ADVANCED TECHNIQUES FOR THE DECIPHERMENT OF ANCIENT SCRIPTS

SILVIA FERRARA FABIO TAMBURINI

ABSTRACT: This contribution explores modern and traditional approaches to the decipherment of ancient writing systems. It surveys methods used by paleographers and epigraphers and state-of-the art applications of computational linguistics, such as models based on neural networks. It frames the contextual problems scholars encounter in dealing with ancient codes, the situations and preconditions of the unknown codes, their idiosyncrasies and peculiarities, and the potential solutions afforded by both traditional and novel methods of investigation.

KEYWORDS: decipherment, scripts, epigraphy, paleography, neural networks.

1. INTRODUCTION: DECIPHERING A SCRIPT, IDENTIFYING A LANGUAGE[1]

Script and language are two very different things. Often, in common parlance, they tend to be confused and merged into the same realm, when in fact, they follow parallel paths that can be completely separate. In the history of writing, there are plentiful examples of several languages being recorded by one script (one example is the Roman alphabet, which is used to write many languages, not only of the Romance language family, but many more), as much as one language being recorded by more than one script. For instance, the ancient Greek language was recorded first by the Linear B syllabary in the late second millennium BC, then in the first millennium BC by the Cypriot Syllabary on Cyprus and finally by the Greek alphabet on the mainland of Greece. A crucial distinction between language and script necessarily frames our understanding of decipherment. In particular, we may find the following cases (Gelb & Whiting 1975):

[1] This research is an output of the ERC Project "INSCRIBE. Invention of Scripts and Their Beginnings." The project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 771127) granted to Silvia Ferrara.

- Type 0: known script and known language;
- Type I: unknown script and known language;
- Type II: known script and unknown language;
- Type III: unknown script and unknown language.

Type 0 reveals a transparent relationship, as it represents uncoded plaintext and involves no linguistic analysis of any sort, but mere transcription. Type I involves the systematic application of phonetic values, often through the aid of bilingual texts, onto signs whose value was initially unknown, to then recognise a known, albeit coded, language. Phoenician, Ugaritic, Old Persian, Cypriot, whose languages were known, but their signs values were not, represent traditional examples of this process. Type II needs to rely on an inordinate amount of external data to validate the reconstruction of the underlying language, Sumerian and Hittite being successful cases in point thanks to the generous amounts of Akkadian inscriptions which used the same script. Cases such as Eteocypriot or Eteocretan, for which data is scanty, do not prove so fruitful, although their scripts can be read. In terms of language analysis, this type of decipherment is the most painstaking of all. Type III is the near-impossible reconstruction, as it is hindered by a double negative, thus often rendering any hypothesis moot and its necessary validation altogether impossible (Phaistos Disk, Isthmian being two likely examples). This approximate dashboard serves us as a basic opener to the problems we shall be addressing in this contribution when dealing with scripts that to this day remain undeciphered.

Another distinction is in order. Reading a writing system is different from deciphering it. If reading is the process whereby specific phonetic values can be applied onto specific signs systematically, deciphering implies the identification or reconstruction of the language: its internal structure, morphology and, if possible, syntax. If the language in question is already known, its translation will be the ultimate step in the process (Type I above). If such process is neither completed nor validated, no cogent definitive decipherment can be claimed. Etruscan is a salutary reminder of this all-important distinction, as the script is readable (namely we know the phonetic values of its signs), but the language and its internal structure, the grammar, is barely reconstructed (Bonfante & Bonfante 2002). Deciphering, in other words, implies a full validation of a script's linguistic architecture.

Currently there are several scripts from the ancient world that are yet to be deciphered. The most numerous family of related undeciphered scripts is represented by four cases from the Aegean area, all dating to the broad second millennium BC horizon, with three from the island of Crete (Cretan hieroglyphic, Linear A, Phaistos Disk) and one from the island of Cyprus (Cypro-Minoan). Among other writing systems, two further examples are even more problematic, as their status as proper scripts (that is systems that record a natural language through phonography, with signs corresponding to specific sounds by convention) is *sub judice*. These two cases are the Indus Valley script also called Harappan, dating to c. the fourth millennium BC, and the Rongorongo from Easter Island (Rapa Nui), dating to as late as the eighteenth-century CE. If proved to be proper writing systems, Rongorongo and the Indus Valley could be examples of invented writing systems much in the same way that Mesopotamian, Egyptian, Chinese, and Mesoamerican writing are considered pristine, original inventions. Another undeciphered script from ancient Iran, predominantly originating from Susa, named Proto-Elamite, may share a common origin with proto-cuneiform (Englund 2004; Dahl 2018; Kelley *et al.* 2022) and may be tied to the later-attested Elamite language of the same region, although there is no definitive evidence to prove a positive linguistic relation. There are several other ancient scripts whose languages remain unknown, such as the early Chinese pictographs of the civilisation of the Yellow River, the Byblos syllabary of the II millennium Levantine coast, and the famous Voynich manuscript of Renaissance times. Many more are often added to this list, but they generally present either dubious or unstable inventories of signs, piecemeal attestations, or reveal a contested status as proper writing.

2. PALEOGRAPHIC AND EPIGRAPHIC METHODS

In this section traditional methods of analysis will be laid out. These involve paleographic and epigraphic perspectives, starting from the nature of the undeciphered script at hand (typology of script, the signlist, the dataset, the possible presence of deciphered scripts in association, and cognate scripts in the same family). External and internal decipherment processes will be showcased, the former relying on bilingual scripts found in the same context, the latter relying merely on internal evidence: this case will entail six key methodological steps, upon which any tentative internal decoding process must rely.

2.1 Data and Contexts

With the relationships between language and script being equal, the nature of the data at our disposal can play a fundamental role. A decipherment can be hampered by circumstantial factors, largely to be attributed to two critical points: the amount of data available and the nature of the texts. Both can skew the data considerably. Contextualising the inscriptions is therefore key, in terms of quantity, quality of the inscriptions, and historical circumstances. Two decipherment grounds based on different types of evidence, external and internal, for the data in question. The external evidence for decipherment considers known and understandable texts found in the same context, or in direct association, with an undeciphered script. External, or indirect, evidence is also relational, as it should also include the framing of the broad script affiliation, such as scripts that are cognate to the undeciphered one. Typically, these can aid in the palaeographic reconstruction and genesis of the undeciphered script. Internal, or direct, evidence, as we shall see, centres instead on the undeciphered script, the quality and quantity of its data, its inventory of signs. As such, any approach to its potential decipherment should be typological, distributional, and statistical.

External Evidence

Traditionally, scripts have been deciphered through the aid of external evidence, thanks to the attestations of known texts closely associated to the unknown script. These are bilingual, trilingual, bigraphic or trigraphic texts and they have paved the way for the decipherment. It is sufficient to cite, for example, the most famous examples of the Mesopotamian cuneiform script and the Egyptian hieroglyphs, in which known languages and scripts were used as support to reconstruct unknown ones. It must be noted that in bilinguals the portions of known text often do not represent a precise word-by-word translation of the unknown segments, which, in other words, are coded, and as such they may not stand in a linear reversible process that obeys word-by-word correspondences (Gelb & Whiting 1975). This will make any effort to apply automatic translations difficult at best. As a result, any reliance on bilinguals alone will allow basic access into the linguistic contents of the undeciphered text but may not suffice to reach a systematic decipherment. It can however be effective at a later validation stage when a decipherment is achieved through other means.

Bilinguals or trilinguals have proven helpful in identifying personal names (or toponyms), which are not translated into a different language, but segments of texts that are merely transcribed sign-by-sign. If the positions in the known readable text of personal names (or toponyms) correspond to the positions of personal names in the coded text, correspondences can be spotted, and phonetic values drawn and applied directly to the signs of the sequences in the unknown text. This can help shift a Type III decipherment situation into a Type II, gaining access into a preliminary reading process. This still remains a crucial avenue into the labyrinth of coded text.

External, or rather indirect evidence, can be represented by the broader environment of an undeciphered script, which can give clues for the paleographic analysis, the development of sign shapes through time, the process of adaptation to different language, and the possibility of tracing compression and simplification of sign shapes. For instance, the Aegean scripts that are still undeciphered can benefit from an approach that considers a holistic paleographic reconstruction, from a diachronic perspective from the earliest emergence of the Cretan hieroglyphs and Linear A through to the connection with the Cypriot script, Cypro-Minoan, which descends directly from Linear A. This can allow for the possibility also of reconstructing a broad historical background for the scripts.

Internal Evidence

Since most decipherments were historically achieved through the aid of bilinguals, and since this book is primarily concerned with ancient scripts that are still undeciphered and do not offer bilingual aids, we will focus our attention on perspectives that rely on internal data. The starting point is the number of inscriptions available. In quantity lies the entire potential of discovery. One notorious case is the Phaistos disk, which is a unique specimen of text, bearing fewer than 250 signs altogether: with this amount of data no decipherment is possible. But quantity alone is not the sole qualifying factor. Again for the undeciphered scripts of the Aegean, there are in total fewer than 2000 inscriptions for Linear A, *ca.* 500 for Cretan hieroglyphic, fewer than 300 for Cypro-Minoan. We are dealing with a thoroughly modest number of inscriptions, and a modest number of token signs, in the low thousands altogether. The only deciphered script of this family, Linear B was deciphered on the grounds of approximately 3000 inscriptions (Ventris & Chadwick 1976; Judson 2020). However, in terms of a cogent internal decipherment, further aspects need to be considered.

Inscriptions need also so be of a certain *kind*. They ought to be repetitive, schematic, and coherent. They must show grammatical features, such as inflection, or clear and redundant syntactical elements. Such was the case for the highly inflected Mycenaean Greek of Linear B, that allowed first the internal reconstruction of nominal declensions by Alice Kober (the famous ‘Kober triplets’ 1945; 1948) and the diagnostic features identified by Michael Ventris (Ventris & Chadwick 1953) that clinched the dialectal features inherent in the language. Notably, if few texts are attested and these present a predominance of personal names or toponyms (such is the case with Etruscan), little can be achieved to identify the nature of the language and its attribution to a language family. In such a case, the prevalence of votive dedications, epitaphs, names engraved on mirrors can only get us so far, and in this case, not even the attestation of bilinguals can forward any considerable progress in the study of Etruscan.

2.2 Key Passages and Validation of Internal Decipherment Attempts

From a theoretical perspective a decipherment trajectory can be divided into six steps, each representing an intermediary link in a well-defined analytical chain. If one link is missing or unstable, no decipherment is possible. That said, we should hasten to add that this is a theoretical reconstruction, without a real historical referent. Each undeciphered script is undeciphered in its own way, with peculiar idiosyncrasies that need to be contextualised individually. We shall raise specific issues to highlight these critical points for each script.

Step 1. Inventory of signs. The essential foundation is the gathering of all the inscriptions attested in the given writing system. Script and writing system are not strictly interchangeable terms. In order to ascertain whether a script can be defined as a *bona fide* ‘system’, it needs to have a rationalised, organised, normalised group of signs. These signs constitute an inventory, and this in turn produces a not insignificant advantage, as it offers us the *abc* or basis upon which we can build correspondences between individual signs and their sounds. We could call this the phonetic and phonological architecture. Of course, there are implications of typological nature, since a writing system could be alphabetic, syllabic and logo-syllabic (the latter with a series of signs for ‘words’ or morphemes, called logograms).

The typology of a writing system depends on the definitive number of signs in the normalised repertoire – the more numerous the signs, the more likely that the script is predominantly logographic. Alphabets range around a maximum of 30 signs; syllabaries can reach many hundreds. The syllabary with the fewest signs is the Canadian Aboriginal script Cree (45), followed by the Classical Cypriot Syllabary.

As intuitive as this first step can seem, several scripts are yet to be defined as a normalised system: e.g. the Rongorongo of Easter Island, the Cretan hieroglyphs, the Cypro-Minoan script. The nature of the problem for this impasse varies in each case. For Rongorongo many signs appear extremely similar, thus creating a difficulty in assessing whether they are allographs (signs representing the same sound albeit with minuscule graphic variations in their shapes) or signs with a different sound. For Cretan hieroglyphs, the problem is in the highly iconic or figurative graphic appearance of the signs. On seals and seal impressions they show a high degree of figurativeness, in this case raising the issue of ‘art’ versus proper writing. As we shall see, the relationship between image and sign has always been problematic.

Step 2. Positional frequency of signs. The second step involves a distributional analysis of the signs within sequences, provided that sequences can be singled out in their boundaries. This is possible for Cypro-Minoan (though not always consistently), which uses word-dividers, as well as Cretan hieroglyphs (albeit with sometimes odd separators, especially on the seals), but seems a current hurdle for Rongorongo, whereby strings of signs are laid out continually. A case of ideal separation could be seen in the Old Persian at Persepolis that led to the decipherment of cuneiform, the words were clearly divided by a wedge. Equally consistent was the use of a small vertical line in Linear B.

Positional frequency is part of the graphotactics of a writing system; it can offer precious insight into how a system behaves even when we have no clue as to the phonetic values of its signs. Moreover, it can shed light on internal grammatical patterns, as it constitutes the very foundation of bigraphic and trigraphic cluster analysis, which investigates relations between groups of consecutive signs. We can call these sign ‘contexts’. For Cypro-Minoan, this has been a crucial step in assessing the nature of the script and its alleged internal division (Corazza *et al.* 2022; Skelton *et al.* 2022).

Step 3. Grammatical patterns. If sequences or ‘words’ show a rational middle ground between consistent patterns, repetitions, and distributional variety, there is a strong chance of spotting morphological characteristics of the language. This can be achieved even without testing phonological hypotheses, as there is no need to apply experimental sound values to the signs. This step is nevertheless crucial, as it can show if a language is fusional or agglutinative. In the former case, the grammatical pattern will be shown through a series of identical roots correlated by changing endings. The same can in principle apply to affixes in agglutinative languages, although a large amount of material would be necessary to make these transparent. A clear case of transparent inflection comes from Linear B. This pattern identification, made by Alice Kober, was crucial its internal decipherment.

Step 4. Typological concatenations (Network analysis). This step is, to an extent, almost external to the script, as it is based on epigraphic and archaeological factors. It has been noted that ‘extraction’, the decoding of a message and the plausibility of its interpretation (Houston 2004), must involve the also the script’s context, its situation. The agents, the relations between the agents and the readers of an inscription, the environment in which a script operates, are all part of the equation. Not least important are the types of objects upon which a script is inscribed or incised or painted. If this is particularly true for an appreciation of the materiality of scripts (Flouda 2015; Lomas *et al.*

2007; Steele 2020), their types and functions can give us clues and pointers to identify possible subject matters. For instance, high statues objects such as jewellery or metal objects often will bear personal names. The same applies to cylinder seals. Identifying subject matters is, however, only an intermediate step, not in itself decisive, as the case of Etruscan patently shows.

Again, consistency and significant quantities are key. If several inscriptions are systematically found in a given context, e.g., religious, and a number of sequences attested there are also attested in a completely different context or on objects of a different type, e.g. objects related to administrative activities, a logical correlation can be established between the two sets, which can help determine the nature of the texts. In this way, archaeology is wedded to epigraphy and the study of the inscriptions. This allows us to see these texts in the settings of their usage, to understand what purpose they may have served. Progress has been made in the investigation of Cypro-Minoan using network analysis (Ferrara & Valério 2017). See also Pluta & Nakassis (2003) for Linear A.

Step 5. Reconstructing graphic affiliations. Not all scripts fall into a neatly reconstructed family of systems. Equally, some scripts are the result of heavy adaptations (Houston 2012), both graphically and phonologically. This makes any modelling of synchronic and diachronic correspondences sign-sounds even more problematic. Other scripts can benefit from a reconstruction of graphic affiliation. For these scripts derivation, degree of adaptation, graphic similarity can be assessed with fruitful results. For instance, the Aegean family has benefitted from such ‘holistic’ approaches and can still benefit from further probing into paleographic investigation through time and across the three scripts that boasts a direct lineage (Cretan hieroglyphic, Linear A, Cypro-Minoan). Cypro-Minoan can be now traced back directly and sign-by-sign to Linear A (Valério 2016), so that we can claim with a certain degree of confidence that Cypro-Minoan and Linear B are directly tied through Linear A. This can, in turn, lead us to a confident, if not systematic application of phonetic values onto Cypro-Minoan, allowing us to read it to a partial extent (Ferrara & Valério 2017).

Step 6. Applying phonetic values. The final, definitive step is the definitive one, namely the process of applying phonetic values. Michael Ventris for the decipherment of Linear B relied on several grids of hypothetic values, reached via the distributional and positional analysis in Step 2. This was a basic trial and error exercise, with the prior knowledge that he was facing a fundamentally CV graphotactic system. A marginal reliance on external support provided by the deciphered Cypriot script of the first millennium (which too, recorded a dialect of ancient Greek), gave validation, but the process was predominantly triggered by the domino effect of successful trial experiments with phonetic values. While methodologically this was no mean feat, the positive turn of events was facilitated by the numerous attestations of local toponyms, famously resilient to change through time, and some still in use today. As suggested above, proper nouns tend to be relatively easy to identify, if we consider Thomas Young's preparatory work on personal names inscribed within cartouches on the Rosetta stone – a crucial factor that assisted Champollion's definitive decipherment of the Egyptian Hieroglyphs or George Grotefend's progress with the Old Persian cuneiform script.

All of this shows how indispensable any statistical analysis tends to be, when related scripts are not available, and the only support is internal.

2.3 Iconologies

A problem that is inherent in several writing systems, which will also be treated below in reference to neural networks, is that of the interface between iconic shapes of signs and decorative elements. This is a fine line to tread; an interregnum whereby figurative symbols can be *prima facie* confused for artistic decorations. Indeed, the iconic substratum of signs has historically proven to be a complicating factor for all image-based scripts that underwent ultimately successful decipherment attempts.

It applied to the Rosetta stone itself, since prior to the decipherment, the reigning view was that the Egyptian Hieroglyphs were 'sematographic', meaning that they recorded ideas, not sounds. Champollion himself subscribed to this view, stating in his *Précis* that 'a hieroglyph inscription's appearance is true chaos [..]. The most contradictory objects are placed side-by-side, generating monstrous combinations'. The script's iconicity was the very obstacle to decipherment, its own hidden trap. Before Champollion admitted to himself that the script could be phonetic, the evidence for reading personal names had to be presented with conviction, and the leap to assuming that not only personal names but the whole text was written phonetically accepted as a creditable hypothesis.

The same hurdles were faced by the early scholars of Maya writing system and a long delay was to be endured for its decipherment. The stumbling block was accepting that an image-based script could be phonetic. The same to an extent still applies to our appreciation of the Cretan hieroglyphic on the seals, branded as 'ornamental writing' (Olivier 1986), when it may well represent logographic values and or syllables (Grumach 1963, 1964; Ferrara 2015). The study of the Indus Valley script is still marred by the same approach, as will be apparent below.

3. COMPUTATIONAL TECHNIQUES

This section describes the state-of-the-art regarding computational techniques proposed so far to help and support paleographers in deciphering ancient scripts.

Deciphering an ancient script could be, in general, a very complex task and very often this task has been split into different subproblems in order to obtain specific answers or to simplify the task by decomposing it into simpler problems. In literature, we can find various contributions dealing with all these subproblems and propose computational methods for solving them in some way, often in relation to one specific script. In order, we have to (a) decide if a set of symbols actually represents a writing system, then (b) we have to devise appropriate procedures to isolate or segment the stream of symbols into a sequence of single signs and then (c) reduce the set of signs to the minimal set for the given writing system forming the alphabet (or syllabary, or whatever inventory of signs), identifying all the allographs. Once we have such a minimal, but complete, set of symbols, we can start (d) assigning phonetic values and, finally, (e) trying to match phonetic transcriptions to a specific language. The following sections describe in detail the five points outlined above from a computational point of view.

3.1 Pictures or Language?

When confronted with symbols carved on stones or engraved on tablets or other supports, one of the very first steps regards deciding if these symbols represent some sort of language or some other ways of communicating not attributable to a natural language.

In this direction, two main streams of studies have faced this problem from a computational point of view almost in the same period of time: Rao *et al.* (2009, 2010) analysed the the Indus Valley script, still undeciphered, to ascertain whether the script actually encodes natural language. The authors present some evidence for the linguistic hypothesis by showing that the script's conditional entropy is closer to those of natural languages than various types of nonlinguistic systems.

Almost at the same time, Lee *et al.* (2010) applied a two-parameter decision-tree technique able to distinguish the type of communication expressed in very small corpora. When they applied this technique to a hundred stones expertly carved by the Picts (a Scottish, Iron Age culture) with stylised symbols, they were able to conclude that these were not random or sematographic (heraldic) characters, but, rather, exhibit the characteristics of written language.

Unfortunately, we are very far from reaching a consensus regarding such kind of approaches: a study by Sproat (2010) severely criticises this method and, using a larger set of nonlinguistic and comparison linguistic corpora than were used in these and other studies, he showed that none of the previously proposed methods are useful in determining, with a high degree of certainty, if the considered symbols really represent a writing system or not proposing, at the same time, a novel measure based on repetition that classify them as nonlinguistic, contradicting the findings of the previous works.

3.2 Script Segmentation

One significant challenge to undeciphered scripts is word and sign segmentation. These are two basic units that should be clearly identified before starting the decipherment process, either by hand or with the support of computational techniques. The same problem arises when trying to build electronic corpora for undeciphered scripts, which is a necessary starting point for computational epigraphy, and requires a large human effort for their preparation from raw archaeological data.

Palaniappan & Adhikari (2017) proposed an automatic tool based on machine learning algorithms to aid epigraphical research which presents a deep learning pipeline that takes as input images of the undeciphered Indus script and provides as output a string of graphemes, suitable for inclusion in a standard corpus. The input image is (a) decomposed into regions and (b) these regions are classified as containing textual and/or graphical information using a convolutional neural network. Text regions were (c) hierarchically merged and trimmed to remove non-textual information and then (d) segmented using standard image processing techniques to isolate individual graphemes. In the last step, another convolutional neural network classifies the graphemes exhibiting quite high accuracy showing the great potential of deep learning approaches in computational epigraphy. In addition, the work presented by Luo *et al.* (2021) jointly models word segmentation and cognate alignment, by relying on phonological constraints, with a generative stochastic model, also introducing a method for identifying close languages (see Section 3.5).

Figure 1 shows a fragment of Rongorongo, a script from Easter Island probably recording the local Rapanui language. As remarked by Davletshin (2012) and Valério *et al.* (2022), it is still not clear how to segment this script into linguistic units (either sounds, syllables or morphemes) and, moreover, different small shapes, almost identical, appear combined in different ways to form complex signs. The anthropomorphic glyphs are an example: they resemble a similar picture, but small additional elements are added to it.

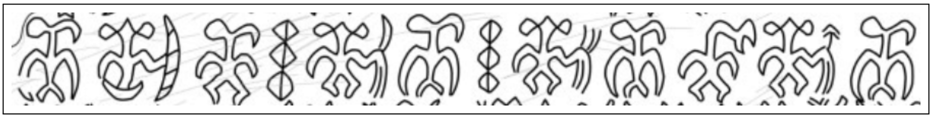

FIGURE 1: A FRAGMENT OF THE RONGORONGO SCRIPT EXTRACTED FROM LINE 3 OF THE ÉCHANCRÉE TABLET, SIDE A (LASTILLA *et al.* 2021).

3.3 Building a Uniform Set of Signs

One of the first problems scholars have to solve, after having found a method for segmenting the script into meaningful linguistic units, regards the identification of a sign-list. This is not an easy task because scribe writing styles and evolution of symbols through time could introduce variations which hinder allograph identification and management.

Skelton (2008) and Skelton & Firth (2016) applied phylogenetic systematics, a technique developed in biology for reconstructing evolutionary histories of organisms, to writing systems, in particular to Linear B, a pre-alphabetic Greek script. By using this technique, they were able to study the evolution of Linear B script through time considering also scribal hands as a further source of variation demonstrating the effectiveness of phylogenetic analysis.

Born *et al.* (2019) applied computational linguistics to analyse the undeciphered Proto-Elamite script. In particular, they used three different clustering algorithms to create and examine groups of signs based on the way they occur and co-occur within texts.

Corazza *et al.* (2022) present a study devoted to the analysis of the Cypro-Minoan syllabary. They applied a method that can aid to shed light on the tripartite division (CM1, CM2, CM3) of Cypro-Minoan, to assess if it holds up against a multi-pronged, multi-disciplinary approach. This involved considerations linked to paleography and epigraphy, and crucially, deep learning-based strategies. The usage of an unsupervised state-of-the-art convolutional neural model not using any prior knowledge of the script allowed them to establish that the use of different media skews to a large extent the uniformity of the sign shapes as the application of several neural techniques highlight graphic proximity among signs inscribed on similar supports. Moreover, all their results point in the same direction, namely the validation of a unitary, single Cypro-Minoan script, rather than the division into three subgroups currently discussed in the literature. This conclusion shows that most signs differences are due to the epigraphic supports, and help to rationalise the sign inventory of Cypro-Minoan script proposed by Olivier (2007), suggesting several sign mergers.

3.4 Assigning Signs Values (phonetic/numeric)

This is a fundamental step in the decipherment, but, as we will see in the following section, most of the works in the literature jointly solved it with the task of determining the language written with the examined script.

As a notable exception Corazza *et al.* (2021) applied computational techniques, mainly constraint programming and optimisation methods, for proposing numerical values to the fraction signs of Linear A. Minoan Linear A is an undeciphered script mainly used for administrative purposes on Bronze Age Crete. One of its most problematic features is the exact mathematical values of its system of numerical fractions. The authors addressed this issue through a multi-stranded methodology that comprises palaeographical examination and statistical, computational, and typological approaches. Building on previous analyses, which suggested hypothetical values for some fractions, they extended their probe into assessing values for some problematic ones. The results, which were based on a close palaeographical analysis and on computational, statistical and typological strategies, show a remarkable convergence and point towards a systematic assignment of mathematical values for the Linear A fraction signs.

3.5 Fixing Signs Values and Match Sequences with a Known Language

Any modern attempt to decipher lost scripts using computational tools is based on the comparison of a lost script/language wordlist with words of a known deciphered language. These computational approaches have to address two main problems:

- the first regards the possibility that the two scripts do not correspond. In this case also the phonological values of the lost symbols could be unknown and the matching between the two wordlists must be preceded by some matching between scripts;
- the two wordlists must be matched in some way searching for “cognate” words, i.e. words in different languages that can share an etymological ancestor in a common parent language.

Some scholarly works concentrate only on cognate detection within the same script (Bouchard-Côté *et al.* 2009) or using directly the International Phonetic Alphabet sound representations (Hall & Klein 2010). In both cases the tested languages were typologically very similar.

Conversely, the most advanced recent studies on the automatic decipherment of lost languages proposed systems producing both sign mappings between different scripts and mapping of words into their corresponding cognates (e.g. Snyder *et al.* 2010; Berg-Kirkpatrick & Klein 2011; Luo *et al.* 2019, 2021). These studies share a common view on the computational approach: they structured the algorithm as a two-step procedure, taking inspiration from the Expectation-Maximization (EM) algorithm, an iterative method to find (local) maxima or minima. The first step proposes a temporary working matching between the two "alphabets"[2] and the second step, by relying on the script-matching, tries to match the two word lists proposing possible cognates. At the beginning of the process the scripts matching will be almost random, and so the cognate matching, but, after several iterations the whole process should converge proposing both a script-matching and a list of possible cognates. The key point hinges on finding an appropriate function, to be optimised by this iterative process, representing in an optimal way the concept of matching between words including also some linguistic constraints regarding scripts, words and possibly sounds. Let us review the most relevant analyses, in our view, which tackle the decipherment problem in an automatic way, all following the general scheme just discussed.

Snyder *et al.* (2010) presented the first paper which adopts the modern approach to the computational decipherment problem: their method requires a non-parallel corpus in a known related language and produces both alphabetic mappings and translations of words into their corresponding cognates, employing a non-parametric Bayesian framework to simultaneously capture both low-level character mappings and high-level morphemic correspondences. They tested this method on Ugaritic, an ancient Semitic language, comparing it with old Hebrew: the model correctly maps 29 of 30 signs to their old Hebrew counterparts, and deduces the correct Hebrew cognate for 60% of the Ugaritic words that have cognates in Old Hebrew. The code for this method is not available.

Berg-Kirkpatrick & Klein (2011) took a different approach: they devised a simple objective function that, when optimised, yields accurate solutions to both decipherment and cognate pair identification problems. Their system requires only a list of words in both languages as input. The proposed solution is both simple and elegant: binary variables govern both the matching between signs in the two languages and the matching between the two lexica. By applying a simple integer combinatorial optimisation procedure, their system was able to obtain good results on the same problem introduced by Snyder *et al.* (2010) and on a new matching task on romance languages. Unfortunately, their code is not available, but reproducing this approach seems quite simple as it is described clearly in the paper.

[2] Here, with the term "alphabet", we indicate a generic notion of inventory of signs, glyphs, etc. used as a writing system.

Luo *et al.* (2019) present a novel neural approach that, in our opinion, is the most sophisticated and promising for the automatic decipherment of unidentified languages. To compensate for the lack of strong supervision information, their model is designed to include known patterns in language change documented by historical linguistics. The mapping between signs is carried out by a bidirectional recurrent neural network while the procedure for matching cognates is formalised as a minimum-cost flow problem. They applied this method to the same problem presented in Snyder *et al.* (2010), a sort of benchmark in this field, and on a brand new dataset that included Linear B and ancient Greek lexica obtaining thus very good mapping results. As an added bonus, all code and datasets for reproducing their results are available to the community.

In the work by Luo *et al.* (2021), the authors faced a more difficult hurdle considering scripts that are not fully segmented into words and contexts in which the closest known language is not determined. By building on rich linguistic constraints reflecting consistent patterns in historical sound change, they were able to capture natural phonetic geometry by learning character embeddings based on the International Phonetic Alphabet. The resulting generative framework jointly models word segmentation and cognate alignment, informed by phonetic/phonological constraints. They tested their method on both identified languages, namely Gothic and Ugaritic, and an undeciphered one, namely Iberian, showing that incorporating phonetic geometry leads to clear and consistent gains. They also proposed a measure for language closeness which correctly identifies related languages for Gothic and Ugaritic. The authors provide their code and data.

3.6 Other Computational Tools to Help Paleographers

Inscriptions are studied by epigraphy to extract evidence of the thought, language, society and history of past civilisations, but many of them have been damaged. Trying to restore, as far as possible, these precious sources would provide some further information to increase and enrich our knowledge on a given population. Assael *et al.* (2022) present Ithaca, a deep neural network for the textual restoration, geographical and chronological attribution of ancient Greek inscriptions. Ithaca is designed to assist the historian's work and was able to improve their accuracy when reading and attributing inscriptions, unlocking the cooperative potential between artificial intelligence and historians. In the same line, Fetaya *et al.* (2020) present a method for modelling the language written on clay cuneiform tablets using recurrent neural networks in order to assist scholars completing the breaks in ancient Akkadian texts from Achaemenid period Babylonia.

A relevant source of variation in ancient writing system interpretation regards the intrinsic variation introduced by scribal hands. One of the problems of palaeography is to determine writer identity, or differences when the writing style is not always the same. Srivatsan *et al.* (2021) investigate the use of neural feature extraction tools in performing scribal hand analysis on the Linear B writing system. A fully unsupervised neural network assigns each sign written by the same scribal hand a shared vector embedding to represent that author's stylistic patterns, and each sign representing the same syllable a shared vector embedding to represent the identifying shape of that character. They trained their system using both a reconstructive loss governed by a decoder that seeks to reproduce glyphs from their corresponding embeddings, and a discriminative loss which measures the model's ability to predict whether or not an embedding corresponds to a given Linear B sign image.

A similar work by Popović *et al.* (2021) examined one of the Dead Sea Scrolls, namely the Great Isaiah Scroll, and, by using pattern recognition and artificial intelligence techniques, demonstrates that two main scribes, each showing different writing patterns, were responsible for writing the scroll. This sheds new light on the Bible's ancient scribal culture and provides evidence that ancient biblical texts were not copied by a single scribe only.

Finally, in Lastilla (2022) we can find further evidence that automatic techniques, namely self-supervised learning applied to convolutional neural networks, can be successfully applied to the problem of handwriting identification for Medieval and modern manuscripts revealing the strong potential of self-supervised techniques in the field of digital paleography, where unlabelled data is common, and labelled data is hard to be produced.

3.7 Wrapping up on Computational Techniques

This review of the most promising studies for the decipherment of ancient scripts might be seen to suggest that these tools can solve all the unsolved problems, of palaeographic, epigraphic and linguistic nature, debated for years by experts. This is naturally not the case. These techniques, even if very promising, also present a large number of problems when applied to real decipherment attempts:

- first of all, segmented and clean corpora are needed. Building a corpus for an ancient undeciphered script, even in the case where we have already solved the segmentation problems and were able to collect single sign images and sign/word sequences, is not an easy task. Most inscriptions are damaged and many signs are not readable. Broken words and/or partial sentences are also frequent. When tested, reliability of some of the cited works decreases considerably with partial or corrupted data;

- an extensive cognate list must be available, but in most real cases we only have two word lists that must be matched without any guarantee that lost language cognates are really present in the known language lexicon;
- the two-step approach based on EM algorithm is ingenious, solving two problems jointly, but it is prone to getting stuck into local maxima/minima during the optimisation process producing sub-optimal results that are not able to provide any real insight;
- in NLP we have to make evaluation on well-known test beds and all the studies we discussed before worked on well known correspondences (e.g. Linear B/Mycenaean Greek, Ugaritic/Old Hebrew, etc.) to prove the system effectiveness. It is an entirely different matter to test the same systems on real cases when we have to deal with unknown writing systems and their corresponding languages?

In the light of these considerations, we agree with Sproat (2020), who suggests that these tools can help paleographers shed light on the decipherment process, but we cannot rely on them only for providing a complete solution to our real problems without any human intervention for guiding the process and interpreting the results.

4. SYNERGIES

The present contribution is animated by a spirit of quiet optimism when it comes to decipherment attempts. A successful one seems to have been achieved for the Linear Elamite script recently (Desset *et al.* 2022). Just as was the case for the decipherment of Maya (Coe 1993), the Linear Elamite success was the result of a synergistic collaboration between many scholars. Decipherment has become a cooperative field. There is not only group action but of intellectual convergence: archaeologists, philologists, epigraphists, historians. Traditional methods – those encompassing the expertise of scholars in the humanities, blending paleography and linguistics – are irreplaceable, but it would be short-sighted to write off other potentially useful approaches. Science-based methods are well worth considering, though a person should not lazily rely exclusively on them to achieve results.

Any reconstruction needs to be multifaceted. It needs to be archaeological, focusing on context and use of a script, explaining it at the macro level. It needs to be paleographic, concerned with the changing shapes of signs, their development, their graphic idiosyncrasies. It needs to be linguistic, seeking to understand which sounds are recorded, and it needs to also be anthropological: we need to discover why these scripts came about in the first place. Deep learning strategies, as we saw, offer something that up until a few years ago was unthinkable: they enable us to take control over what we choose to do manually. In the last five to ten years, deep learning algorithms have proved effective at detecting similar patterns in different entities or realms. The crucial function in the realm of decipherment is disambiguation. In practice, they can validate our reconstructions, they can verify if we are properly grouping like with like via the traditional method. They can aid in assessing whether we are dealing with allographs or separate signs with different sounds. They can cross-check sequences of undeciphered writing *vis à vis* known readable sequences. They can aid in morphological reconstruction. Computers may not be a *deus ex machina*, but can, to a solid degree of methodological soundness, be efficient co-pilots. It is with an attitude of intellectual open-mindedness that we conceived this contribution.

REFERENCES

Assael, Y., T. Sommerschield, B. Shillingford, M. Bordbar, J. Pavlopoulos, M. Chatzipanagiotou, I. Androutsopoulos, J. Prag & N. de Freitas (2022). Restoring and attributing ancient texts using deep neural networks. *Nature*, 603. 280–283.

Berg-Kirkpatrick, T. & D. Klein (2011). Simple effective decipherment via combinatorial optimization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 313–321.

Bonfante, G. & L. Bonfante (2002). *The Etruscan Language: An Introduction*. Manchester: Manchester University Press.

Born, L., K. Kelley, N. Kambhatla, C. Chen & A. Sarkar (2019). Sign clustering and topic extraction in Proto-Elamite. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis, USA: Association for Computational Linguistics, 122–132.

Bouchard-Côté, A., T.L. Griffiths & D. Klein (2009). Improved reconstruction of protolanguage word forms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics, 65–73.

Coe, M. (1993). *Breaking the Maya Code*. London: Thames and Hudson.

Corazza, M., S. Ferrara, B. Montecchi, F. Tamburini & M. Valério (2021). The mathematical values of fraction signs in the linear a script: A computational, statistical and typological approach. *Journal of Archaeological Science*, 125. 105214.

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

Dahl, J. (2018). The Proto-Elamite Writing System. In J. Álvarez Mon, G. Basello & Y. Wicks (eds.) *The Elamite World*, 383–396. london: Routledge.

Davletshin, A. (2012). Numerals and phonetic complements in the kohau rongorongo script of easter island. *Journal of the Polynesian Society*, 121(3). 243–274.

Desset, F., K. Tabibzadeh, M. Kervran, G. Basello & G. Marchesi (2022). The decipherment of linear elamite writing. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 112(1). 11–60.

Englund, R. (2004). The State of Decipherment of Proto-Elamite. In S. Houston (ed.) *The First Writing: Script Invention as History and Process*, 100–149. Cambridge, UK: Cambridge University Press.

Ferrara, S. (2015). The beginnings of writing on crete: Theory and context. *The Annual of the British School at Athens*, 110. 27–49.

Ferrara, S. & M. Valério (2017). Contexts and repetitions of cypro-minoan inscriptions: Function and subject matter of clay balls. *Bulletin of the American Schools of Oriental Research*, 378. 71–94.

Fetaya, E., Y. Lifshitz, E. Aaron & S. Gordin (2020). Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37). 22743–22751.

Flouda, G. (2015). Materiality and script: constructing a narrative on the minoan inscribed axe from the arkalochori cave. *SMEA Nuova Serie*, 1. 43–56.

Gelb, I. & R. Whiting (1975). Methods of decipherment. *The Journal of the Royal Asiatic Society of Great Britain and Ireland*, 2. 95–104.

Grumach, E. (1963). Neue hieroglyphische siegel/nachtrag zu 'zwei hieroglyphische seigel'. *Kadmos*, 2(1). 7–13.

Grumach, E. (1964). Studies in the structure of some ancient scripts. iii. the structure of the cretan hieroglyphic script. *Bulletin of the John Rylands Library*, 46. 346–384.

Hall, D. & D. Klein (2010). Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, 1030–1039.

Houston, S.E. (2012). *The Shape of Script: How and Why Writing Systems Change*. School for Advanced Research Advanced Seminar Series.

Judson, A. (2020). *The Undeciphered Signs of Linear B: Interpretation and Scribal Practices (Cambridge Classical Studies)*. Cambridge: Cambridge University Press.

Kelley, K., L. Born, M. M.W. & A. Sarkar (2022). Image-Aware Language Modeling for Proto-Elamite. *Lingue e Linguaggio*, XXI(2). This issue.

Kober, A. (1945). Inflection in linear class b: I-declension. *AJA*, 50. 268–276.

Kober, A. (1948). The minoan scripts: Fact and theory. *AJA*, 52. 82–103.

Lastilla, L. (2022). Enhancement of scribal hands identification via self-supervised learning. In *Italian Research Conference on Digital Libraries - IRCDL 2022*. Padova.

Lastilla, L., R. Ravanelli, M. Valério & S. Ferrara (2021). Modelling the Rongorongo tablets: A new transcription of the Échancrée tablet and the foundation for decipherment attempts. *Digital Scholarship in the Humanities*, 37(2). 497–516.

Lee, R., P. Jonathan & P. Ziman (2010). Pictish symbols revealed as a written language through application of shannon entropy. *Proceedings of The Royal Society A*, 466. 2545–2560.

Lomas, K., R. Whitehouse & J. Wilkins (2007). *Literacy and the state in the ancient Mediterranean*, *Accordia specialist studies on the Mediterranean*, volume 7. Accordia Research Institute.

Luo, J., Y. Cao & 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*. Florence, Italy: Association for Computational Linguistics, 3146–3155.

Luo, J., F. Hartmann, E. Santus, R. Barzilay & Y. Cao (2021). Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9. 69–81.

Olivier, J.P. (1986). Cretan writing in the second millennium b.c. *World Archaeology*, 17(3). 377–389.

Olivier, J.P. (2007). *Édition holistique des textes chypro-minoens*. Pisa/Roma: Fabrizio Serra Editore.

Palaniappan, S. & R. Adhikari (2017). Deep learning the indus script. *arXiv e-print*, 1702.00523.

Pluta, N. & D. Nakassis (2003). Linear A and Multidimensional Scaling. In K. Foster & R. Laffineur (eds.) *Metron. Measuring the Aegean Bronze Age, Proceedings of the 9th International Aegean Conference*, 335–342. Aegaeum, 24.

Popović, M., M.A. Dhali & L. Schomaker (2021). Artificial intelligence based writer identification generates new evidence for the unknown scribes of the dead sea scrolls exemplified by the great isaiah scroll (1qisaa). *PLOS ONE*, 16(4). 1–28.

Rao, R.P.N., N. Yadav, M.N. Vahia, H. Joglekar, R. Adhikari & I. Mahadevan (2009). Entropic evidence for linguistic structure in the indus script. *Science*, 324(5931). 1165–1165.

Rao, R.P.N., N. Yadav, M.N. Vahia, H. Joglekar, R. Adhikari & I. Mahadevan (2010). Commentary and discussion: Entropy, the indus script, and language: A reply to R. Sproat. *Computational Linguistics*, 36(4). 795–805.

Skelton, C. (2008). Methods of using phylogenetic systematics to reconstruct the history of the linear b script. *Archaeometry*, 50. 158–176.

Skelton, C. & R. Firth (2016). A study of the scribal hands of knossos based on phylogenetic methods and find-place analysis. *Minos*, 39. 159–188.

Skelton, C., L. Selvig, D. Chen, N. Srivastan & T. Berg-Kirkpatrick (2022). Cypro-minoan:one language or three? an exercise in phonology-based statistical analysis. *Lingue e Linguaggio*, XXI(2). This issue.

Snyder, B., R. Barzilay & K. Knight (2010). A statistical model for lost language decipherment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, 1048–1057.

Sproat, R.W. (2010). A statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2). 1–27.

Sproat, R.W. (2020). Translating lost languages using machine learning? Wellformedness. https://doi.org/http://www.wellformedness.com/blog/translating-lost-languages-machine-learning/.

Srivatsan, N., J. Vega, C. Skelton & T. Berg-Kirkpatrick (2021). Neural representation learning for scribal hands of linear b. In E.H. Barney Smith & U. Pal (eds.) *Document Analysis and Recognition – ICDAR 2021 Workshops*. Springer International Publishing, 325–338.

Steele, P. (2020). Material entanglements of writing practices in the bronze age aegean and cyprus. *Sustainability*, 12(24).

Valério, M. (2016). *Investigating the Signs and Sounds of Cypro-Minoan*. Ph.D. thesis.

Valério, M., L. Lastilla & R. Ravanelli (2022). The Rongorongo Tablet C: New Technologies and Conventional Approaches to an Undeciphered Text. *Lingue e Linguaggio*, XXI(2). This issue.

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

Ventris, M. & J. Chadwick (1976). *Documents in Mycenaean Greek*. Cambridge: Cambridge University Press.

Silvia Ferrara
University of Bologna
Department of Classical Philology and Italian Studies (FICLIT)
via Zamboni, 32 - Bologna
Italy
e-mail: s.ferrara@unibo.it
https://orcid.org/0000-0003-2498-7666

Fabio Tamburini
University of Bologna
Department of Classical Philology and Italian Studies (FICLIT)
via Zamboni, 32 - Bologna
Italy
e-mail: fabio.tamburini@unibo.it
https://orcid.org/0000-0001-7950-0347