Analysis of Glyph and Writing System Similarities using Siamese Neural Networks

Claire Roman[1], **Philippe Meyer**[2]
[1]Independent Researcher
[2]Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350, Jouy-en-Josas, France
[1]claire.roman.91@gmail.com, [2]philippe.meyer@inrae.fr

Abstract

In this paper we use siamese neural networks to compare glyphs and writing systems. These deep learning models define distance-like functions and are used to explore and visualize the space of scripts by performing multidimensional scaling and clustering analyses. From 51 historical European, Mediterranean and Middle Eastern alphabets, we use a Ward-linkage hierarchical clustering and obtain 10 clusters of scripts including three isolated writing systems. To collect the glyph database we use the Noto family fonts that encode in a standard form the Unicode character repertoire. This approach has the potential to reveal connections among scripts and civilizations and to help the deciphering of ancient scripts.

Keywords: siamese neural network, writing system, clustering

1. Introduction

The study of the comparison of scripts is interesting as it unveils links between alphabets and between glyphs, shedding light on the evolution of languages. This helps in comprehending the evolution and historical narratives of civilizations, including their migrations and interactions (Hooker, 1990; Salomon, 1998). Furthermore, it plays a pivotal role in deciphering ancient scripts and inscriptions, for example by identifying the writing systems most closely related to an undeciphered alphabet. Employing this methodological approach, Ventris and Chadwick (1953) successfully deciphered the Linear B script through a meticulous comparison with the Greek alphabet.

To apply computational linguistics and artificial intelligence to glyph comparison several issues have to be considered when choosing an appropriate model. On the one hand related graphemes could vary considerably and so a similarity function more robust to variations than usual image quality metrics such as the mean-squared error or the structural similarity (Wang et al., 2004) is needed. On the other hand artificial neural networks are widely known for their resilience to fluctuations for classification tasks (LeCun et al., 2015) but require a lot of labelled data per class which poses a challenge when comparing glyphs since thousands of classes have to be considered.

Siamese neural networks are a particular class of deep learning models that focus on discerning similarities between entries instead of classifying them into distinct categories. This makes them effective when labeled data is scarce and therefore efficient for one-shot learning (Bromley et al., 1993). They find recent applications in various fields such as intrusion detection systems (Bedi et al., 2020) or blood cell classification (Tummala and Suresh, 2023).

In this paper, we use the siamese neural networks developed by Koch et al. (2015) which have been trained and tested on the Omniglot dataset (Lake et al., 2015) in order to compare similarities between graphemes and study the space of writing systems. For that purpose we use 51 historical European, Mediterranean and Middle Eastern writing systems that we have collected from the Noto font families that encode the Unicode characters. Then we visualize the space of glyphs by multidimensional scaling analyses and we perform a Ward-linkage hierarchical clustering to define 10 families of writing systems. The dataset and codes are released at `https://github.com/PhilippeMeyer68/glyph-SNN`.

2. Related work

Various computational studies of the script evolution and comparison with the tools of mathematics, computer science and artificial intelligence have been performed. For example, families of writing systems have been obtained using clustering algorithms by minimizing the necessary topological transformations between glyphs (Hosszú and Kovács, 2016) and by using convolutional neural networks (Daggumati and Revesz, 2023). Clustering algorithms have also been used to study subgroups of a given writing system such as in Corazza et al. (2022) where unsupervised deep learning is used to classify the Cypro-Minoan writing system in one single group or in Bogacz et al. (2018) where 3D scanning and object identification are applied to visualize links between Maya glyphs.

LT4HALA 2024@LREC-COLING-2024, pages 98–104
25 May, 2024.

On the other hand deep learning models have also shown their efficiencies for glyph recognition and translation (Barucci et al., 2021, 2022; Moustafa et al., 2022; Guidi et al., 2023; Hamplová et al., 2024). In particular, Liu et al. (2022) extended the work of Koch et al. (2015) by using siamese neural networks for ancient character recognition. For an overview of published research using machine learning for ancient languages one can see the survey of Sommerschield et al. (2023).

Other approaches to decipher old scripts such as algorithms based only on non-parallel data in known languages (Luo et al., 2019) or natural phonological geometry, word segmentation and cognate alignment (Luo et al., 2021) have been conducted.

3. Method

3.1. Distances between glyphs and scripts via siamese neural networks

The model developed by Koch et al. (2015) consists of two identical convolutional neural networks that share the same set of parameters and weights. Each subnetwork takes a 105x105 pixels image as input and processes it independently through convolutional layers to generate a feature vector. These feature vectors are then compared to measure the similarity between the two input images. The network is trained using pairs of images, where one is compared to another, belonging to the same class or not, that is to say considered as positive or negative sample. A regularized cross-entropy objective loss function is employed during training to encourage the network to minimize the distance between feature vectors for images of the same class and maximize it for images of different classes. This way, the siamese network learns to extract meaningful and discriminative features that facilitate accurate similarity measurements, enabling effective one-shot learning.

To train the siamese neural network, the authors of Koch et al. (2015) use the Omniglot dataset (Lake et al., 2015) composed of 1,623 characters handwritten by 20 different individuals and from 50 alphabets, both real and invented writing systems such as the Aurebesh and Tengwar. In this work we use the same siamese neural network model, except that we train it only on the 15 invented languages of Omniglot to avoid introducing bias by comparing glyphs that would have already been used during the training phase. We still select the same random number of input pairs, that is to say 150,000 pairs of glyphs augmented with 8 distortion copies, which give 1,350,000 effective pairs.

For two glyphs g_1 and g_2 we denote by $\mathrm{SNN}(g_1, g_2)$ the similarity predicted by this siamese neural network and by d_g the dissimilarity measure, or distance-like function, defined by

$$d_g(g_1, g_2) := 1 - \mathrm{SNN}(g_1, g_2). \tag{1}$$

Let s_1 and s_2 be two scripts. We define the similarity of s_1 to s_2 by

$$\tilde{d}_s(s_1, s_2) := \frac{1}{n} \sum_{g_1 \in s_1} \min_{g_2 \in s_2} \Big(d_g(g_1, g_2) \Big), \tag{2}$$

where n is the number of glyphs of s_1. We symmetrize it to obtain the distance-like function d_s between s_1 and s_2 defined by

$$d_s(s_1, s_2) := \frac{1}{2} \Big(\tilde{d}_s(s_1, s_2) + \tilde{d}_s(s_2, s_1) \Big). \tag{3}$$

In this definition a glyph of s_1 can be associated with several glyphs of s_2. We believe that imposing a 1-1 mapping in the definition of d_s, such as for the bottleneck distance between persistence diagrams (Cohen-Steiner et al., 2005), is less appropriate since several glyphs can be historically related to a single glyph. For example it is known that the letters U, Y, V and W of the Latin alphabet have as ancestor the upsilon greek character Υ (Daniels and Bright, 1996).

3.2. Font-driven database

We have selected 51 historical European, Mediterranean and Middle Eastern writing systems and obtained the database of corresponding characters from their Unicode identifiers and the Noto Sans Regular family fonts.

The Unicode repertoire is an inventory of characters maintained by the Unicode Consortium and encompassing text from every writing system worldwide, facilitating global communication and interoperability across different devices and platforms.

The Noto font collection is designed and engineered for typographically correct and aesthetically pleasing global communication in more than 1,000 languages and over 160 scripts. It supports more than 77,000 characters and includes nearly all non-CJK characters included in the actual Unicode Standard version. Each supported script has at least one font in a basic style called Noto Sans Regular. This allows characters to have a standardized form, of the same size and quality.

By this process we have a database of 1,649 standardized centered glyphs as 105x105 pixels image from 51 alphabetic and syllabic writing systems. These chosen scripts are listed in Appendix A.

4. Results

4.1. Space of glyphs and scripts

In this section, we use the dissimilarity measures d_g and d_s to compare and visualize glyphs and scripts from our database. We have found that the two scripts which are the closest are the Old Sogdian and the Pahlavi Psalter with a distance of 0.05 while the most distant pair is the Coptic and the Old Persian with a distance of 0.88. Looking at how distant a script is to other writing systems by summing its distance to all other scripts we see that the Old Persian is actually the most isolated alphabet, see Table 1.

Script	Distance to other scripts
Old Persian	33.37
Glagolitic	27.69
Meroitic Hieroglyphs	26.07
Ogham	22.04
Tifinagh	21.48

Table 1: The 5 most isolated scripts with respect to the siamese-based distance.

In order to visualize graphemes and alphabets and the distances separating them we use multidimensional scaling (MDS) analysis. This is a technique in dimension reduction that aims to represent complex, high-dimensional data in a lower-dimensional space while preserving the pairwise distances between data points as accurately as possible (Kruskal, 1964).

In this way, we can represent the glyphs of one or several scripts. In Figure 1 is given the 2-dimensional scaling analysis of the Latin and Old Italic scripts, which have a distance d_s equal to 0.26. Several glyphs of these alphabets are similar and close, illustrating the real connections between these scripts, the Old Italic used in the Italian Peninsula from the 8th to the 1st century BC being known as an ancestor of the Latin, see Bonfante (1996). In Figure 2 we represent a 2-dimensional scaling analysis of the Coptic and Old Persian scripts which is the most distant pair of alphabets of the database and we notice that the alphabets essentially form two distinct clusters.

4.2. Comparison and clustering of writing systems

In this section we perform a Ward-linkage hierarchical clustering (Ward Jr., 1963) on the 51 writing systems with respect to the siamese-based distance function d_s. This agglomerative clustering algorithm analyzes the variance of the clusters and is known to be less sensitive to noise and outliers than the other hierarchical clustering algorithms.

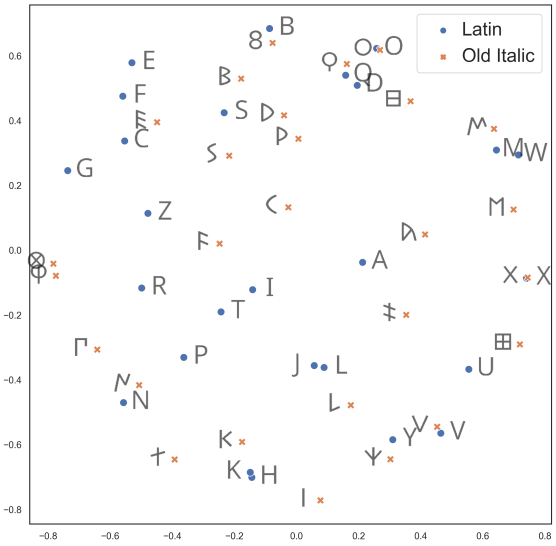

Figure 1: Multidimensional scaling in dimension 2 of the Latin and Old Italic glyphs which are close scripts with respect to the siamese-based distance.

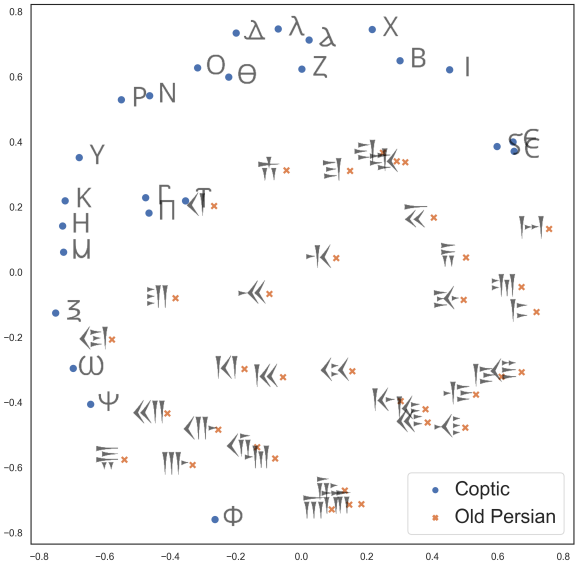

Figure 2: Multidimensional scaling in dimension 2 of the Coptic and Old Persian glyphs which are distant scripts with respect to the siamese-based distance.

The associated dendrogram of the clustering is given in Figure 3.

The Elbow method clearly indicates to truncate the dendrogram at 10 clusters. For this truncation the clustering quality Dunn index (Dunn, 1973) is 0.81. Information about size, medoid, diameter and mean distance of all pairs of each cluster is given in Table 2.

As noticed in Section 4.1, the Old Persian cuneiform, the old slavic Glagolitic and the Meroitic Hieroglyphs are isolated scripts and define their

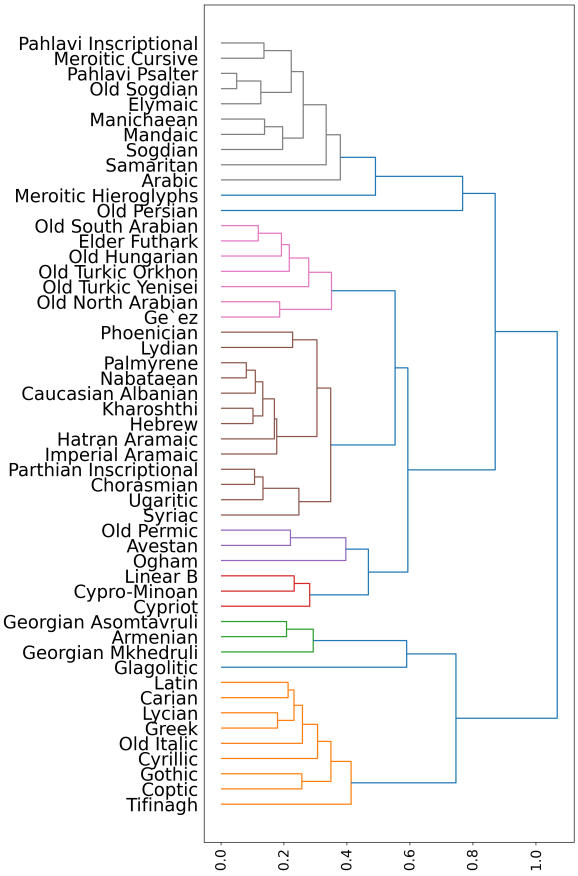

Figure 3: Dendrogram associated to a Ward-linkage hierarchical clustering of the scripts with the siamese-based distance.

Cluster	Size	Medoid	Diam.	Mean d.
C1	9	Greek	0.41	0.28
C2	3	Georgian Asomtavruli	0.28	0.25
C3	1	Glagolitic	0	0
C4	3	Cypro-Minoan	0.29	0.26
C5	3	Avestan	0.43	0.31
C6	13	Nabataean	0.40	0.19
C7	7	Old South Arabian	0.34	0.23
C8	10	Pahlavi Psalter	0.43	0.21
C9	1	Meroitic Hieroglyphs	0	0
C10	1	Old Persian	0	0

Table 2: Size, medoid, diameter and mean distance of all pairs of each cluster.

own families in the clustering. There is a cluster composed of the 3 Cypriots writing systems and another one composed of the Armenian and Georgian scripts. The three rather distant Old Permic, Avestan and Ogham scripts are grouped together. Several Middle Eastern writing systems such as the Pahlavi, the Arabic and the Sogdian form another cluster. The Greek alphabet is the medoid of a cluster composed of 9 scripts, such as the Latin or the Cyrillic and other Greek extensions such as the Carian. Another group of scripts is given of the Old Arabian and Turkic scripts. The last cluster is the biggest one, composed of Aramaic scripts that could be divided into subfamilies.

To represent how distant or close the scripts and the clusters are to each other, we perform a 2-dimensional scaling analysis and the associated visualization is given in Figure 4. We see that the distribution of the scripts respects the clusters defined by the Ward-linkage hierarchical clustering with little overlap between groups.

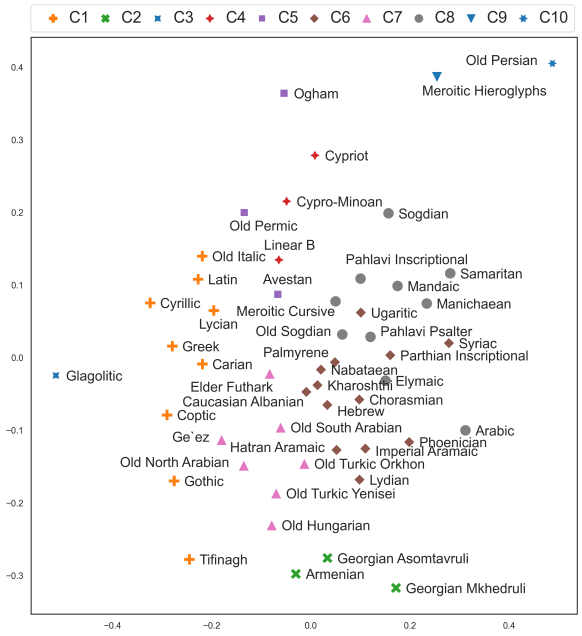

Figure 4: Multidimensional scaling in dimension 2 of all the scripts with respect to the siamese-based distance where the colors represent the 10 clusters of writing systems.

4.3. Comparison of our results with the literature

In Hosszú and Kovács (2016), 58 different historical Mediterranean and Asian scripts are classified by clustering algorithms applied on topological features of glyphs. The main similarities with our work are that there is a Latin-Greek group, a Hebrew-Nabataean group and a Cypriot group with both approaches. However, the Lydian and Phoenician scripts are in different clusters in Hosszú and Kovács (2016) while they are close with a distance of 0.23 by our metric which seems to be in agreement with the work of Woudhuizen (2020). Furthermore, the Carian script is an isolated point in Hosszú and Kovács (2016) while it is classified in the Greek family in our work. The similarities and the possible historical connection between graphemes of the Carian and the Greek scripts have been remarked and extensively studied, see Chapter 4.B *The Greek Alphabetic Era* of Adiego

(2006). Finally the Dunn index of our clustering is 0.81 which is slightly better than the Dunn index of 0.76 of Hosszú and Kovács (2016).

In the work of Daggumati and Revesz (2023), 8 ancient scripts are classified with convolutional neural networks combined with support vector machines and a hierarchical clustering. The main difference is that the Greek and the Phoenician scripts are very close with their metric whereas they are in two different clusters in our work with a siamese-based distance of 0.46. Indeed, it is known that these writing systems are related and that several glyphs of the Greek alphabet are vertical mirror reflections of Phoenician glyphs, see Swiggers (1996). It turns out that this phenomenon of boustrophedon writing is taken into account in the metric of Daggumati and Revesz (2023) but not in ours.

5. Conclusion

In this study, we introduce a two-step process for comparing glyphs and writing systems. Firstly, we present a method for generating a clean alphabet database from the Noto Sans fonts and the Unicode inventory. Then a distance-like function defined by a siamese neural network is given. This allows us to consider space of glyphs and scripts to compare them. Then a Ward-linkage hierarchical clustering of 51 alphabets resulted in the identification of 10 clusters representing related writing systems. These groups very often represent real historical connections, such as the Georgian and Armenian cluster or the Latin cluster composed of the Latin, Carian, Lycian, Greek, Old Italic, Cyrillic, Gothic, Coptic and Tifinagh scripts. This demonstrates the effectiveness of the approach in identifying links between alphabets and motivates future research to its application in deciphering ancient scripts and inscriptions.

We now discuss limitations of this approach. The comparison explained in this article is only based on the graphical aspect of the graphemes and scripts, there is no knowledge about the phonetic facet of the associated languages that intervenes. Furthermore, this work uses Unicodes and fonts and then requires an implementation of the writing systems which is not the case for all of them. For example until now there is no Unicode for the Paleohispanic scripts. Moreover, we mostly have compared segmental scripts. It is not clear if it makes sense to extend this type of comparison to logographic writing systems composed of thousands of signs such as the Chinese characters.

In future work, we would like to include all the scripts encoded in the Unicode repertoire to obtain a larger taxonomy of world's writing systems in order to contribute to the study of historical connections between civilizations. It would be particularly interesting to apply this approach to the decipherment of ancient scripts by comparing them with deciphered writing systems.

6. Bibliographical References

I. Adiego. 2006. *The Carian Language*, volume 86 of *Handbook of Oriental Studies. Section 1 The Near and Middle East*. Brill, Leiden, The Netherlands.

A. Barucci, C. Canfailla, C. Cucci, M. Forasassi, M. Franci, G. Guarducci, T. Guidi, M. Loschiavo, M. Picollo, R. Pini, L. Python, S. Valentini, and F. Argenti. 2022. Ancient egyptian hieroglyphs segmentation and classification with convolutional neural networks. In *The Future of Heritage Science and Technologies: ICT and Digital Heritage*, pages 126–139, Florence, Italy. Springer International Publishing.

A. Barucci, C. Cucci, M. Franci, M. Loschiavo, and F. Argenti. 2021. A deep learning approach to ancient egyptian hieroglyphs classification. *IEEE Access*, 9:123438–123447.

P. Bedi, N. Gupta, and V. Jindal. 2020. Siam-ids: Handling class imbalance problem in intrusion detection systems using siamese neural network. In *Third International Conference on Computing and Network Communications (CoCoNet'19)*, volume 171, pages 780–789, Trivandrum, Kerala, India.

B. Bogacz, F. Feldmann, C. Prager, and H. Mara. 2018. Visualizing networks of maya glyphs by clustering subglyphs. In *Eurographics Workshop on Graphics and Cultural Heritage*, pages 105–111, Vienna, Austria. The Eurographics Association.

L. Bonfante. 1996. The scripts of italy. In P. T. Daniels and W. Bright, editors, *The World's Writing Systems*, chapter 23, pages 297–311. Oxford University Press, Oxford, United Kingdom.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 737—-744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. 2005. Stability of persistence diagrams. In *Proceedings of the Twenty-First Annual Symposium on Computational Geometry*, SCG '05, pages

263—271, New York, NY, USA. Association for Computing Machinery.

M. Corazza, F. Tamburini, M. Valério, and S. Ferrara. 2022. Unsupervised deep learning supports reclassification of bronze age cypriot writing system. *PLOS ONE*, 17(7):e0269544.

S. Daggumati and P. Z. Revesz. 2023. Convolutional neural networks analysis reveals three possible sources of bronze age writings between greece and india. *Information*, 14(4):227.

P. T. Daniels and W. Bright. 1996. *The World's Writing Systems*. Oxford University Press, Oxford, United Kingdom.

J. C. Dunn. 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57.

T. Guidi, L. Python, M. Forasassi, C. Cucci, M. Franci, F. Argenti, and A. Barucci. 2023. Egyptian hieroglyphs segmentation with convolutional neural networks. *Algorithms*, 16(2):79.

A. Hamplová, A. Romach, J. Pavlíček, A. Veselý, M. Čejka, D. Franc, and S. Gordin. 2024. Cuneiform stroke recognition and vectorization in 2d images. *Digital Humanities Quarterly*, 18(1).

J. T. Hooker. 1990. *Reading the Past: Ancient Writing from Cuneiform to the Alphabet*. Barnes & Noble, Inc., New York, NY, USA.

G. Hosszú and F. Kovács. 2016. Topological analysis of ancient glyphs. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 002248–002253, Budapest, Hungary. IEEE.

G. Koch, R. Zemel, and R. Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *32nd International Conference on Machine Learning*, volume 37, Lille, France. JMLR: W&CP.

J. B. Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

X. Liu, W. Gao, R. Li, Y. Xiong, X. Tang, and S. Chen. 2022. One shot ancient character recognition with siamese similarity network. *Scientific Reports*, 12(1):14820.

J. Luo, Y. Cao, and R. Barzilay. 2019. Neural decipherment via minimum-cost flow: From Ugaritic to Linear B. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3146–3155, Florence, Italy. Association for Computational Linguistics.

J. Luo, F. Hartmann, E. Santus, R. Barzilay, and Y. Cao. 2021. Deciphering Undersegmented Ancient Scripts Using Phonetic Prior. *Transactions of the Association for Computational Linguistics*, 9:69–81.

R. Moustafa, F. Hesham, S. Hussein, B. Amr, S. Refaat, N. Shorim, and T. M. Ghanim. 2022. Hieroglyphs language translator using deep learning techniques (scriba). In *2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 125–132, Cairo, Egypt. IEEE.

R. Salomon. 1998. *Indian Epigraphy: A Guide to the Study of Inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan Languages*. Oxford University Press, Oxford, United Kingdom.

T. Sommerschield, Y. Assael, J. Pavlopoulos, V. Stefanak, A. Senior, C. Dyer, J. Bodel, J. Prag, I. Androutsopoulos, and N. de Freitas. 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, 49(3):703–747.

P. Swiggers. 1996. Transmission of the phoenician script to the west. In P. T. Daniels and W. Bright, editors, *The World's Writing Systems*, chapter 21, pages 261–270. Oxford University Press, Oxford, United Kingdom.

S. Tummala and A. K. Suresh. 2023. Few-shot learning using explainable siamese twin network for the automated classification of blood cells. *Medical & Biological Engineering & Computing*, 61:1549—-1563.

M. Ventris and J. Chadwick. 1953. Evidence for greek dialect in the mycenaean archives. *The Journal of Hellenic Studies*, 73:84–103.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

J. H. Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

F. C. Woudhuizen. 2020. The lydian yod-sign. In Gür B. and Dalkılıç S., editors, *A Life Dedicated to Anatolian Prehistory. Festschrift for Jak Yakar*, chapter 32, pages 465–477. Bilgin Kültür Sanat, Ankara, Turkey.

A. List of scripts

To collect the glyph database we have selected all the European, Mediterranean and Middle Eastern writing systems that are implemented in the version 15.0 of the Unicode Standard, see Table 3. Many of these writing systems are alphabetic such as the Latin and Lycian scripts while some of them are abjad, abugida or syllabic writing systems such as the Arabic, Cypriot and Ge`ez scripts (Daniels and Bright, 1996).

Script	Number of glyphs
Arabic	36
Armenian	38
Avestan	54
Carian	49
Caucasian Albanian	52
Chorasmian	21
Coptic	25
Cypriot	55
Cypro-Minoan	97
Cyrillic	32
Elder Futhark	24
Elymaic	22
Ge`ez	26
Georgian Asomtavruli	38
Georgian Mkhedruli	33
Glagolitic	47
Gothic	27
Greek	24
Hatran Aramaic	21
Hebrew	27
Imperial Aramaic	22
Kharoshthi	37
Latin	26
Linear B	60
Lycian	29
Lydian	26
Mandaic	25
Manichaean	36
Meroitic Cursive	24
Meroitic Hieroglyphs	30
Nabataean	31
Ogham	20
Old Hungarian	51
Old Italic	27
Old North Arabian	29
Old Permic	38
Old Persian	36
Old Sogdian	18
Old South Arabian	29
Old Turkic Orkhon	42
Old Turkic Yenisei	31
Pahlavi Inscriptional	19
Pahlavi Psalter	18
Palmyrene	23
Parthian Inscriptional	22
Phoenician	22
Samaritan	22
Sogdian	21
Syriac	26
Tifinagh	31
Ugaritic	30

Table 3: The writing systems used in this work.