are necessary. Firstly, it is essential to re-evaluate the features that define the scripts. Secondly, the database should be expanded by deciphering the inscriptions that archaeologists have discovered but remain undeciphered. This will enhance the resolution of the analysis and assist archaeologists in identifying the scripts used for the undeciphered inscriptions.

The composite phylogenetic analysis can be applied to any script group, not just Rovash scripts. It helps to determine whether the a priori classification of the tested inscriptions into each script is correct. The method can be generalized to include pattern systems with evolutionary properties and graph sequences formed from them, in addition to scripts and inscriptions, respectively.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40494-024-01211-7.

Additional file 1. Application of various hierarchical clustering methods to the inscriptions using different dissimilarity measures.

Author contributions

The manuscript was authored by Gábor Hosszú.

Funding

Open access funding provided by Budapest University of Technology and Economics.

Data availability

The dataset supporting the conclusions of this article developed data processing software (Matlab application) are available in the GitHub repository, at https://github.com/hosszu/2024_Multivariate_analysis_of_Rovash_inscriptions/ [67].

Declarations

Competing interests

The author declares that he has no conflict of interest.

Received: 18 December 2023 Accepted: 13 March 2024
Published online: 03 April 2024

References

1. Hosszú G. Mathematical statistical examinations on script relics. In: Bhatnagar V, editor. Data mining and analysis in the engineering field. Hershey: Information Science Reference; 2014. p. 142–58. https://doi.org/10.4018/978-1-4666-6086-1.ch008.
2. Hosszú G. A novel computerized paleographical method for determining the evolution of graphemes. In: Khosrow-Pour M, editor. Encyclopedia of information science and technology. Hershey: Information Science Reference; 2015. p. 2017–31. https://doi.org/10.4018/978-1-4666-5888-2.ch194.
3. Hosszú G. Scriptinformatics. Extended phenetic approach to script evolution. Budapest: Nap; 2021. http://napkiado.hu/media/Hosszu_Gabor-Scriptinformatics.pdf. Accessed 9 February 2021.
4. Hosszú G. Data-driven phenetic modeling of scripts' evolution. In: Liu S, Bohács G, Shi X, Shang X, Huang A, (Eds). Proc. 10th Int. Conf. Logistics, Informatics and Service Sciences, LISS 2020. Springer; 2021. p. 389–403; https://doi.org/10.1007/978-981-33-4359-7_28.
5. Tóth LL, Hosszú G. A new topological method for examining historical inscriptions. J Inf Technol Res. 2019;12:1–16. https://doi.org/10.4018/JITR.2019040101.
6. Hosszú G. Phenetic approach to script evolution. In: Busch H, Fischer F, Sahle P (eds). Kodikologie und Paläographie im digitalen Zeitalter 4. Codicology and palaeography in the digital age 4. Norderstedt: Books on Demand; 2017. p. 179–252.
7. Rezende EL, Diniz-Filho JAF. Phylogenetic analyses: comparing species to infer adaptations and physiological mechanisms. Compr Physiol. 2012;2:639–74. https://doi.org/10.1002/cphy.c100079.
8. Nakhleh L, Warnow T, Ringe D, Evans SN. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. Trans Philol Soc. 2005;103:171–92. https://doi.org/10.1111/j.1467-968X.2005.00149.x.
9. Deepak A, Fernández-Baca D, Tirthapura S, Sanderson MJ, McMahon MM. EvoMiner: frequent subtree mining in phylogenetic databases. Knowl Inf Syst. 2014;41:559–90. https://doi.org/10.1007/s10115-013-0676-0.
10. Phillips-Rodríguez WJ. The evolution of a Sanskrit epic: Some genetic considerations about scripts. In: The evolution of texts: Confronting stemmatological and genetical methods. Istituti editoriali e poligrafici internazionali; 2006. p. 175–90.
11. Phillips-Rodríguez WJ. Scripts and manuscripts: Two independent speciation processes in the Mahābhārata textual tradition. In: The churning of the epics and purāṇas: Proceedings of the Epics and Purāṇas Section at the 15th World Sanskrit Conference. Dev Publishers & Distributors; 2018. p. 1–13.
12. Biolcati V, Wooley J, Lévêque É, Rossi A, Hoffmann AG, Visentin A, Macháin Ó, P, Iacopino D. Establishing the original order of the poems in Harward's Almanac using paleography, codicology, X-ray fluorescence spectroscopy, and statistical analysis. Herit Sci. 2023;11:265. https://doi.org/10.1186/s40494-023-01107-y.
13. Ciula A. Digital palaeography: using the digital representation of medieval script to support palaeographic analysis. Digit Mediev. 2005. https://doi.org/10.16995/dm.4.
14. Ciula A. The palaeographical method under the light of a digital approach. In: Rehbein M, Sahle P, Schaßen T, editors. Kodikologie und Paläographie im digitalen Zeitalter 1—codicology and palaeography in the digital age 1. Norderstedt: Books on Demand; 2009. p. 219–35.
15. Azmi MS, Omar K, Nasrudin MF, Muda AK, Abdullah A. Digital Paleography: Using the Digital Representation of Jawi Manuscripts to Support Paleographic Analysis. In: 2011 International Conference on Pattern Analysis and Intelligent Robotics, 28–29 June 2011, Putrajaya, Malaysia; 2011. p. 71–7.
16. Levy N, Wolf L, Dershowitz N, Stokes P. Estimating the distinctiveness of graphemes and allographs in paleographic classification. In: Levy N, editor. Proceedings of Digital Humanities DH 2012. Hamburg: Hamburg University Press; 2012. p. 264–7.
17. Wolf L, Potikha L, Dershowitz N, Shweka R, Choueka Y. Computerized Paleography: Tools for Historical Manuscripts. In: 18th IEEE International Conference on Image Processing (ICIP). Brussels (Belgium); 2011. p. 3545–8.
18. Stokes PA. Computer-Aided Palaeography, Present and Future. In: Rehbein M, Sahle P, Schaßen T, editors. Kodikologie und Paläographie im digitalen Zeitalter 1—Codicology and palaeography in the digital age 1. Norderstedt: Books on Demand; 2009. p. 309–38.
19. Hassner T, Sablatnig R, Stutzmann D, Tarte S. Digital palaeography: new machines and old texts (Dagstuhl Seminar 14302). Dagstuhl Rep. 2014;4(7):112–34.
20. Aussems M, Brink A. Digital Palaeography. In: Rehbein M, Sahle P, Schaßen T, editors. Kodikologie und Paläographie im digitalen Zeitalter 1—codicology and palaeography in the digital age 1. Norderstedt: Books on Demand; 2009. p. 293–308.
21. Gelb I. A Study of Writing. Chicago: University of Chicago Press; 1952.
22. Blatner A. Commentary: a call for "scriptology." Vis Lang. 1989;23:415.
23. Buneman P. The recovery of trees from measures of dissimilarity. In: Hodson FR, Kendall DG, Tăutu P, editors. Mathematics in the Archaeological and Historical Sciences. Edinburgh: Edinburgh University Press; 1971. p. 387–95.
24. Platnick NI, Cameron HD. Cladistic methods in textual, linguistic, and phylogenetic analysis. Syst Zool. 1977;26:380–5.
25. Reeve MD. Shared innovations, dichotomies, and evolution. In: Ferrari A, editor. Filologia classica e filologia romanza: esperienze ecdotiche a confronto: Atti del Convegno Roma 25–27 maggio 1995. Spoleto: Centro Italiano di Studi sull'Alto Medioevo; 1998. p. 445–505.
26. Mabee C. The American Leonardo: A life of Samuel F. B. Morse. New York: Purple Mountain Press; 2000. https://doi.org/10.1017/S0022050700081341.
27. Hennig W. Phylogenetic systematics. Urbana (IL): University of Illinois Press; 1966.
28. Podani J, Morrison DA. Categorizing ideas about systematics: alternative trees of trees and related representations. Rendiconti Lincei Scienze Fisiche e Naturali. 2017;28:191–202.
29. Michener CD, Sokal RR. A quantitative approach to a problem of classification. Evolution. 1957;11:490–9.
30. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. 1958;38:1409–38.
31. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–25. https://doi.org/10.1093/oxfordjournals.molbev.a040454.
32. Ward JH Jr. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58:236–44. https://doi.org/10.1080/01621459.1963.10500845.
33. Hosszú G. The rovas: a special script family of the central and eastern European languages. Acta Philologica. 2013;44:91–102.
34. Looijenga T. Runes around the north sea and on the continent AD 150–700; texts & contexts. Groningen: SSG Uitgeverij; 1997.
35. Jansson SBF. Runes in Sweden. Stockholm: Gidlund; 1997.
36. Looijenga T. Texts and contexts of the oldest runic inscriptions. Leiden, Boston: Brill; 2003.
37. Barnes MP. Runes A Handbook. Woodbridge: The Boydell Press; 2012.
38. Erdal M. A grammar of Old Turkic Handbook of Oriental Studies Central Asia. Leiden: Koninklijek Brill; 2004.
39. Vékony G. Későnépvándorláskori rovásfeliratok a Kárpát-medencében [Rovash inscriptions from the Late Migration Period in the Carpathian Basin]. Szombathely, Budapest: Életünk szerkesztősége; 1987 (in Hungarian).
40. Vékony G. A székely írás emlékei, kapcsolatai, története [Relics, relationships and the history of the Szekely script]. Budapest: Nap; 2004 (in Hungarian).
41. Vékony G. A Bodrog-Alsóbűi felirat [The Bodrog-Alsóbű inscription]. Somogyi Múzeumok Közleményei. 2000;14:219–25.
42. Gömöri J. Az avar kori és X-XI. századi vaskohászat régészeti emlékei Somogy megyében [Archaeological monuments of iron metallurgy in Somogy county from the Avar age and the 10th–11th centuries]. Somogyi Múzeumok Közleményei. 2000;14:163–218.
43. Hosszú G. Heritage of Scribes. The Relation of Rovas Scripts to Eurasian Writing Systems, Budapest: Rovas Foundation. 2013. https://google.hu/books?id=TyK8azCqC34C&pg. Accessed 11 March 2024.
44. Konkobaev K, Useev N, Šabdanaliev N. [Конкобаев К, Усеев Н, Шабданалиев Н] Atlas of ancient Turkic written monuments of the

Altai Republic [Атлас древнетюркских письменных памятников Республики Алтай]. Astana: Ғылым; 2015.
45. Hosszú G. The appearance of Székely law in a Rovash relic, MSc Thesis, Budapest: Pázmány Péter Catholic University, Faculty of Law and Political Science; 2010. https://www.academia.edu/2256595. Accessed 12 March 2024.
46. Arauzo-Azofra A, Jiménez-Vílchez A, Molina-Baena J, Luque-Rodriguez M. Algorithmic cache of sorted tables for feature selection. Speeding up methods on consistency and information theory measures. Data Min Knowl Discov. 2019;33:964–94. https://doi.org/10.1007/s10618-019-00620-8.
47. Xu KS, Kliger M, Hero AO. Adaptive evolutionary clustering. Data Min Knowl Discov. 2014;28:304–36. https://doi.org/10.48550/arXiv.1104.1990.
48. Zhang W, Li R, Feng D, Chernikov A, Chrisochoides N, Osgood C, Ji S. Evolutionary soft co-clustering: formulations, algorithms, and applications. Data Min Knowl Discov. 2015;29:765–91. https://doi.org/10.1007/s10618-014-0375-9.
49. Rizoiu M-A, Velcin J, Bonnevay S, Lallich S. ClusPath: a temporal-driven clustering to infer typical evolution paths. Data Min Knowl Discov. 2016;30:1324–49. https://doi.org/10.48550/arXiv.1512.03501.
50. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Burlington: Morgan Kaufmann; 2011. https://doi.org/10.1016/C2009-0-61819-5.
51. Daggumati S, Revesz PZ. Data mining ancient scripts to investigate their relationships and origins. In: Proc. 23rd Int. Database Applications & Engineering Symp, IDEAS'19. ACM; 2019. p. 1–10; doi:https://doi.org/10.1145/3331076.3331116.
52. Marchesini S. The Elymian language. In: Tribulato O, editor. Language and linguistic contact in ancient Sicily. Cambridge: Cambridge University Press; 2012. p. 95–114. https://doi.org/10.1017/CBO9781139248938.005.
53. Maggi M. Some remarks on the history of the Khotanese orthography and the Brāhmī script in Khotan. In: Kudo N, editor. Annual report of the international research institute for advanced Buddhology at Soka University for the academic year 2021, vol. XXV. Aliso Viejo: Soka University; 2022. p. 149–72.
54. Stutzmann D, Tensmeyer C, Christlein V. Writer identification and script classification: two tasks for a common understanding of cultural heritage. Manuscr Cult. 2020;15:11–24.
55. Christlein V, Marthot-Santaniello I, Mayr M, Nicolaou A, Seuret M. Writer retrieval and writer identification in Greek papyri. In: Carmona-Duarte C, Diaz M, Ferrer MA, Morales A, editors. Intertwining Graphonomics with Human Movements. Berlin: Springer; 2022. p. 76–89. https://doi.org/10.1007/978-3-031-19745-1_6.
56. Metzner C, Schilling A, Traxdorf M, Tziridis K, Maier A, Schulze H, Krauss P. Classification at the accuracy limit: facing the problem of data ambiguity. Sci Rep. 2022;12:22121. https://doi.org/10.1038/s41598-022-26498-z.
57. Nakao EK, Levada ALM. Entropic principal component analysis using Cauchy-Schwarz divergence. Knowl Inf Syst. 2023;65:5375–85. https://doi.org/10.21203/rs.3.rs-1499062/v1.
58. Tan PN, Steinbach M, Kumar V. Introduction to data mining. London: Pearson; 2018.
59. Podani J. Introduction to the exploration of multivariate biological data. Leiden: Backhuys; 2000.
60. Warnow T. Computational phylogenetics. An introduction to designing methods for phylogeny estimation. Cambridge: Cambridge University Press; 2017. https://doi.org/10.1017/9781316882313.
61. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967;32:241–54. https://doi.org/10.1007/BF02289588.
62. Milligan GW. Ultrametric hierarchical clustering algorithms. Psychometrika. 1979;44:343–6.
63. Sokal RR, Rohlf FJ. The comparison of dendrograms by objective methods. Taxon. 1962;11:33–40. https://doi.org/10.2307/1217208.
64. Wheeler WC. Systematics: a course of lectures. Hoboken: Wiley-Blackwell; 2012.
65. Sober E. Ockam's razor: a user's manual. Cambridge: Cambridge University Press; 2015. https://doi.org/10.1017/CBO9781107705937.
66. The MathWorks, Inc. MATLAB version: 23.2.0.2515942 (R2023b); 2024. https://www.mathworks.com. Accessed 27 February 2024.
67. Hosszú G. Multivariate analysis of rovash inscriptions. Dataset on Github. 2023. https://github.com/hosszu/2024_Multivariate_analysis_of_Rovash_inscriptions. Accessed 12 March 2024.

Publisher's Note