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

1. Introduction

We use **siamese neural networks** to compare **glyphs** and **writing systems** with multidimensional scaling and clustering analyses. From **51 alphabets**, we use a Ward-linkage hierarchical clustering and obtain **10 families of scripts** including 3 isolated scripts. This approach has the potential to reveal connections between civilizations and to help the deciphering of ancient scripts.

2. Font-driven database

We create a **database of 1,649 standardized glyphs** from 51 historical European, Mediterranean and Middle Eastern writing systems by using their **Unicode identifiers** and **Noto Sans Regular fonts**.

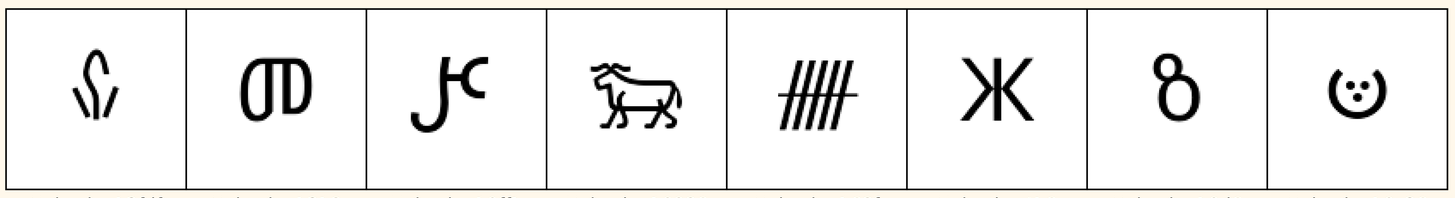

Figure 1: Examples of graphemes from the database.

3. Siamese-based distance

We use the siamese neural network developed in [1] which has been trained and tested on the **Omniglot dataset** [2] to compare **similarities between glyphs**.

For two glyphs g_1 and g_2 we denote by $\mathrm{SNN}(g_1,g_2)$ the similarity predicted by this model and for two scripts s_1 and s_2 we define the distance $d_s(s_1,s_2)$ by

$$d_g(g_1,g_2) = 1-\mathrm{SNN}(g_1,g_2),$$
$$d_s(s_1,s_2) = \frac{1}{2}\left(\frac{1}{|s_1|}\sum_{g_1\in s_1}\min_{g_2\in s_2} d_g(g_1,g_2)+\frac{1}{|s_2|}\sum_{g_2\in s_2}\min_{g_1\in s_1} d_g(g_1,g_2)\right).$$

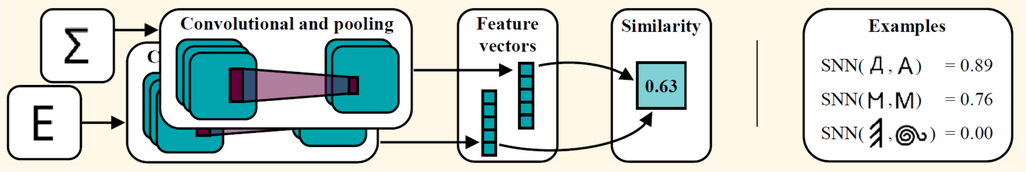

Figure 2: Siamese-type neural network.

4. Comparison of graphemes

Closest pair	Old Sogdian & Pahlavi Psalter	0.05
Farthest pair	Coptic & Old Persian	0.88

Isolated scripts

- Old Persian
- Meroitic Hieroglyphs
- Glagolitic
- Ogham

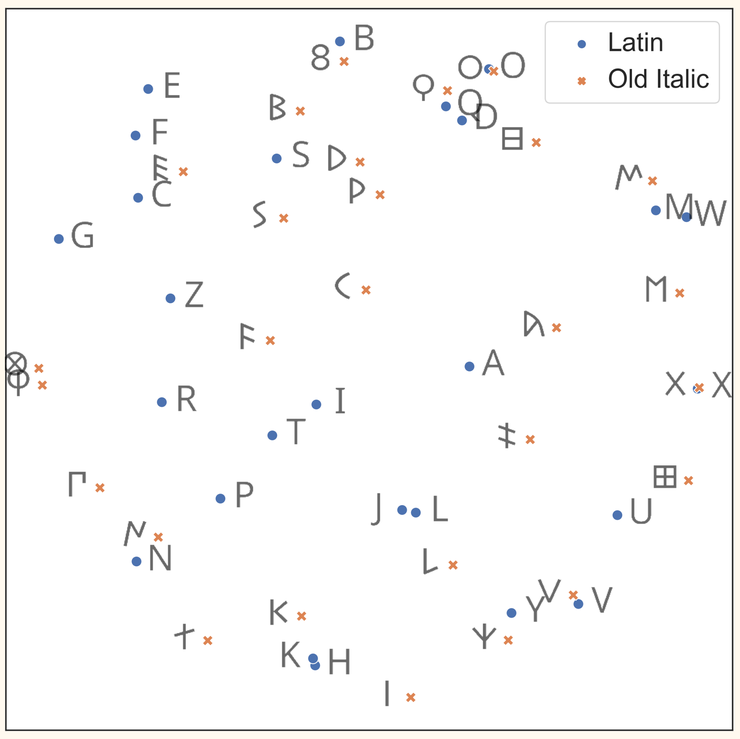

Figure 3: Two-dim. scaling of the Latin and Old Italic scripts (0.26).

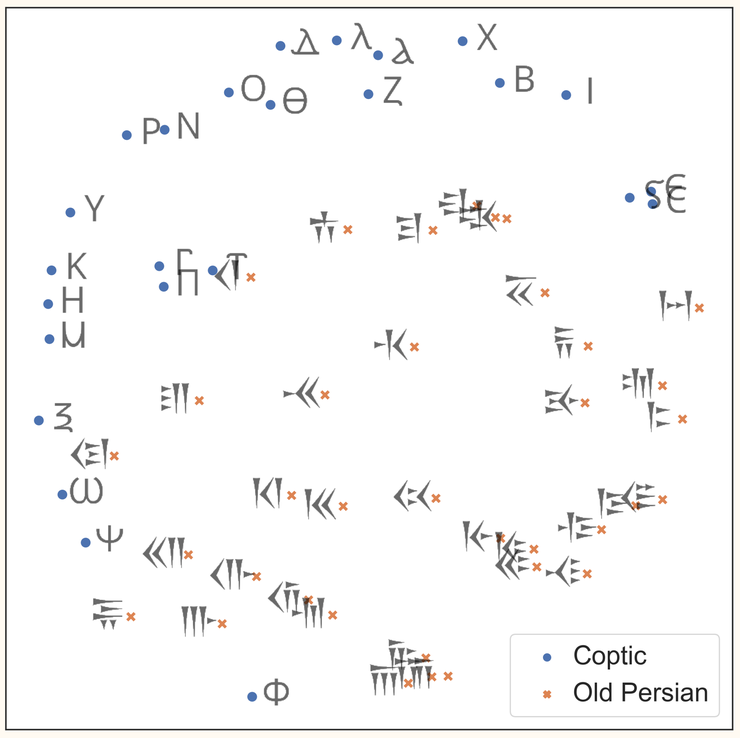

Figure 4: Two-dim. scaling of the Coptic and Old Persian scripts (0.88).

5. Clustering of writing systems

We perform a Ward-linkage **hierarchical clustering** on the 51 scripts. The Elbow method indicates to truncate the **dendrogram** at **10 clusters** where the clustering quality Dunn index is 0.81.

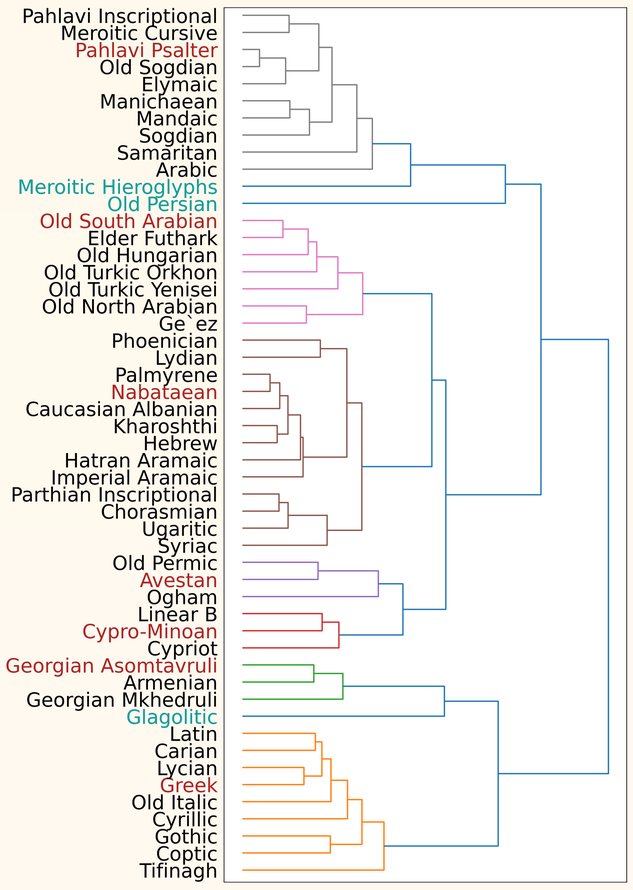

Figure 5: Dendrogram of the Ward-linkage hierarchical clustering of the 51 scripts. Color chart: medoid, isolated script.

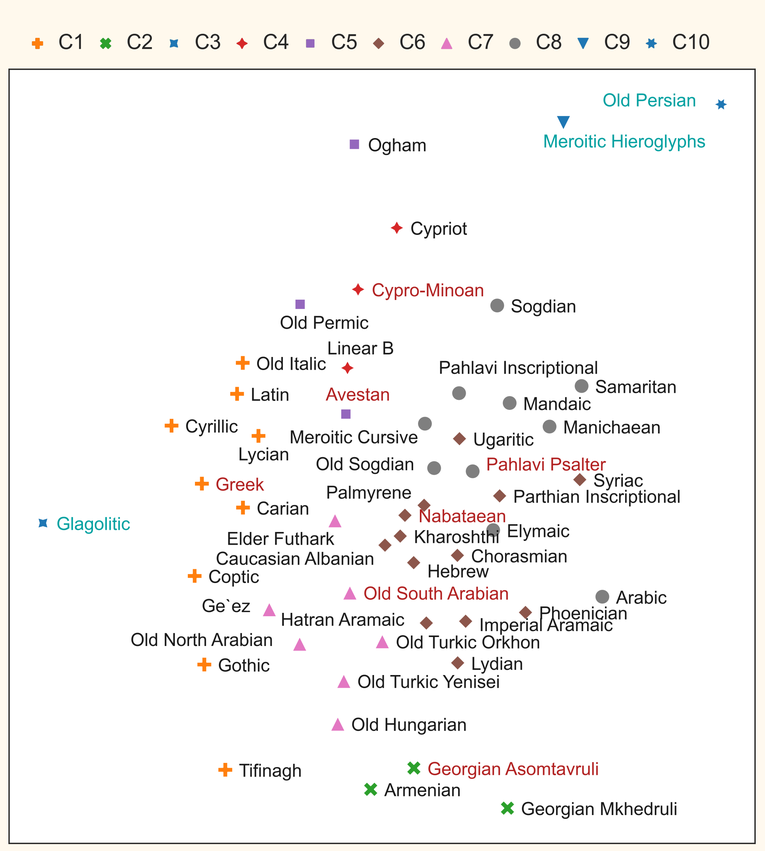

Figure 6: Two-dim. scaling of the 51 scripts. Marker chart: clusters. Color chart: medoid, isolated script.

6. Conclusion

Effectiveness

- The clusters very often represent **real historical connections** such as:
  - ⋆ The Georgian-Armenian group
  - ⋆ The Hebrew-Nabataean group
  - ⋆ The Greek-Latin-Cyrillic group
  - ⋆ The Cypriot group

Limitations

- **Some scripts have no Unicode implementation** such as Paleohispanic scripts.
- **Unclear if applicable to logographic systems** like Chinese characters.

Future work

- Include more scripts to obtain a **larger taxonomy** of world's writing systems.
- Apply it to the **decipherment of lost scripts** by comparing them to known scripts.

7. References

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

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