Topological Analysis of Ancient Glyphs

Gábor Hosszú*, Ferenc Kovács**
* Budapest University of Technology and Economics, Budapest, Hungary
** Pázmány Péter Catholic University, Budapest, Hungary
hosszu@eet.bme.hu, kovacsf@itk.ppke.hu

***Abstract*—This paper presents a machine learning approach to explore the phenetic relations of historical scripts and their glyphs. Its first step is the identification of the observable topological transformations in the development of the glyphs, and with the use of these transformations, the method collects the possible cognate glyphs by minimizing the necessary topological transformations between the glyphs. In these investigations, the phonetic properties of the graphemes were consistently considered. The second step of our method is selecting similarity groups of possible cognate glyphs by minimizing the differences of their topological properties. The third step is multidimensional scaling and different cluster analyses based on the similarity groups of the glyphs of the historical scripts in order to explore the phenetic relationships between these scripts. The resulting phenetic structure of the scripts could be used for paleographical research, especially in deciphering ancient hard-to-read inscriptions.**

I. Introduction

Archaeological objects with inscriptions are discovered from time to time. Their significant part often remains undeciphered, especially if they are short, and hence, the *script* (writing system) and the language of the text remains typically unknown. Nowadays, however, the knowledge base of the paleography has dramatically increased and therefore, a need has arisen to apply mathematical algorithms to analyze the vast amounts of known data, which may assist the paleographers in deciphering the hard-to-read inscriptions. This paper details an approach that utilizes the topological information of the glyphs in ancient scripts in order to provide data about undeciphered inscriptions.

For the sake of clarity, first we define some terms. *Grapheme* is the smallest semantically distinguishing element in a script [1]. Graphemes, in other word *characters*, may be in form of letters, ligatures, numerical digits, or punctuation marks. The grapheme is taken as an object with different features, including its shape variations called glyphs, sound values, the age it was used in, its geographical distribution area and the script it belongs to. The *glyph* refers to a unique shape that can be described by topological information. A significant part of the surviving historical relics have inscriptions. The inscriptions are composed of symbols. *Symbols* are the minimum individual units of the inscriptions from a visual perspective. Consequently, a symbol is the materialization of a particular glyph of a grapheme, and the grapheme is the abstraction of a symbol. It is noteworthy that [2] and [3] use the term *graph* in the same meaning as the previously described term *symbol.*

The scripts are constantly changing after being established [4] in the form of a changing set of graphemes, including the shape transformation of glyphs. One cause of this alteration is the evolution of writing technology, which impacts the glyph shapes. A special case is when the script used in making the inscription is known, but the inscription is composed of realizations of glyph variants which are significantly altered from the typical ones. Consequently, describing the glyph variations could help in deciphering inscriptions. Therefore, we model the topological transformations between the glyph variants of graphemes. Pardede et al. [5] created a multilayer grapheme model to distinguish between the visual identity of a grapheme within an alphabet and the everyday topological variations of the glyph of that grapheme.

Revesz [6] compared some Mediterranean and Eurasian scripts using a translation of the scripts to a DNA encoding, and a hypothetical evolutionary tree reconstruction algorithm was applied. Note that Revesz generated similarity groups of glyphs with partly unknown or different sound values. In such a way, genetically independent, accidentally similar glyphs could be taken as relatives. Then Revesz created the sequence of the similarity groups of glyphs based on the alphabetic list of the graphemes of the appropriate scripts. However, the evolution of the graphemes surely happened individually, and not according to their place in an alphabetic list.

In our research, we intended to filter out the *homoplasies,* which mean the appearance of similar glyphs without genetic relations. Nevertheless, in order to gain reliable results, we restricted the investigations to the phenetics, which does not need evolutionary consideration.

This paper is organized as follows: Section II presents the exploration of similarities among graphemes of various ancient scripts and introduces the typical topological transformations on the glyphs. Section III presents the phenetic analysis of scripts based on the explored similarity groups of glyphs to produce a concise representation of the phenetic relations of the ancient scripts. Section IV draws conclusions.

II. Topological Analysis of Glyphs

In script evolution, glyphs can usually be deduced from earlier glyphs—unprecedented glyphs are extraordinary. It is noteworthy that the typical precondition of a genealogical tie between two glyphs is that their sound values must be identical or phonetically related.

Based on the comparisons of the appropriate glyphs in various scripts, we identified glyph-forming transformations that were applied consciously or otherwise during the evolution of the analyzed scripts, e.g. line insertion, removal, extension, shortening, shifting, mirroring, rotating, ligature formation, duplication, line merger, ornamenting, straight to curve, and curve to straight [7].

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

Each glyph can be topologically characterized by geometrical parameters, e.g. number of vertices of degree one/two/three/four or higher, number of loops, number of horizontal/vertical/approximately horizontal/approximately vertical edges, and number of separate pieces of the glyph. Using these topological parameters, the shape of a glyph can be described by the N_g-element vector g, where and N_g is the number of topological properties ($N_g = 13$ in the present model), and each $g(k)$ element ($k = 1, \dots, N_g$) is a topological property of the glyph g.

Then the relation between the glyphs g_1 and g_2 is stated in (1), where m_i denotes the topological transformations ($i = 1,2, \dots, N_t$), N_t is the number of types of m_i topological transformations ($N_t = 12$ in the present model), and M denotes the number of subsequent topological transformations in transforming g_1 to g_2. Typically, different sequences of topological transformations can lead to the same glyph; therefore, the minimum M is chosen according to the principle of parsimony (Occam's Razor).

$$g_2 = \min_M \left[\left(\prod_{\xi=1}^{M} m_i \right) g_1 \right], 0 \le i \le N_t \qquad (1)$$

The correspondence between glyphs in a similarity set could be genetic relation or only homoplasy. The more complex a glyph topologically is, the more reliably do its similarity set indicates descent relations. In order to measure the topological complexity, the Glyph Complexity Parameter (GCP) is introduced, which is calculated for a glyph as in (2).

$$GCP = (\operatorname{sgn} N)\left[\left(\sum_{i=1}^{N} e_i^{(N)} \right) - 1 \right] + \sum_{k=1}^{R} e_k^{(R)} + [\operatorname{sgn}(S-1)]S^2 + (\operatorname{sgn} L)\left[\left(\sum_{j=1}^{L} e_j^{(L)} \right) + 1 \right] + A, \qquad (2)$$

where N is the number of vertices; $e_i^{(N)}$ is the number of edges connected to vertex i; R is the number of reflections or translations of edges; $e_k^{(R)}$ is the number of edges of the k'th reflection or translation of edges; S is the number of separated partitions of the glyph topology; L is the number of the non-overlapping loops in the glyph; $e_j^{(L)}$ is the number of edges between the vertices of the j'th loop; and A is the number of arches not belonging to any loop. If the edges are curved, the concentricity is considered as parallelism. Equation (2) has been tested on several glyphs.

In order to develop a method to systematically explore the descent relationships among the graphemes, the sets of possible cognate glyphs have to be selected from the glyphs of the investigated scripts. We supposed that for the cognate glyphs the sequence of topological transformation as in (1) can be found, and is only composed of a small number of transformations. Therefore, selecting the sets of possible cognate glyphs should be based on the minimization of the sequence of topological transformation between them.

It is noteworthy that the glyphs of the possible cognate graphemes could vary greatly; therefore, further elimination is necessary for identifying those similarity groups of glyphs as subsets of the possible cognate sets of glyphs, which could be handled as the most probable direct cognates. In the second step, our method minimizes the difference of the topological properties of glyphs, which belong to one of the possible cognate sets of glyphs. If the difference is under an arbitrary threshold, then the investigated glyphs belong to a common similarity group. If the difference exceeds the threshold, two or more glyph similarity groups are selected from the examined possible cognate set of glyphs. For calculating the difference of the topological properties of glyphs, (3) was applied.

$$d(g_1, g_2) = \sqrt{\frac{1}{N_g} \sum_{k=1}^{N_g} w_k [g_2(k) - g_1(k)]^2}, \qquad (3)$$

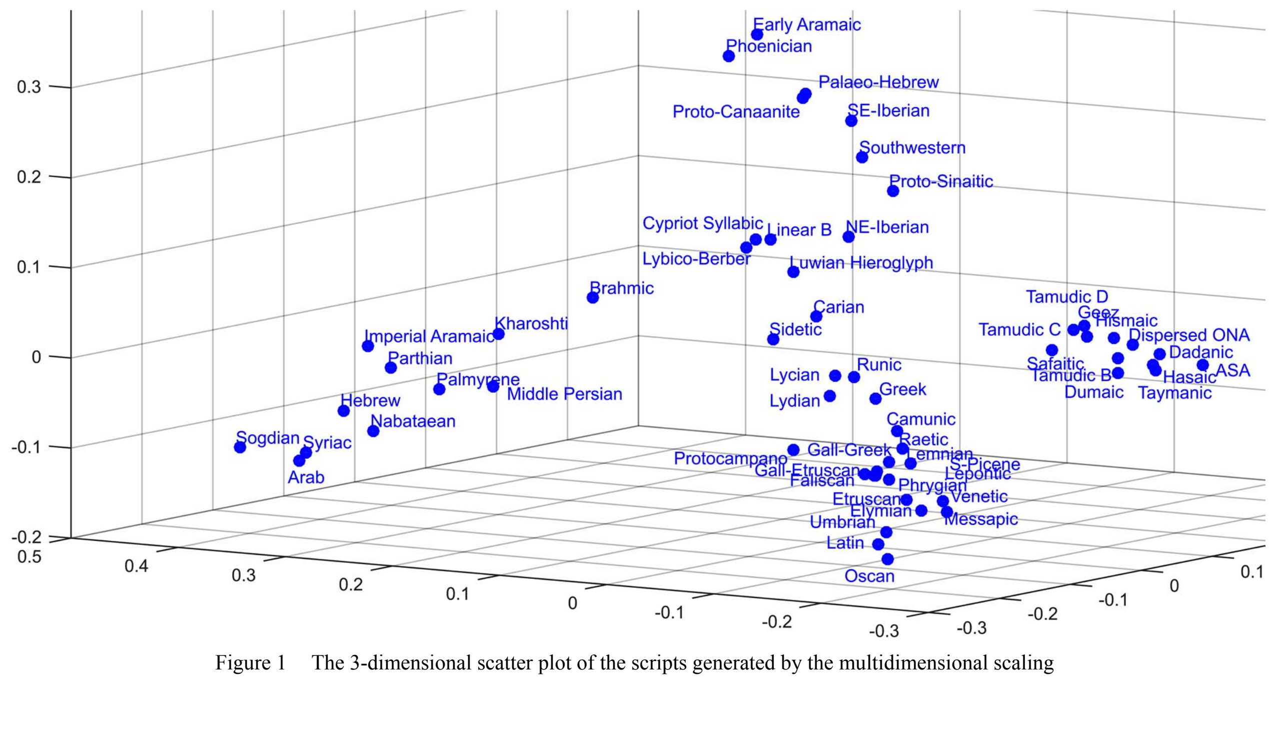

Figure 1 The 3-dimensional scatter plot of the scripts generated by the multidimensional scaling

where w_k $(k = 1, \ldots, N_g)$ is the weight representing the significance of each topological property. The value of w_k was determined experimentally based on the descent relations of certain well-known scripts from international paleography; moreover, $\sum_{k=1}^{N_g} w_k = N_g$. Equation (3) takes into account that in case of more complex shapes of glyphs the smaller differences are not yet significant, and contrarily, even the smallest differences can be important for less complex shapes.

III. Phenetic Analysis of Scripts

Based on the similarity groups of glyphs selected based on (3), the phenetic analysis of the investigated scripts were carried out as the third step of our method. We analyzed 58 different historical Mediterranean and Asian scripts as data points. These data points were described with the use of several features (variables), including 342 different similarity groups of the glyphs and 12 orthographical properties of the scripts as variables. The examined glyphs were obtained from a comprehensive set of paleographical publications [8]–[23]. There are two abbreviations used: ASA for Ancient South Arabian script, and ONA means Oasis North Arabian, a group of scripts [16].

In the phenetic analysis the Jaccard distance was used for calculating the 58×58 distance matrix of the scripts. This matrix was converted by an excessive multidimensional scaling to 2- and 3-dimensional data structures without significant data loss. The resulting 3-dimensional scatter plot is presented in Fig. 1.

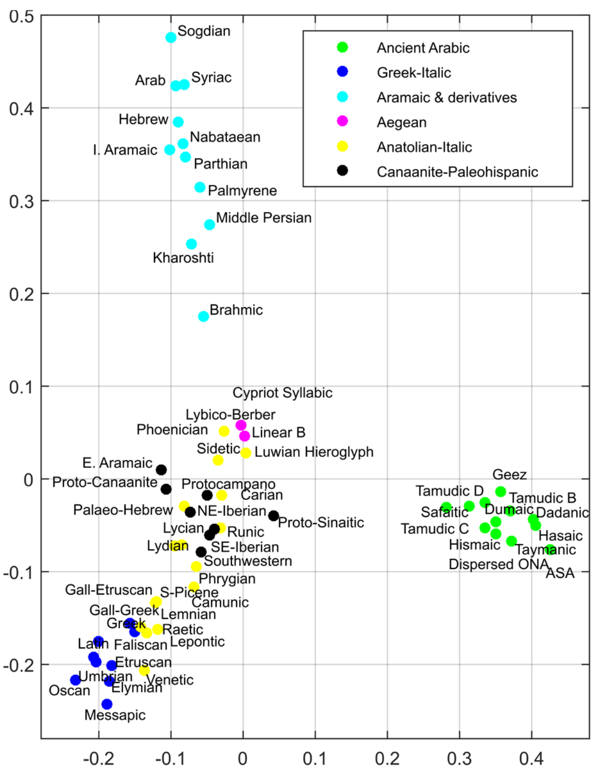

Figure 2 Scatter plot of early scripts based on MDS to 2-dimension and then k-means clustering

To determine phenetic relationships of the scripts, we applied the k-means clustering algorithm [24] with 6 clusters on the result of the 2- and 3-dimensional multidimensional scaling. In the k-means clustering algorithm the squared Euclidean distance was used; therefore, each centroid is the mean of the data points in that cluster. The resulting scatter plot in the case of 2-dimensional scaling is presented in Fig. 2; the computation was carried out with the use of MATLAB [25]. The cluster structure was validated by the Dunn index [26], which was 0.76 in the presented case. Other phenetic analyses were also carried out, using various hierarchical clustering algorithms [27], including single linkage, complete linkage, and two average linkages: UPGMA [28] and WPGMA [29]. The appropriateness of these algorithms strongly depends on the data structure to be clustered. The investigated scripts were developed based on an evolution. Therefore, some branches of the scripts remained close to each other during the history. That is the reason why the single linkage clustering is not efficient, since it cannot distinguish the clusters with elements close to each other. Moreover, there are outlier members of the script branches; therefore, the complete linkage clustering is also not optimal. Particularly, certain scripts had several descendants (e.g. Aramaic script), while others remained singular (e.g. Lybico-Berber script). Consequently, the number of elements of the clusters is inhomogeneous. The UPGMA gives weights to each cluster according to the number of elements of the

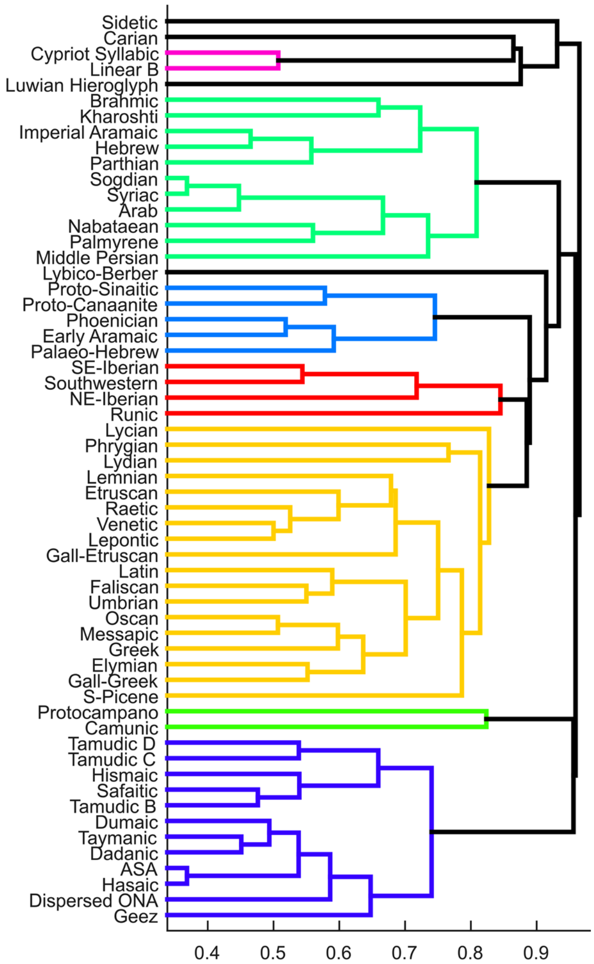

Figure 3 Dendrogram of early scripts based on phenetic analysis with WPGMA clustering

cluster in each step. Sneath and Sokal [30] demonstrated that the UPGMA would favor clusters more similar in size. On the contrary, the WPGMA is appropriate when there is a reason a priori to eliminate size differences between the resulting clusters. Fig. 3 presents the phenogram of the scripts calculated by using WPGMA. In the hierarchical clustering investigations the Jaccard distance was used and an optimal leaf ordering for hierarchical binary cluster tree [31] was applied.

Besides the crisp clustering algorithms, the fuzzy c-means clustering [32], [33] was also applied to explore the behavior of the scripts. The fuzzy clustering was carried out based on iterative minimization of (4).

$$J(\mu, a) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{m} \left\| x_j - a_j \right\|^2,$$
$$\mu_{ij} = \frac{\left\| x_i - a_j \right\|^{-2/(m-1)}}{\sum_{k=1}^{C} \left\| x_i - a_k \right\|^{-2/(m-1)}}, a_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} x_j}{\sum_{j=1}^{n} \mu_{ij}^{m}}, \tag{4}$$

where n is the number of data points (scripts), c is the number of clusters, x_i is the i'th script as data point, u_{ij} is the degree of membership of x_i in the j'th cluster, a_j is the j'th cluster center, $\| \quad \|$ is any norm that expresses the similarity between x_i and a_j, $i = 1, \ldots, c$, and $j = 1, \ldots, n$. The m weighting exponent is a fuzziness index (fuzzifier) that influences the performance of the c-means, $m \in \mathcal{R}$, $1 \leq m < \infty$. Fig. 4 presents the result of the fuzzy clustering, where $n = 58$ scripts were grouped into $c = 6$ clusters, and the fuzzifier $m = 1.3$.

The cluster centers are denoted with colored × marks in Fig. 4. This result is consistent with Fig. 2, since the same groups of scripts are represented by the two most separated clusters (Aramaic and Ancient Arabic), while centers of other clusters became close to each other. Consequently, the separation of these clusters became very weak. The clusters of the Aegean and the Canaanite-Paleohispanic scripts merged into one cluster, which could indicate the weak separation of these two groups of scripts. Since the number of clusters was predefined (6), the Runic script created an individual cluster. In case of the k-means clustering, the Runic belonged to the Anatolian-Ancient Italic cluster (Fig. 2), while in case of the hierarchical WPGMA clustering; the Runic was in the same cluster with the Paleohispanic scripts (Fig. 3). The difference between the two different clustering algorithms is not special. Considering the leaf ordered dendrogram in Fig. 3, the Runic script is located between the Paleohispanic and the Anatolian (Lycian, Phrygian, Lydian) scripts. This position expresses the similarity of Runic with both groups of scripts. In case of the fuzzy clustering, when the Paleohispanic scripts are clustered together with other groups (Canaanite and Aegean), and the Anatolian scripts are clustered with Ancient Italic scripts, Runic remained alone in a separate cluster. This example demonstrated that the results of the various clustering algorithms are not contradictive despite their differences.

IV. Application of the Results

In case of homoplasy, glyphs of unrelated graphemes of different scripts could be identical. In our research we were able to filter out the homoplasies, since the sound values of the graphemes of the same glyphs belonging to the same cluster are usually identical or at least similar to each other, as Table I presents some examples for demonstrating the diversity of graphemes with identical glyphs. The glyphs are based on the paleographical literature [8]–[23]. On Table I, the transliteration values

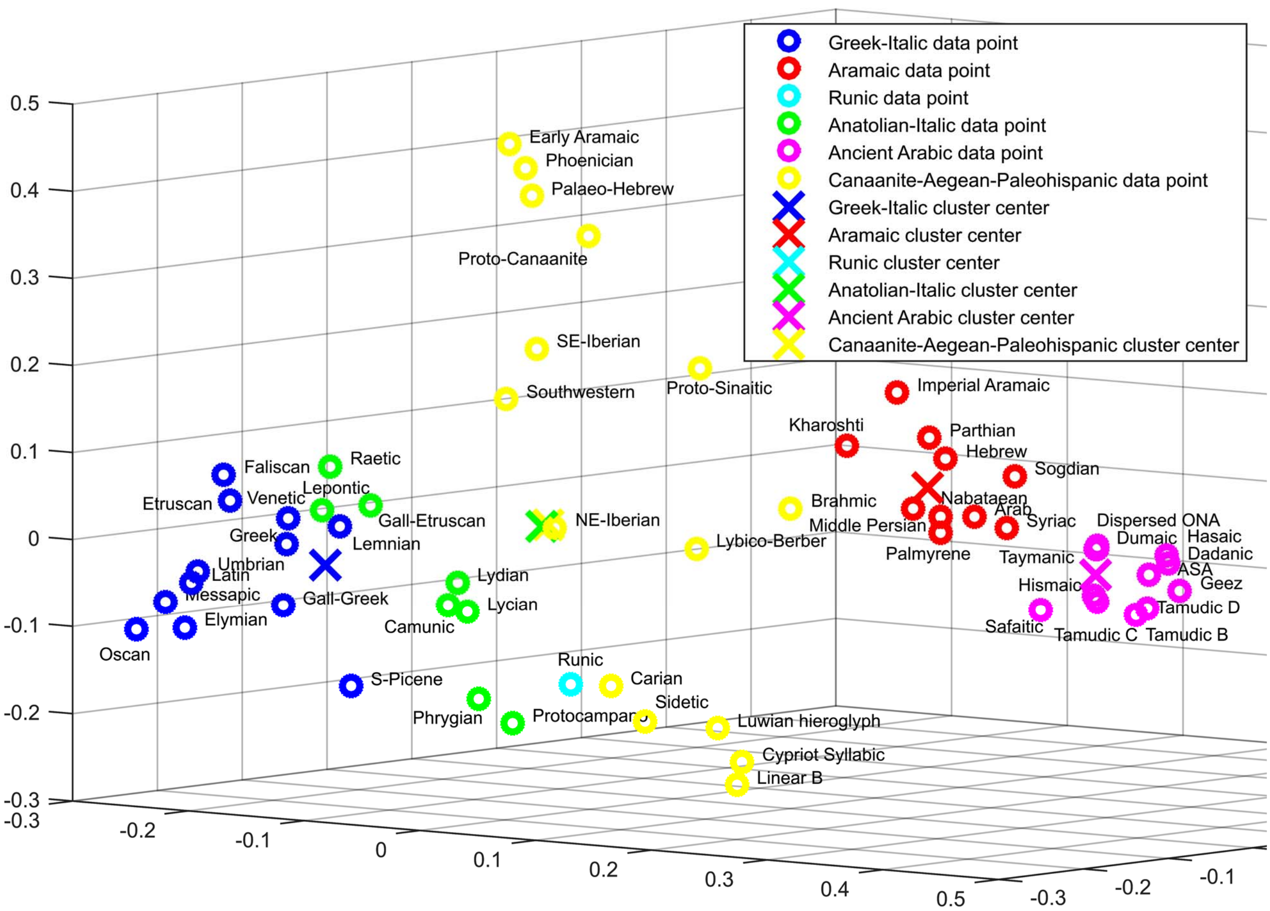

Figure 4 Scatter plot of early scripts based on MDS to 3 dimension and then c-means clustering

(usually corresponding to the sound values) are placed in angle brackets, and the names of the clusters are in italics. Table I shows that the graphemes belonging to the same cluster of scripts have typically cognate sound values. However, there are exceptions, e.g. in the last row, where two Paleohispanic scripts, the SE-Iberian and the NE-Iberian have different sound values for the same glyph: <e> and <ř>, respectively. In case of an undeciphered inscription the determination of the script used for producing the inscription is difficult due to the differences of the signs of the inscription and the glyphs of the appropriate graphemes. However, determining which cluster of scripts is the most appropriate for the undeciphered inscription is much easier than choosing only one script. Then the deciphering process can be carried out based on the sound values of all graphemes of each script in the same cluster. As it was demonstrated above, the cluster structure depends on the applied machine learning algorithm; therefore, additional investigation could be necessary to decide which clustering algorithm is the most appropriate.

TABLE I.
SOME EXAMPLES FOR THE DIVERSITY OF GRAPHEMES

Glyph	Scripts with transliteration values
Ŧ	*Aegean:* Linear B <to> *Canaanite:* Proto-Canaanite <s> *Paleohispanic:* Southwestern, SE-Iberian <o>
ǂ	*Aegean:* Linear B, Cypro-Syllabic <pa> *Anatolian:* Lydian <s> *Ancient Italic:* Lepontic <z> *Paleohispanic:* Southwestern <o>, SE-Iberian <o>
⧖	*Aegean:* Cypro-Syllabic <le> *Anatolian:* Carian <γ> *Ancient Arabic:* ASA <z> *Canaanite:* Proto-Sinaitic <z> *Paleohispanic:* Southwestern <k^o>; NE-Iberian <go/ko> *Lybico-Berber:* Lybico-Berber <S>
8	*Anatolian:* Phrygian <b/p^h> *Ancient Italic:* Etruscan <fh>, Oscan, Umbrian <f>
ȣ	*Anatolian:* Lydian <f> *Ancient Arabic:* ASA, Tamudic B, Safaitic, Hasaitic <ṯ>
ϙ	*Canaanite:* Proto-Sinaitic <w, q>; Proto-Canaanite <q> *Anatolian:* Carian <t> *Ancient Arabic:* ASA, Dispersed ONA, Dumaitic, Taymanitic, Dadanitic, Hismaitic, Thamudic B, Safaitic, Hasaitic, Geʿez <y> *Paleohispanic:* SE-Iberian <e>, NE-Iberian <ř> *Ancient Italic:* Etruscan, Faliscan, Messapic <q> *Luwian:* Luwian Hieroglyph <wa/i_9> *Greek:* Greek <q>

V. CONCLUSIONS

The paper presented a machine learning approach, including a composite method. Its first step is searching for sets of the possible cognate glyphs. It utilizes the determined typical topological transformations of the glyphs, which were used in the evolution of the scripts. This method selects the possible cognate glyphs from the phonetically similar graphemes in such a way that it minimizes the necessary topological transformations between glyphs.

The second step of our method generates similarity group of each set of the possible cognate glyphs by minimizing the differences between the topological properties of the glyphs belonging to the same group.

The third step is a set of various machine learning methods in order to obtain a phenetic model for the investigated scripts based on the explored similarity relations of their graphemes (based on the topological similarity of their glyphs). In this step, the multidimensional scaling and various crisp and fuzzy clustering algorithms were applied.

The presented results give an overall picture about the phenetic relations of the examined scripts. Our research does not include phylogenetic or cladistic investigations; notwithstanding, the resulting phenograms give a rough estimation of their genealogical ties.

The present study concentrated on the phenetic analysis of the Mediterranean and Aramaic-based Asian scripts; but the presented method could be extended to other writing systems. The introduced approach may give a support for the paleographers in exploring the relations among scripts and deciphering ancient inscriptions.

REFERENCES

[1] Sukkarieh, Jana Z., Matthias von Davier, Kentaro Yamamoto, "*From Biology to Education: Scoring and Clustering Multilingual Text Sequences and Other Sequential Tasks,*" Educational Testing Service, Princeton, NJ, ETS Research Report No. RR–12-25, Dec. 2012.

[2] G. August, "Descriptively and explanatory adequate models of orthography," in *New Trends in Graphemics and Orthography*, G. August, Ed. Berlin, New York: de Gruyter, 1986, pp. 25–42.

[3] M. Kohrt, "The term 'grapheme' in the history and theory of linguistics," in *New Trends in Graphemics and Orthography*, G. August, Ed. Berlin, New York: de Gruyter, 1986, pp. 80–96.

[4] H. Rogers, "Sociolinguistic factors in borrowed writing systems," *Toronto Working Paper in Linguistics*, vol. 17, pp. 247–262, 1999.

[5] R. E. I. Pardede, L. L. Tóth, A. G. Jeney, F. Kovács, and G. Hosszú, "Four-layer grapheme model for computational paleography," *J. Information Technology Research.*, in press.

[6] P. Z. Revesz, "A computational study of the evolution of Cretan and related scripts," in *Mathematical Models and Computational Methods, Proc. Int. Conf. Applied Mathematics, Computational Science and Engineering*, 2nd ed., I. J. Rudas, ed., Montclair, NJ: Institute for Natural Sciences and Engineering, 2015, pp. 101–105.

[7] G. Hosszú, "A novel computerized paleographical method for determining the evolution of graphemes," in *Encyclopedia of Information Science and Technology*, M. Khosrow-Pour, Ed., 3rd ed. Hershey, PA: IGI Global, 2015, ch. 194, pp. 2017–2031.

[8] I.-J. Adiego Lajara, *The Carian Language.* Leiden, The Netherlands: Brill, 2007.

[9] P. T. Daniels and W. Bright, Eds., *The World's Writing Systems.* New York, Oxford: Oxford University Press, 1996.

[10] B. Davis, "Introduction to the Aegean pre-alphabetic scripts," *Kubaba* vol. 1, pp. 38–61, 2010.

[11] A. Glass, "A preliminary study of Kharoṣṭhī manuscript paleography," MA Thesis, Dept. Asian Languages and Literature, Univ. of Washington, Seattle, WA, 2000.

[12] E. Laroche, *Les hiéroglyphes hittites,* Première partie, Paris, France: L'écriture, 1960.

[13] LBI: Libyco-Berber Inscriptions Online Database. © LBI-Projekt, Wien: Institutum Canarium [Online]. Available: http://www.institutum-canarium.org

[14] M. Lejeune, "Les inscriptions de Gordion et l'alphabet phrygien," *Kadmos* vol. 9 no. 1, pp. 51–74, 1970.

[15] M. C. A. Macdonald, "Ancient North Arabian," in *The Cambridge Encyclopedia of the World's Ancient Languages*, R. D. Woodard, Ed., Cambridge, UK: Cambridge University Press, 2004, pp. 488–533.

[16] M. C. A. Macdonald, (2015): "On the uses of writing in ancient Arabia and the role of palaeography in studying them," *Arabian*

Epigraphic Notes [Online] vol. 1, (2015): 1–50. Available: http://hdl.handle.net/1887/32745

[17] S. Marchesini, "The Elymian language," in *Language and Linguistic Contact in Ancient Sicily,* O. Tribulato, Ed., Cambridge: Cambridge University Press, 2012, pp. 95–114.

[18] S. Marchesini, "Über die Rätische Inschrift aus Pfatten/Vadena im Tiroler Landesmuseum Ferdinandeum, Innsbruck," in *Wissenschaftliches Jahrbuch der Tiroler Landesmuseen*, W. Meighörner, Ed., Innsbruck, Wien, Bozen: StudienVerlag, 2014, pp. 202–217.

[19] MNAMON: Antiche Scritture del Mediterraneo. Guida critica alle risorse elettroniche. Pisa: Scuola Normale Superiore, section Laboratorio Informatico per le Lingue Antiche (LILA) [Online]. Available: http://lila.sns.it/mnamon

[20] T. G. Palaima, "Scribes, scribal hands and palaeography," in *A Companion to Linear B. Mycenaean Greek Texts and their World,* vol. 2, Y. Duhoux and A. M. Davies, Eds., Louvain-La-Neuve, Walpole, MA: Peeters, 2011, pp. 33–136.

[21] N. Sims-Williams and F. Grenet, "The Sogdian inscriptions of Kultobe," *Shygys* vol. 1, pp. 95–111 and color pages without pagination, 2006.

[22] P. O. Skjærvø, "The joy of the cup: A pre-Sasanian Middle Persian inscription on a silver bowl," *Bulletin of the Asia Institute.* New Series, vol. 11, pp. 93–104, 1997.

[23] M. Valério, "Origin and development of the Paleohispanic scripts: the orthography and phonology of the Southwestern alphabet," *Revista Portuguesa de Arqueologia*. vol. 11, no. 2, pp. 107–138, 2008.

[24] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proc. 5th Berkeley Symp. Math. Statistics and Probability*, vol. 1, Berkeley: University of California Press, 1967, pp. 281–297.

[25] MATLAB Release R2015a, The MathWorks, Inc., Natick, MA, United States.

[26] J. C. Dunn, "Well separated clusters and optimal fuzzy partitions," *J. Cybernetics* vol. 4, no. 1, pp. 95–104, 1974.

[27] S. C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika,* vol. 32, no. 3, pp. 241–254, Sep. 1967.

[28] R. R. Sokal and C. D. Michener, A statistical method for evaluating systematic relationships. *The University of Kansas Science Bulletin* vol. XXXVIII, no. 22, pp. 1409–1438, 1958.

[29] L. L.,McQuitty, "Expansion of similarity analysis by reciprocal pairs for discrete and continuous data," *Educ. Psychol. Measurement*, vol. 27, no. 2, pp. 253–255, 1967.

[30] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy.* San Francisco, CA: Freeman, 1973.

[31] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics*, vol. 17, no. Suppl 1, pp. S22–S29, Jun. 2001.

[32] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybernetics* vol. 3, no. 3, pp. 32-57, 1973.

[33] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms.* New York, NY: Plenum Press, 1981.